

Philippe Blanchard
Felix Bühlmann
Jacques-Antoine Gauthier *Editors*

Advances in Sequence Analysis: Theory, Method, Applications

Advances in Sequence Analysis: Theory, Method, Applications

Life Course Research and Social Policies

Volume 2

Series Editors:

Laura Bernardi

Dario Spini

Michel Oris

Life course research has been developing quickly these last decades for good reasons. Life course approaches focus on essential questions about individuals' trajectories, longitudinal analyses, cross-fertilization across disciplines like life-span psychology, developmental social psychology, sociology of the life course, social demography, socio-economics, social history. Life course is also at the crossroads of several fields of specialization like family and social relationships, migration, education, professional training and employment, and health. This Series invites academic scholars to present theoretical, methodological, and empirical advances in the analysis of the life course, and to elaborate on possible implications for society and social policies applications.

For further volumes:
<http://www.springer.com/series/10158>

Philippe Blanchard · Felix Bühlmann
Jacques-Antoine Gauthier
Editors

Advances in Sequence Analysis: Theory, Method, Applications



Editors

Philippe Blanchard
Institute of Political and International
Studies
University of Lausanne
Lausanne
Switzerland

Jacques-Antoine Gauthier
Centre LINES/LIVES
University of Lausanne
Lausanne
Switzerland

Felix Bühlmann
Centre LINES/LIVES
University of Lausanne
Lausanne
Switzerland

ISSN 2211-7776
ISBN 978-3-319-04968-7
DOI 10.1007/978-3-319-04969-4
Springer Cham Heidelberg New York Dordrecht London

ISSN 2211-7784 (electronic)
ISBN 978-3-319-04969-4 (eBook)

Library of Congress Control Number: 2014936639

© Springer New York Heidelberg Dordrecht London 2014

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

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

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

Printed on acid-free paper

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

Contents

1	Introduction: Sequence Analysis in 2014	1
	Jacques-Antoine Gauthier, Felix Bühlmann and Philippe Blanchard	
Part I How to Compare Sequences		
2	Motif of Sequence, Motif in Sequence	21
	Shin-Kap Han	
3	Using Optimal Matching Analysis in Sociology: Cost Setting and Sociology of Time	39
	Laurent Lesnard	
4	Distance, Similarity and Sequence Comparison	51
	Cees H. Elzinga	
5	Three Narratives of Sequence Analysis.....	75
	Brendan Halpin	
Part II Life Course Sequences		
6	New Perspectives on Family Formation: What Can We Learn from Sequence Analysis?	107
	Anette Eva Fasang	
7	Developmental Psychologists' Perspective on Pathways Through School and Beyond	129
	Julia Dietrich, Håkan Andersson and Katariina Salmela-Aro	
8	Sequence Analysis and Transition to Adulthood: An Exploration of the Access to Reproduction in Nineteenth-Century East Belgium	151
	Michel Oris and Gilbert Ritschard	

Part III Political Sequences

- 9 Trajectories of the Persecuted During the Second World War: Contribution to a Microhistory of the Holocaust.....** 171
Pierre Mercklé and Claire Zalc
- 10 A Contextual Analysis of Electoral Participation Sequences** 191
François Buton, Claire Lemercier and Nicolas Mariot
- 11 Governance Built Step-by-Step: Analysing Sequences to Explain Democratization** 213
Matthew Charles Wilson

Part IV Visualisation of Sequences and Their Use for Survey Research

- 12 Sequence as Network: An Attempt to Apply Network Analysis to Sequence Analysis.....** 231
Ivano Bison
- 13 Synchronising Sequences. An Analytic Approach to Explore Relationships Between Events and Temporal Patterns** 249
Denis Colombi and Simon Paye
- 14 Graphical Representation of Transitions and Sequences.....** 265
Christian Brzinsky-Fay
- 15 Patterns of Contact Attempts in Surveys.....** 285
Alexandre Pollien and Dominique Joye

Contributors

Håkan Andersson Department of Psychology, Stockholm University, Stockholm, Sweden

Ivano Bison Department of Sociology and Social Research, University of Trento, Trento, Italy

Philippe Blanchard Institute of Political and International Studies, University of Lausanne, Lausanne, Switzerland

Christian Brzinsky-Fay WZB Berlin Social Science Center, Berlin, Germany

Felix Bühlmann Institute of Social Sciences, National Centre of Competence in Research LIVES, University of Lausanne, Lausanne, Switzerland

François Buton Center for the Political Study of Latin Europe (CEPEL), CNRS, Montpellier, France

Denis Colombi Centre de sociologie des organisations, Science-Po-CNRS, Paris, France

Julia Dietrich Institute of Educational Science, Department of Educational Psychology, University of Jena, Jena, Germany

Cees H. Elzinga VU University Amsterdam, Amsterdam, The Netherlands

Anette Eva Fasang Social Sciences, Humboldt University Berlin, Berlin, Germany; Demography and Inequality, WZB Social Science Research Center Berlin (WZB), Berlin, Germany

Jacques-Antoine Gauthier Institute of Social Sciences, Life Course and Inequality Research Centre, University of Lausanne, Lausanne, Switzerland

Brendan Halpin Department of Sociology, University of Limerick, Limerick, Ireland

Shin-Kap Han Department of Sociology, Seoul National University, Seoul, Korea

Dominique Joye FORS-UNIL, Lausanne, Switzerland

Claire Lemercier Center for the Sociology of Organizations (CSO), CNRS, Paris, France

Laurent Lesnard Sciences Po, Paris, France

Nicolas Mariot European Center for Sociology and Political Science (CESSP), CNRS, Paris, France

Pierre Mercklé Centre Max Weber, ENS de Lyon, CNRS, Lyon, France

Michel Oris Institute for Demographic and Life Course Studies, LIVES, University of Geneva, Geneva, Switzerland

Simon Paye Centre de sociologie des organisations, Science-Po-CNRS, Paris, France

Alexandre Pollien FORS-UNIL, Lausanne, Switzerland

Gilbert Ritschard Institute for Demographic and Life Course Studies, LIVES, University of Geneva, Geneva, Switzerland

Katriina Salmela-Aro Department of Psychology, University of Jyväskylä, Helsinki, Finland; Helsinki Collegium for Advanced Studies, University of Helsinki, Helsinki, Finland

Matthew Charles Wilson The Pennsylvania State University, Pennsylvania, USA

Claire Zalc Institut d'histoire moderne et contemporaine, ENS Ulm, CNRS, Paris, France

About the Authors

Håkan Andersson did his thesis work within developmental psychology at the Department of Psychology at the University of Stockholm, Sweden. He is now an analyst at the Swedish Higher Education Authority. His research has mainly focused on self-regulation during childhood and adolescence using both larger longitudinal data sets and short-term intensive data. He has also done research on the effect of different forms of therapy on psychological well-being and symptom severity. He has a special interest in person-oriented methodology and intra-individual micro dynamics.

Ivano Bison is Assistant Professor of Sociology at the University of Trento (Italy). His research interests include inter- and intra-generational career mobility and transition from school to work. Currently he is concentrating on the theoretical and methodological aspects of analysis of processes of social change analyzed from the point of view of narrative and sequences analysis ('Lexicographic index: A new measurement of resemblance among sequences', in M. Williams and P. Vogt (eds.), *The SAGE Handbook of Innovation in Social Research Methods*, London, SAGE, 2011; with Franzosi R., Temporal Order: Sequence Analysis; in R. Franzosi, *Quantitative Narrative Analysis*, SAGE, 2010).

Philippe Blanchard is a political scientist at the University of Lausanne. He works on political sociology, political communication, and methods for social and political sciences. He is presently conducting sequence research about Moroccan party activists, French AIDS activists, gendered careers in Swiss trade unions and board members of European international companies. He has taught sequence analysis in Austria, Denmark, France, Switzerland and the US. He is chair of the Standing Group on Political Methodology of the European Consortium for Political Research and member of the Advisory Board of the ECPR Methods Program.

Christian Brzinsky-Fay is senior researcher in the research unit "Skill Formation and Labour Markets" at the WZB Berlin Social Science Center, where he coordinates the "College for Interdisciplinary Educational Research". This is a PostDoc programme for sociologists, economists, educational scientists and psychologists, which is funded by the German Federal Ministry of Education and Research and the Jacobs Foundation. He studied Political Science at the Free University Berlin and

did his doctoral studies in Social Sciences at the University of Tampere (Finland). Together with Ulrich Kohler and Magdalena Luniaik, he created the Stata module for sequence analysis. His research interests are labour market transitions, educational inequality and welfare state impacts.

François Buton is a researcher at the National Center for Scientific Research (CNRS, France). Among his publications are *L'administration des faveurs. L'État, les sourds et les aveugles, 1789–1885* (2009), *Sociologie des combattants* (ed.), Pôle Sud 36 (2012), “Instituer la police des risques sanitaires. Mise en circulation de l'épidémiologie appliquée et agencification de l'État sanitaire”, *Gouvernement et action publique* 4 (2012) 67–90 (with Frédéric Pierru), and, in English, “The Household Effect on Electoral Participation. A Contextual Analysis of Voter Signatures from a French Polling Station (1982–2007)” (with Claire Lemercier and Nicolas Mariot), *Electoral Studies*, 31 (2012) 434–447. His web page: <http://www.cepef.univ-montp1.fr/spip.php?article66>.

Felix Bühlmann studied sociology and political science at the universities of Geneva, Berlin, Lausanne and Manchester. He is Assistant Professor of Life Course Sociology at the University of Lausanne, Switzerland. His research interests include the life course, economic sociology and political sociology. Economic, political and economic elites, as well as trajectories of vulnerability and precariousness are currently his main interests. He has published in the *European Sociological Review*, *Economy and Society* and *Actes de la recherche en sciences sociales*.

Denis Colombi is currently pursuing a PhD research on international mobility and labor markets transformation at the Centre de Sociologie des Organisations (Sciences-Po, Paris) after his graduation at the École des hautes études en sciences sociales (EHESS, Paris). He studies skilled migrants' careers and compare them to low-skilled migrants and international elites, as way to link social classes and economic sociology. His main topics of research are economic sociology, careers studies and globalization.

Julia Dietrich is a research associate (equivalent to assistant professor) at the Friedrich-Schiller-University of Jena, Germany. Being a developmental psychologist, she also collaborates with colleagues in educational and social psychology, and economics. Her current research focuses on educational transitions, especially the transition from school to employment or higher education, and on young people's motivation and goals. She is also interested in methods for the analysis of change and the assessment of context effects on young people's development.

Cees H. Elzinga In 2002, inspired by Wil Dijkstra, Cees H. Elzinga (1950) got involved in Sequence Analysis and Social Demography when he became an assistant professor in the Dept. of Social Science Research Methods of the VU University in Amsterdam. Today he is a Full Professor in Pattern Recognition, Vice-Dean and Educational Program Manager of the Faculty of Social Sciences of that university. Over the last decade, he published on theory and applications of sequence analysis in *Sociological Methods & Research*, in Demography and in the *European*

Journal of Population. He also published on kernel algorithms in *Information Sciences*, in *Theoretical Computer Science* and in *Pattern Recognition Letters*.

Anette Eva Fasang is assistant professor of Sociology at Humboldt-University Berlin and head of the project group *Demography and Inequality* the Social Science Research Center Berlin (WZB). She obtained her Ph.D. from Jacobs University Bremen and did postdoctoral research at Yale University and Columbia University. Her research interests include social demography, stratification, life course sociology, family demography, and methods for longitudinal data analysis.

Jacques-Antoine Gauthier is a senior lecturer at the Life course and inequality research center at the University of Lausanne. His current researches focus on modeling and analyzing longitudinal data on life courses. The central issue of his researches concern the time related construction of individual life trajectories—notably when experiencing major transitions such as conjugality and parenthood. A second field of research focuses on the range of application of sequence analyses (costs and multidimensionality issues) and on the visualization of their results.

Brendan Halpin is a lecturer in sociology at the University of Limerick, Ireland. His research has involved lifecourse data and longitudinal methods since his doctoral studies, and he has spent many years working with the British Household Panel Survey. Current research interests include labour market dynamics, family formation and dissolution, education and educational homogamy, and imputation for categorical time-series data.

Shin-Kap Han is Professor of Sociology and Director of Social Science Library/Social Science Information Center at Seoul National University in Korea. Social organization, broadly defined, is his area of research and teaching, which includes social networks, social stratification, organizations and institutions. His current research focuses on, among others, analyzing collective action on various scales, both historical and contemporary; integrating cultural and textual materials into sociological analysis; and developing methodological tools for structural analysis of social space and social time. (Homepage: <http://sociology.snu.ac.kr/skhan/>)

Dominique Joye is Professor of Sociology at the University of Lausanne. He works on survey methodology, in particular in a comparative context (ISSP, EVS, etc.) but also on social inequalities and life course studies in a Swiss context.

Claire Lemercier is a research professor (directrice de recherche) of history at the French National Center for Scientific Research (CNRS). She works along with sociologists and political scientists in the Center for the Sociology of Organizations (CSO) in Sciences Po, Paris. Originally a specialist of the relationships between the State, firms and entrepreneurs in nineteenth-century France, she is now expanding her investigations to the twentieth century, England and the USA. She also teaches and uses various formal methods, prominently network analysis and sequence analysis.

Laurent Lesnard is a CNRS associate professor of sociology at Sciences Po (Paris, France). He is the director of the Centre for Socio-Political Data (Sciences

Po & CNRS) and of the equipment Data, Infrastructures and Survey Methods for the Social Sciences (Sciences Po) which is funded by a “Équipement d’excellence” grant (2011–2019). His research is in the fields of sociology of time and of sequence analysis. He studies questions of individual and couple working hours and the transformation of social ties. In 2011, Laurent Lesnard was awarded the Bronze medal of the National Centre for Scientific Research (CNRS) that heralds his arrival as a talented new specialist in his field.

Nicolas Mariot is a senior researcher at the National Center for Scientific Research (CNRS, France). Among his publications are *Bains de foule. Les voyages présidentiels en province, 1888–2002* (2006), *Face à la persécution. 991 Juifs dans la guerre* (with Claire Zalc 2010), *Tous unis dans la tranchée? 1914–1918, les intellectuels rencontrent le peuple* (All united in the trench? 1914–1918, how intellectuals met the people, 2013) and, in English, Does acclamation equal agreement? Rethinking collective effervescence through the case of the Presidential tour de France during the 20th century, *Theory & Society*, 40-2, march 2011: 191–221. His website: <http://www.jourdan.ens.fr/~mariot/>.

Pierre Mercklé is a sociologist. He is an Associate Professor at the Ecole Normale Supérieure (ENS) de Lyon (France), where he is currently teaching social sciences and quantitative methods. Since 1995, he has been conducting research within the “Cultures, Dispositions, Pouvoirs, Socialisation” (DPCS) team at the Centre Max Weber (CNRS UMR 5283, Lyon). Pierre Mercklé’s research interests range from sociology of culture and social network analysis to history of social sciences. Among his recent publications: *L’enfance des loisirs* (La Documentation française 2010), and *Sociologie des réseaux sociaux* (La Découverte 2011).

Michel Oris, PhD in History of the University of Liège, is professor at the Faculty of Social and Economic Sciences of the University of Geneva since March 2000. Codirector of the NCCR LIVES (Overcoming Vulnerability. Life Course Perspectives), director of the Centre for the Interdisciplinary Study of Gerontology and Vulnerability and of the Institute of Socioeconomics, president of the Société de Démographie Historique and the Association Internationale des Démographes de Langue Française, his research interests are related to the living conditions of the elderly, the interactions between individual life trajectories and the social structures, vulnerabilities and the life course.

Simon Paye has recently completed a PhD research on the transformation of academic work and careers in the UK. He has since joined the Collège de France as temporary lecturer, where he intends to apply new methods to fundamental topics of sociology of work: working times, boundaries between work, “leisure” and other social times, boundaries between salaried and independent/illegal workforce, etc. He is also involved in a collective endeavour geared to take stock of current approaches towards the engineering profession (edited volume forthcoming: “*La production de l’ingénieur. Contributions à l’histoire sociale d’une catégorie professionnelles*”).

Alexandre Pollien is a sociologist and researcher at the Swiss Centre of Expertise in the Social Sciences (FORS). His research focuses on non-response in surveys. He also works in sociology of the life course and published articles on career, family transitions and training itineraries.

Gilbert Ritschard, PhD in Econometrics, is Professor of Statistics for the social sciences and Vice-dean of the Faculty of Economics and Social Sciences at the University of Geneva. He has carried out teaching and research in data analysis and statistical modeling for social sciences, including Longitudinal data analysis, Event history analysis, and Sequence analysis. His recent applied contributions concern life course studies. He currently leads a methodological project within the Swiss NCCR “LIVES: Overcoming vulnerability, a life course perspective,” and developed with his team the now worldwide used TraMineR toolbox for the exploration of state and event sequence data.

Katariina Salmela-Aro is professor in psychology at the University of Jyväskylä, Finland and visiting professor in the Institute of Education, University of London, UK. Her research has involved life-span approach using longitudinal methods. Current research interest include life-span model of motivation during critical life transitions, engagement, burnout and related interventions.

Matthew Charles Wilson As the first member of his family to earn a college degree, Matthew Wilson pursues graduate-level research in the Department of Political Science at the Pennsylvania State University. He is a National Science Foundation Graduate Research Fellow. His research concerns the interactions of leaders and institutions, and their relation to conflict outcomes. He is particularly interested in methods of authoritarian rule. His focus on regimes harkens to the literature on democratic transition and consolidation. As a comparativist, he has a regional interest in Latin America. He is exploring ways to develop sequence analysis to enable theory testing on qualitatively distinct institutional sequences.

Claire Zalc is a research fellow in history at the Institut d’histoire moderne et contemporaine (CNRS-ENS). She specializes on the history of immigration in twentieth-century France, the history of business and business owners, the history of work and labor, and the history of Jews in France during World War Two. She is currently teaching quantitative methods and published, with Claire Lemercier, Méthodes quantitatives pour l'historien (La Découverte 2008). She is the author of Melting Shops. Une histoire des commerçants étrangers en France (Perrin 2010) and the coauthor (with Nicolas Mariot) of *Face à la persécution. 991 Juifs dans la guerre* (Odile Jacob 2010).

Chapter 1

Introduction: Sequence Analysis in 2014

Jacques-Antoine Gauthier, Felix Bühlmann and Philippe Blanchard

Sequences for Social Sciences

Sequence analysis (SA), the statistical study of successions of states or events, is one of the promising venues of sociological methodology. Since its introduction in the social sciences in the mid-1980s it has developed steadily and spread to many social science disciplines. It is now central to the life-course perspective where it has been used to understand trajectories of cohabitation and housing and of occupational careers or crucial transitions, such as from school to work or from employment to retirement. More recently SA has also been employed to shed light on chains of historical events, on historical evolution of political institutions and on historical life courses. The method has also been employed to investigate other temporal scales, such as sequences of daily time use or response patterns in surveys.

In a nutshell, SA models processes. It compares chronological sequences of states within a holistic conceptual model instead of observing allegedly independent observations over time. SA accounts both for individual and structural dynamics.

This publication benefited from the support of the Swiss National Centre of Competence in Research “LIVES—Overcoming vulnerability: life course perspectives”, which is financed by the Swiss National Science Foundation. The authors are grateful to Laure Bonnevie for reviewing the manuscript.

J.-A. Gauthier (✉)

Institute of Social Sciences, Life Course and Inequality Research Centre,
University of Lausanne, Lausanne, Switzerland
e-mail: Jacques-Antoine.Gauthier@unil.ch

F. Bühlmann

Institute of Social Sciences, National Centre of Competence in Research LIVES,
University of Lausanne, Lausanne, Switzerland
e-mail: felix.buhlmann@unil.ch

Ph. Blanchard

Institute of Political and International Studies, University of Lausanne, Lausanne, Switzerland
e-mail: philippe.blanchard@unil.ch

Each individual displays a unique trajectory, defined as a string of states of specific nature, with specific durations and a specific order. At the same time sequences resemble each other along common subsequences showing how individuals' trajectories are contingent on social structures.

Sequence analysis has the potential to fundamentally enhance our understanding of a broad range of social processes. Partly for methodological reasons, phenomena such as family configurations, employment patterns, poverty and precariousness, political engagement or social inequalities are traditionally analysed with synchronic models. This leads to a static understanding of these phenomena that prevents researchers from understanding (a) the real temporal nature of these phenomena and (b) how these are constructed through interdependent biographical processes. This also has practical consequences: as long as a phenomenon such as poverty is analysed using synchronic data, one fails to understand intra-individual variability. Why do some people fall into poverty and others escape from it? What typical previous biographical patterns lead to such an event, and what other patterns do not? SA—along with methods such as event history analysis or structural equation models—expands our understanding of biographical trajectories and has a role to play among major, current and substantial debates in the social sciences.

This book intends to promote a broad and systematic debate that takes stock of the advances and the specific range of application of SA, that encourages a careful standardisation of the approach beyond diverging orientations, and that discovers and explores new methodological paths and combinations. The project was initiated in the aftermath of the Lausanne Conference on Sequence Analysis (LaCOSA), in June 2012, which brought together scholars from Finland, France, Germany, Italy, South Korea, the Netherlands, Switzerland, the United Kingdom, and the United States¹. The conference revealed the international spreading of the method, and the need for concertation and harmonisation beyond national clusters of sequence analysts and some inter-cluster connections. Indeed, sociologists of the life course, historians of the family, political scientists, comparativists and social statisticians had to share beyond what could be done by publishing separately in specific disciplinary journals. Distinct approaches needed to be compared, good practices needed to be established and further developments needed to be based on collectively approved directions. Finally, the conference intended to connect the sequence approach with other, more traditional approaches to longitudinal data, which is one way SA can gain greater popularity.

The book addresses several audiences: experienced SA users looking for up-to-date developments by renowned, as well as innovative, younger scholars; new SA users searching for a larger view on the diverse ways of applying the method; and sociologists, historians, political scientists and specialists of other social sciences using longitudinal data and interested in new ways of dealing with them.

¹ The authors wish to thank the Swiss National Science Foundation (NSF), and the following units of the University of Lausanne: the Institute for Social Sciences (ISS), the Institute for Political and International Studies (IEPI), the Research Center on Political Action of the University of Lausanne (CRAPUL) and the Foundation of the 450th for their financial support to the LaCOSA conference.

Accordingly, it may be read either as a whole, by sections or by chapters. In the remainder of this introduction, we first outline some epistemological and theoretical principles that are shared by many sequence analysts. Emphasis is put on the developments of the method by specifically using concepts from social science rather than purely formal and technical terms. Second, we trace the main developments in SA over the last twenty years. We argue that due to the specific structure of the scientific field of SA, one can distinguish a relatively consensual core program and a series of innovative, emerging alternatives. Finally, the last part will present the main debates which have animated the SA communities in the last years and give the reader an overview of how these debates are reflected in the book.

What is a Sequence?

Many theories in the social sciences hypothesise historical changes and place a special emphasis on processes, orders, transitions and developments. The centrality of both institutional and biographical order and processes was particularly important in the Chicago School of sociology, which constitutes the theoretical starting point for SA in this field. Howard Becker (1963) pleaded for sequential models of social reality. In his sequential theory of deviance, he insists that explanatory factors (causes) are not simultaneously influential. A social process is divided into phases that have their own specific explanations. The final behaviour can only be explained by the chronological succession of these phases. From this perspective, the various elements of such narratives are bound not only to the immediate, but also to the remote past. Explicitly paying tribute to this school of thought, the sequence theorist Abbott (2001) claims that “social reality happens in stories”, which he calls *narratives*. Considering social processes as “stories” does not mean that they are free from constraints or predetermined. They unfold in a web of constraints and opportunities.

Most narratives may be formally described with four concepts: trajectory, stage, event, and transition (Levy et al. 2005). A *trajectory* may be seen as a model of stability and long-term change (George 1993) or as a sequence of participation profiles (Levy 1991) that represents the variation over time of the different roles an individual holds in various life spheres. It is used to describe the structure of change and stability characteristics of an individual life, in one or more of its dimensions. A *stage* refers to a life period of relative structural or functional stability (family life stages, child development stages, periods of economic growth, etc.). A *transition* means a change between two stages (for example, role and structural changes during the transition into parenthood or during a political revolution). An *event* is what occurs at a given time in a given place and is something to which human beings attribute meaning. Events can be normative (e.g. birth, marriage, graduation) or non-normative (e.g. disease, war, death). Important events are often associated with a transition.

Narratives are structured patterns that have a beginning, a middle, and an end. Given the large number of possible expressions of a specific social phenomenon, looking for ideal types is a convenient solution to apprehend it as a narrative. Indeed, social processes contain typical events that unfold in a typical order. This order is influenced by factors such as personality traits, family configuration, occupation type, health status, or institutional power relationships. Narratives are therefore, in most cases, shaped both by endogenous and exogenous factors. The relative weight of these two structuring factors is variable. For instance, the current health status of an individual is very much structured by internal, (epi)genetical factors, whereas her/his current residential situation may vary according to external factors, such as the dynamics of the housing market and its links to local economical and/or political constraints. To be relevant, however, narratives must have a unity and a coherence that is analogous to that of variables. This unity of narratives forms the basis for generalisation. This is precisely why, once generalised, a narrative may become a variable again, for example to be fed into a regression analysis (e.g., Levy et al. 2006). According to Hull (1989), narratives form coherent units because the events that build them concern a central subject (an individual, a couple, an occupation, an organisation, or a corpus of texts). He proposes another way to describe the exogenous and endogenous structure of narratives, as for him, a central subject must be seen in two ways. On the one hand there is a linkage between the events entering the narrative (internal structuring dynamic). On the other hand, one has to consider the whole relationship which links the central subject to its surrounding—for example, historical, political, organisational—narratives. In this perspective, the elements of a narrative are defined by their function in the narrative line, considered as a whole. These elements are therefore not independent from the context. A systemic perspective—taking the macro-micro link into account—is therefore central to an understanding of social narratives.

Abbott (2001, Chap. 4) differentiates three levels of complexity for sequential models with respect to the recurring or non-recurring nature of events and according to the number of dimensions taken into account:

“Stage theories” or “natural histories” describe biographical or historical stages through which most of the units of observation seem to pass. They obey an internal process logic, relatively independent from contextual influences (for example, child developmental stages, compulsory education in highly standardised countries, legislative or public agenda processes).

“Careers” show more variability than natural histories. They are dependent on the context, more contingent, and also include recurrent events and states. They typically include trajectories of education, occupational careers, residential trajectories, or activist careers.

At the more complex level of “Interactional fields”, a whole network of interdependent sequences belongs to various relevant dimensions that form a system (family, occupation, residence, health, welfare regime, political stability, etc.).

The Core Program of Sequence Analysis

The research trend in sequence analysis was structured in the 1990s and 2000s around a core program with a limited number of methodological options. This core program studies life trajectories of individuals living in a contemporary period. It mostly refers to sequences of equal length, with age as a time axis and year as a time unit. Sequence samples comprise hundreds to thousands of individuals. The data processing runs from coding to sequence formatting, to sequence comparison usually by means of optimal matching analysis (OMA), followed by clustering to build typologies of sequences. Typical representatives of the core program are Blair-Loy (1999), about US female executive careers in finance; Halpin and Chan (1998), about professional mobility in Ireland and the United Kingdom; Pollock (2007), about joint employment, housing and family careers in British households; Bühlmann (2008) about the careers of engineers and managers; Robette (2010), about pathways to adulthood in France; or Stovel et al. (1996), about the long-term transformation of banking careers. The next section summarises the main steps of sequence comparison as developed in the above-cited studies.

Sequence Comparison in a Nutshell²

Sequence analysis can be applied to any type of time-related phenomenon. As these phenomena are often multidimensional, the considered processes must first be decomposed in distinct dimensions or fields. The various positions that characterise a given dimension must be reduced to a finite, non-ambiguous number of states called an alphabet. An individual sequence in this perspective is the succession of the observed states for one unit of observation over a given time period. There are several ways to compare sequences. In the following we will describe the one usually employed in the core program: the optimal matching analysis (Kruskal 1983). Let us consider one possible way to model individual occupational status(es) in a sequential perspective. We first need to define a finite set of such statuses, as for instance $O = \{G = \text{education}, P = \text{part time employed}, F = \text{full time employed}, H = \text{at home}\}$. We need longitudinal data in order to attribute a specific status to each time unit (e.g., years) for a given time frame, e.g. between the ages of 16 and 35 (Table 1.1).

Having done this, we need to know how (dis-)similar these sequences of statuses are. This is done by estimating the number of elementary operations of substitution, insertion or deletion (generically called *indel*), which are necessary to transform one sequence (source) into another one (target). Each elementary operation is associated with a given cost, that can be unitary, theory driven or data based (Gauthier et al 2009). The number of elementary operations that are necessary to perform the

² All computations presented in this example are made using the R statistical environment (R Development Core Team 2011) and the associated TraMineR package for the sequence comparison (Gabadinho et al. 2011).

Table 1.1 A set of four sequences of length 20, built with four symbols (F, G, H and P)

ID	Age	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1		G	G	G	G	G	F	F	F	F	F	P	P	H	H	H	H	H	H	H	
2		G	G	F	F	F	F	F	F	F	F	F	F	F	P	F	F	F	P	P	
3		G	G	G	G	G	G	G	F	F	F	F	F	F	P	P	P	P	P	P	
4		G	G	G	G	F	F	F	F	H	H	H	H	H	H	H	H	H	H	H	

Table 1.2 The substitution costs matrix

	G->	F->	H->	P->
G->	0	2	2	2
F->	2	0	2	3
H->	2	2	0	2
P->	2	3	2	0

Indel cost=1

transformation is called the distance between the two sequences; the greater the distance, the more dissimilar the sequences. As there are many different transformation strategies to make this transformation, the recursive dynamic programming algorithm (Levenshtein 1966) enables one to find the optimal matching between two sequences—that is, the particular combination of elementary operations needed to transform one into the other at a minimal cost. For sequences of equal lengths, when substitution costs are unitary and indel cost values are set to half of that of a substitution, the resulting distance is equivalent if one uses a substitution or an insertion followed by a deletion. If sequences are of different lengths, *indels* fill the gaps and distance normalisation strategies may be considered. In the following example we compute a user-defined substitution cost matrix in which all costs are set to 2, except the one between symbols F and P, which is set to 3. This means that in certain cases, insertion-deletion is less costly than substitution (Table 1.2).

The resulting alignment (Table 1.3) needed four substitutions and twelve indels to transform the sequence 1 into the sequence 2, hence the (optimal distance between them is $(4*2 + 12*1) = 20$).³

We then compute all the pairwise distances between sequences and gather them in a distance matrix (Table 1.4).

The next step is to group the sequences at hand (e.g. by means of a cluster analysis) in order to build typologies of sequences by reducing the complexity of the data. A convenient way to represent the grouping process is to display a dendrogram (Fig. 1.1) of Ward's hierarchical ascending clustering procedure (Ward 1963).

The dendrogram shows the greater proximity of sequences 1 and 4 compared to 2 and 3. In this simple example, by cutting the tree above the first branching and below the second one, we can create two groups of structurally homogeneous sequences.

³ Note that the use of indel costs changes the length of the alignment (that is, the temporality) while keeping its structure, whereas using substitutions only keeps the temporality, but changes the structure of the sequences.

Table 1.3 The optimal alignment

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Seq 1	G	-	-	-	G	G	G	G	F	F	F	F	F	F	P	-	-	-	P	H	H	H	H	H	H	
Trans.	E	E	I	I	S	S	S	S	E	E	E	E	E	E	I	I	I	I	E	D	D	D	D	D	D	
Costs	0	0	1	1	2	2	2	2	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	
Seq 2	G	F	F	F	F	F	F	F	F	F	F	F	F	F	P	F	F	F	P	F	F	P	-	-	-	

Trans. = transformation, where E = equivalent, S = substitution, I = insertion, D = deletion

Table 1.4 The matrix of pairwise distances between sequences

	Seq1	Seq2	Seq3	Seq4
Seq1	0	20	12	10
Seq2	20	0	18	26
Seq3	12	18	0	22
Seq4	10	26	22	0

Fig. 1.1 The clustering step: Dendrogram of the Ward's agglomerative procedure

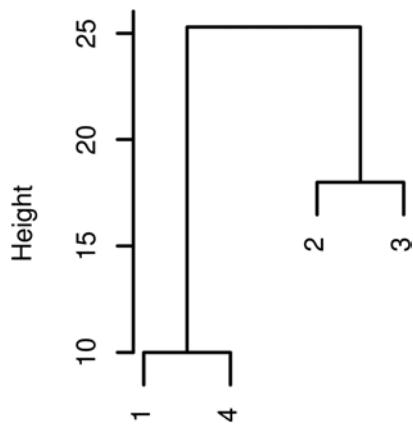


Table 1.5 Assignment of sample units to typical group of sequences

ID	Cluster membership
1	1
2	2
3	2
4	1

This reduction of complexity allows us to create a categorical variable that assigns each sample unit to a specific, typical group of sequences (Table 1.5). Eventually, in order to contextualise the individual sequences, this variable may be used in further statistical analyses, as for instance logistic regression or multiple correspondence analyses.

The Core Program: How did we End Up Here?

One may wonder how the standard options of the core program emerged. They result from a process of convergence and selection between research programs led by scholars from different disciplines and different countries. This process was not managed or coordinated by some central individual or team. The one who evidently could have played such a role, Abbott, left the field at the end of the 1990s without formally proposing a specific guideline to follow. His book *Time Matters* (Abbott

2001), in the shape of a testimony of his work on sequences, deals with the limits of standard methods used to treat longitudinal data in the social sciences and what objectives social sciences should seek to attain in order to formalise social life as a narrative or sequence. His expectation was that from his own empirical experiments on sequences, diverse techniques and tools would burgeon, compete with each other, and eventually contribute to improving our understanding of this new object (Demazière et al. 2011). For the time being, the collective aggregation and convergence of efforts around a standard line in social sciences resulted from a series of factors (Blanchard Forthcoming 2014):

1. The dominance of Abbott's legacy from his methodological and empirical papers—still one of the most cited papers in the field;
2. A theoretical need for statistical sequential approaches that had to be fulfilled empirically;
3. The strength of applications in the sociology of work and employment and their specific expectations in terms of sequences;
4. The availability of good-quality sequential datasets based on surveys, whether retrospective questionnaires or panels;
5. The modest number of software implementations, which gave a decisive advantage to those applications first to be developed for the social sciences, namely *Optimize* (Abbott and Prellwitz 1997) and *tda* (Rohwer and Poetter 2007), then the *TraMineR* package in R (Gabadinho et al. 2011);
6. The limited number of sequence analysts who are facing a much larger population of sociologists who are using traditional methods, which compelled the former to stick to basics and promote a minimum core program so as to meet the journals' reviewers' requirements—chosen mainly among the latter.

All these factors reinforce each other and explain the present state of SA in social sciences: cumulativity, yet limited to some kinds of data, research questions, research designs, and treatment strategies. At this stage no exhaustive recipe book is available, but the (implicit) consensus around a set of core options can be found in several venues. The beginner may refer to Abbott and his colleagues' introductory articles (Abbott 1995; Abbott and Tsay 2000; Macindoe and Abbott 2004), to two rather short textbooks (Aisenbrey 2000; Robette 2011), and to a few papers partly following some pedagogic purpose (Lesnard 2006; Blanchard 2011; Gauthier 2013). The most comprehensive overview on the methodological advances of SA in the last decade can be found in Aisenbrey and Fasang (2010). Social and political science courses within academic method schools are also proposed in Germany (Berlin, 2007), the United Kingdom (Essex, 2007; Manchester, 2009), France (Lille, 2009, 2010, 2011), Switzerland (Geneva, 2009, 2012), Austria (ECPR-Vienna, 2012), the United States (Columbia, 2011) and Canada (Ontario 2012). The core program is progressively establishing itself, although at the present stage of development, the (uncoordinated) priority is put on the development of new lines and variants.

Variants to the Core Program

The provocative question “What Have You Done for Us Lately?” addressed by Joel Levine to Abbott in 2000 can be enlarged: What have we done in the last 14 years? What alternatives have been proposed to the core program? Which ones have also been tested empirically? The following is a short review of sequence data, methods for the calculation of dissimilarity between sequences and methods used to process and interpret dissimilarities.

First, data may diverge from the standard frame. They may relate to non-contemporary individuals, as did Oris and Ritschard’s transposing life course interrogations to Nineteenth-Century East Belgium (see Chap. 8). They may not deal with human individual data, such as Stovel’s (2001) analysing sequences of local lynching events in the Deep South (1882–1930) as marks of micro-level narratives of political violence. Data may use time units other than years or months; for example 10-minute slots in time-budgets may be used (Lesnard 2008). They may not relate directly to standard clock time, as with Buton et al. (see Chap. 10), who consider sequences of votes from ballots with irregular timing over 25 years. Data may also encompass small-N populations, which is the case with Abbott and DeViney’s (1992) study of welfare development in 18 industrialised countries. Sequence lengths may also be unequal, such as life-long AIDS activists’ careers with varied ages in Fil lieule and Blanchard’s study (2012). Some of these alternative data bring new challenges to the standard approach, regarding data formatting and treatment, as well as the interpretation of results. For example non-standard time measurement creates specific event sequences that require specific tools; small-N populations change the way the results are generalised to a larger population (as long as the sample is not the full population, as in Abbott and DeViney 1992); and sequences of uneven lengths affect the interpretation of elementary operations applied by OM, especially in terms of balance between indel costs and substitution costs.

A second field of variants comes with the way sequences are processed. It may take quite different directions from the one privileged today in the majority of studies. Much has been written about the possible variations regarding costs, including empirically machine-trained substitution costs (Gauthier et al. 2009), indel costs varying with states, substitution costs varying with data-based time (Lesnard 2010), or indel costs varying with the state’s neighbourhood in the sequence (Hollister 2009), although no consensual doctrine has yet emerged on costs. Less attention has been paid to other variations. For example, instead of incremental, state-by-state comparison, as in OM, sequence comparison may use holistic criteria based on the search for chunks of common subsequences. One may use the longest common subsequence (actually a variation of OMA with lower indel costs), or the longest common prefix (Elzinga 2006). One may also compare without aligning, either by introducing alternative sets of elementary operations, such as swaps, by comparing

pairs of ordered elements, or by applying a correspondence analysis to the matrix of time spent by period in each state (e.g., Deville and Saporta 1980). These different approaches rely on diverging conceptions of sequences and give privilege to different mathematical tools. As a result, they put the accent on distinct aspects of social time, as Robette and Bry (2012) do in their synthetic comparison of sequence metrics with simulated and real datasets. The strategy of comparison itself may not rely on complete pairwise OM but on a comparison with ex ante ideal-typical sequences (Wiggins et al. 2007). The lengths of sequences may be standardised (Abbott and Hrycak 1990, and the dissimilarity matrix normalised (Han and Moen 1999), in both cases with noticeable consequences on how the results are interpreted. One may even imagine an optimal path inside the OM pairwise comparison array that would be different from simply the cheapest one resulting from the Needleman and Wunsch (1970) algorithm. This would lead again to a different pairwise dissimilarity structure.

Third, the choice of methods associated with the dissimilarity calculation also varies notably, as well as the way they are combined. These complementary methods may take diverse roles relative to the SA core program, as described above. They may be more or less autonomous from SA, previously developed in other fields, or reversely specific to SA; they may be used at various stages of the treatment process, either before or after the calculation of pairwise dissimilarities (in case any dissimilarity is calculated). For example, measures of sequences' complexity are imported from the theory of entropy, then applied to algorithmic information theory and molecules comparison, and later adapted to social sequences (Elzinga 2010). They can be used at the exploration stage or after some OM-based clusters have been distinguished. Similarly, multidimensional scaling, not specific to sequences either, is useful to sort out both sequence populations and clusters (Piccarreta and Lior 2010). Validity and sensitivity analyses (e.g., Hollister 2009; Aisenbrey and Fasang 2010) as well as simulation experiments (Bison 2009; Robette and Bry 2012), at the end of the treatment process, can tremendously improve the robustness of OMA and SA.

As Aisenbrey and Fasang show (2010), waves of innovation have run through the sequence community (if we may call it so) over the last 15 years. Yet, looking back, it appears that many of the alternative data and methods had been explored or at least suggested by the works of Abbott and his colleagues, on empirical (Abbott and Barman 1997; Abbott and DeViney 1992; Abbott and Forrest 1986; Abbott and Hrycak 1990) as well as methodological aspects (Abbott 1995; Abbott and Tsay 2000). Sequence analysts have been right to give more concrete and systematic insights into what was often left at the stage of intuitions by the pioneer. What global picture emerges from these variants? What federating concepts can bring unity back to the picture? This is what this book intends to bring answers to. Let us sketch some of them in the last part of this introduction.

Some Currently Debated Issues—the Chapters of this Book

The contributions of this book address central current debates in the field of SA. A particular emphasis was put on contributions' using new and unusual data, applying the method to phenomena beyond the mainstream life course research and on an extension of the disciplinary scope of SA. A few chapters rely on previously published fieldwork, but they either extend the empirical field further, or propose new methodological options that change the results. The book is organised in four sections: (a) How to Compare Sequences, (b) Life Course Sequences, (c) Historical and Political Sequences, (d) Visualisation of Sequences and their Use for Survey Research.

How to Compare Sequences

Sequence comparison is core business for SA. It also connects with the way sequences are generated and with their structural characteristics. What do sequences mean and how should they be treated from a social science perspective? The first part includes reflections about the genesis and the explanation of sequences, about the definition of what a meaningful sequence is and about the mathematical premises of comparison. The chapters of the first part examine sequence comparison from three different angles: Shin-Kap Han (see Chap. 2) in his chapter “Motif of Sequence, Motif *in* Sequence” tries to identify motifs as dominant and salient elements of sequences. He asks how and why we might be able to identify a part of a sequence that can effectively represent the sequence as a whole. Borrowing from music theory, social network analysis and molecular biology, he explores possible futures of the method and shows how Abbott’s work can be carried on. Laurent Lesnard (see Chap. 3) proposes a deepening of the theoretical anchoring of SA. Challenging some of the technical solutions of cost setting that have been developed in the last years, he urges SA to return to sociological theories of time and to set costs according to the sociological meaning of time, as developed by Durkheim, Elias and Bourdieu. Complementary to this theoretical reinforcing of sequence comparison, Cees Elzinga’s chapter (see Chap. 4) clarifies the mathematical foundations of sequence comparison. He discusses the axiomatic of sequence distances, both OM-based and subsequence-based, before articulating distance and similarity and showing that the two concepts are not exactly symmetrical and may lead to diverging partitions for instance. Brendan Halpin (see Chap. 5) investigates different narratives and measures of similarity and dissimilarity between sequences, comparing the OM algorithm, a duration-weighted combinatorial subsequence algorithm, with a newly developed time-wrapping algorithm. It appears that the time-wrapping algorithm can be parameterised in a way that provides a bridge between algorithms such as OM (for which time is a flexible scale) and combinatorial subsequence algorithms (for which time is a scale-less order). The time-wrapping algorithm therefore has good chances for an empirical link between different “narratives” of similarity, which to date have been too distant from each other within the SA universe.

Sequence Analysis and the Life Course

The second section is located within the life course paradigm, the most fruitful paradigm for SA for the last fifteen years. As the large number of applications of SA on biographical process has shown, life course trajectories—occupational careers, residential trajectories or family trajectories—often feature structures of timing and order that are particularly suited to be understood by SA. Nevertheless, this line of work is not exhausted, as demonstrated by these three chapters. Anette Fasang (see Chap. 6) attempts to more seriously take into account the initial theoretical promises of SA on the non-linearity of processes. She identifies three important domains where SA could make a major contribution: multidimensional lives, linked lives and destandardisation. She shows for instance how Elder's (1985) idea of linked lives can be enriched by a cross-fertilisation with recent trends in network analysis and an emphasis on dyadic relationships among siblings, parents and partners. In summary, she argues—both with and against Abbott—for a more theoretically informed and deductive use of SA and refuses to pigeonhole it as an exclusively explorative method. In their chapter Julia Dietrich, Håkan Andersson and Katariina Salmela-Aro (Chap. 7) explore how SA can be used in a discipline that to date has not been very receptive towards sequence analytical approaches: psychology. This neglect is far from self-evident. To the contrary, developmental psychology and life-span psychology are considered to be important elements of an interdisciplinary approach to life course research. Hence, Dietrich et al. explicitly use concepts from a developmental psychological approach to study educational trajectories and to investigate how these trajectories are related to psychological resources and career goals. They show that the links between trajectories and mental dispositions—such as beliefs, attitudes, personality attributes and so on—are still seldom examined. We can learn from their chapter that psychology (in particular developmental psychology) is a promising disciplinary area of further application of SA in the future. Finally, Michel Oris and Gilbert Ritschard (see Chap. 8) employ SA to enhance our historical knowledge about family formation. They are able to show that in the first half of the 19th century, among peasants and daily labourers from Ardennes and the Pays de Herve, despite a general high age at the moment of marriage, one can observe an impressive variety of trajectories. In particular, they find evidence for a group with very late motherhood—a discovery which questions and nuances the mainstream explanations to history of family formation in Europe.

Political Sequence

The third section opens up the empirical perspectives beyond contemporary life course to historical and political phenomena. On the one hand, it becomes a part of the prosopographical approaches in history. When studying historical periods by means of collective biographies of specific birth cohorts or particular social groups, historians have begun to employ SA in order to understand types of careers and trajectories. On the other hand, the sequential chains of historical events themselves

are being studied. Following the pioneering study of Katherine Stovel on sequences of lynching in the US south (2001), scholars are trying to understand institutional change through SA.

In this part of the book, we present two contributions from the first type and one from the second type: Pierre Merklé and Claire Zalc (see Chap. 9), in an uncommon contribution to Holocaust research, investigate trajectories of persecution of a group of Jewish inhabitants of the French town of Lens. Their micro-historical approach of the Holocaust sheds new light on the trajectories and patterns of persecution, and they also raise a series of important methodological and ethical issues—related to the very particular factors which gave form to these trajectories. François Buton, Claire Lemercier and Nicolas Mariot (see Chap. 10) present another innovative application of SA: they investigate electoral participation on the basis of a signature list of a single French polling station—during more than 25 years and over 44 ballots. This yields in a dynamic account of electoral participation, allows the researchers to understand electorate households and thus poses intriguing questions to traditional, static measures of voting behaviour. Matthew Wilson (see Chap. 11) shows how superficially political scientists usually treat sequence phenomena. Regarding order and timing, existing theories were more speculative than empirical. SA can contribute a lot to explain the much debated democratisation processes. Wilson takes a macroscopic perspective, both geographically and historically, following Abbott and DeViney (1992) on welfare state regimes. Taking into account the order of the steps of democratisation, SA improves the classification of political regimes.

Visualisation of Sequences and Their Use for Survey Research

Visualisation issues have a long tradition and are still debated. Synoptic capacity of visualisation has an enormous impact on the diffusion of research findings, especially if these findings are based on innovative analytical approaches (Tufte 1997). However, as stated above, SA procedures still lack a consensus on how to best represent sequences. The fact that an increasing number of large longitudinal data sets are now available to researchers in social sciences may boost the use of SA and therefore the need for adequate and effective ways to think graphically about sequences and communicate the results. Thus, there is a need to evaluate the possible mapping strategies that may apply to results stemming from SA, their respective strengths and weaknesses, and their specific range of application.

Chapters in the fourth section discuss visualisation from different angles. In an innovative contribution, Ivano Bison (see Chap. 12) proposes a morphologic approach to individual careers by converting sets of individual sequences into network graphs. Based on the occupational careers of women and men, this formal exploration aims at uncovering to-date hidden configurational patterns by a complete change of data visualisation perspective. This analytical shift provides a new structural basis for interpreting individual and collective life course dynamics. In their contribution, Denis Colombi and Simon Paye (see Chap. 13) emphasise the

importance of considering alternative time references and therefore focus on the links existing between events and trajectories. To this aim, individual sequences are synchronised not according to age, but to some important common event, which can be endogenous or exogenous to the considered sequences. In a study of academic careers, they show how such an event-based synchronisation of trajectories can create visually powerful new insights into trajectories. Christian Brzinsky-Fay (see Chap. 14) presents and discusses the advantages and limitations of the main, as well as the less conventional, visualisation options associated with the specific sequential data typically used in the field of social science. His chapter brings together the need for researchers to capture specific features of (or types of) sequences and the formal visualisation strategies derived from Cleveland's rules of perception.

Alexandre Pollien and Dominique Joye (see Chap. 15) apply SA to survey metadata. They analyse contact attempts as social interactions between interviewer and interviewee, each with specific socio-demographic characteristics, a process which leads to diverse degrees of accessibility and cooperation. Using information about the sampled population and a non-response survey, they map prototypical sequences on key variables, thereby showing how SA can enhance sample correction and open new pathways towards data quality improvement.

References

- Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21, 93–113.
- Abbott, A. (2001). *Time matters: On theory and method*. Chicago: University of Chicago Press.
- Abbott, A., & Barman, E. (1997). Sequence comparison via alignment and Gibbs sampling. *Sociological Methodology*, 27, 47–87.
- Abbott, A., & DeViney, S. (1992). The welfare state as transnational event: Evidence from sequences of policy adoption. *Social Science History*, 16(2), 245–274.
- Abbott, A., & Forrest, J. (1986). Optimal matching for historical sequences. *Journal of Interdisciplinary History*, 16, 471–494.
- Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology*, 96, 144–185.
- Abbott, A., & Prellwitz, G. (1997). Optimize software. <http://home.uchicago.edu/aabbott>. Accessed 25 June 2006 (not available anymore).
- Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology. Review and prospect. *Sociological Methods and Research*, 29(1), 3–33.
- Aisenbrey, S. (2000). *Optimal matching analyse*. Opladen: Leske-Budrich.
- Aisenbrey, S., & Fasang, A. (2010). New life for old ideas: The 'Second Wave' of sequence analysis bringing the 'Course' back into the life course. *Sociological Methods and Research*, 38(3), 420–462.
- Becker, H. S. (1963). *Outsiders: studies in the sociology of deviance*. New York: Free Press.
- Bison, I. (2009). OM matters: The interaction effects between indel and substitution costs. *Methodological Innovations Online*, 4(2), 53–67.
- Blair-Loy, M. (1999). Career patterns of executive women in finance: An optimal matching analysis. *American Journal of Sociology*, 104, 1346–1397.
- Blanchard, P. (2011). Sequence analysis for political science. *Working Papers of the Committee on Concepts and Methods, International Political Science Association*. <http://www.concepts-methods.org/WorkingPapers/PDF/1082>. Accessed 4 Nov. 2013.

- Blanchard, P. (Forthcoming 2014). "Validity, falsifiability, parsimony, consistency, precision, and so on": Les vicissitudes de l'innovation méthodologique. In M. Jouvenet & D. Demazière (Eds.), *La sociologie d'Andrew Abbott*. Paris: EHESS.
- Bühlmann, F. (2008). The corrosion of career? Occupational trajectories of business economists and engineers in Switzerland. *European Sociological Review*, 24(5), 601–616.
- Demazière, D., Jouvenet, M., & Abbott, A. (2011). "Les parcours sociologiques d'Andrew Abbott". Conference, University of Versailles Saint-Quentin-en-Yvelines.
- Deville, J. C., & Saporta, G. (1980). Analyse harmonique qualitative, [Qualitative harmonic analysis]. In E. Diday (Ed.), *Data analysis and informatics* (pp. 375–389). Amsterdam: North Holland Publishing.
- Elder, G. H. (1985). Perspectives on the life course. In Glen H. Elder Jr. (Ed.), *Life course dynamics, trajectories and transitions* (pp. 23–49). Ithaca: Cornell University Press.
- Elzinga, C. H. (2006). *Sequence analysis: Metric representations of categorical time series*. Amsterdam: Department of social science research methods.
- Elzinga, C. H. (2010). Complexity of categorical time series. *Sociological Methods and Research*, 38, 463–481.
- Fillieule, O., & Blanchard, P. (2012). Fighting together. Assessing continuity and change in social movement organizations through the study of constituencies' heterogeneity. In N. Kauppi (Ed.), *A political sociology of transnational Europe* (pp. 79–108). Basingstoke: ECPR Press.
- Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. (2011). *Mining sequence data in R with the TraMineR package: A user's guide*. Geneva: Department of Econometrics and Laboratory of Demography.
- Gauthier, J.-A. (2013). Optimal matching, a tool for comparing life-course sequences. In: R. Levy & E. D. Widmer (Eds.), *Gendered life courses between standardization and individualization. A European approach applied to Switzerland* (pp. 37–52). Wien: LIT.
- Gauthier, J., Widmer, E., Bucher, P., & Notredame, C. (2009). How much does it cost? Optimization of costs in sequence analysis of social science data. *Sociological Methods and Research*, 38(1), 197–231.
- George, L. K. (1993). Sociological perspectives on life transitions. *Annual Review of Sociology*, 19, 353–373.
- Halpin, B., & Chan, T. W. (1998). Class careers as sequences: An optimal matching analysis of work-life. *European Sociological Review*, 14(2), 111–130.
- Han, S.-K., & Moen, P. (1999). Clocking out: Temporal patterning of retirement. *American Journal of Sociology*, 105(1), 191–236.
- Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods and Research*, 38, 235–264.
- Hull, D. L. (1989). *The metaphysics of evolution*. NY: SUNY Press.
- Kruskal, J. (1983). An overview on sequence comparison. In D. Sankoff & J. B. Kruskal (Ed.), *Time warps, string edits, and macromolecules. The theory and practice of sequence comparison* (p. 1–44). United States: CSLI Publications.
- Lesnard, L. (2006). *Optimal matching and social sciences*. CREST Working Papers 2006–01, INSEE: Paris.
- Lesnard, L. (2008). Off-scheduling within dual-earner couples: An unequal and negative externality for family time. *The American Journal of Sociology*, 114(2), 447.
- Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods and Research*, 38, 389–419.
- Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Cybernetic Control Theory*, 10(8), 707–710.
- Levy, R. (1991). Status passages as critical life course transition: A theoretical sketch. In W. R. Heinz (Ed.), *Theoretical advances on life course research*. Weinheim: Deutscher Studien Verlag.
- Levy, R., Ghisletta, P., Le Goff, J. M., Spini, D., & Widmer, E. (2005). Towards an interdisciplinary perspective on the life course. *Advances in life course research*. Amsterdam: Elsevier JAI.

- Levy, R., Gauthier, J. A., & Widmer, E. (2006). Entre contraintes institutionnelle et domestique: Les parcours de vie masculins et féminins en Suisse. *The Canadian Journal of Sociology*, 31(4), 461–489.
- Macindoe, H., & Abbott, A. (2004). Sequence analysis and optimal matching techniques for social science data 387–406. In M. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. Thousand Oaks: Sage.
- Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48, 443–453.
- Piccarreta, R., & Lior, O. (2010). Exploring sequences: A graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society Series A*, 173(1), 165–84.
- Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 167–183.
- Robette, N. (2010). The diversity of pathways to adulthood in France: Evidence from a holistic approach. *Advances in Life Course Research*, 15, 89–96.
- Robette, N. (2011). *Explorer et décrire les parcours de vie: Les typologies de trajectoires*, Paris: CEPED.
- Robette, N., & Bry, X. (2012). Harpoon or bait? A comparison of various metrics to fish for life course patterns. *Bulletin de Méthodologie Sociologique*, 116(1), 5–24.
- Rohwer, G., & Poetter, U. (2007) *Transition data analysis*. Bochum: Ruhr University. <http://www.stat.ruhr-uni-bochum.de/tda.html>. Accessed 4 Nov. 2013.
- Stovel, K. (2001). Local sequential patterns: The structure of lynching in the deep South, 1882–1930. *Social Forces*, 79, 843–880.
- Stovel, K., Savage, M., & Bearman, P. (1996). Ascription into achievement: Models of career systems at Lloyds bank, 1890–1970. *The American Journal of Sociology*, 102(2), 358–399.
- Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative* (1st Ed). Graphics Press.
- Wiggins, R., Erzberger, C., Hyde, M., Higgs, P., & Blane, D. (2007). Optimal matching analysis using ideal types to describe the lifecourse: An illustration of how histories of work, partnerships and housing relate to quality of life in early old age. *International Journal of Social Research Methodology*, 10(4), 259–278.

Part I

How to Compare Sequences

Chapter 2

Motif of Sequence, Motif *in* Sequence

Shin-Kap Han

What Makes a Motif a *Motif*?

Richard Coles, a radio host and chaplain of the Royal Academy of Music, recalls an edition of the weekly classical music quiz show on BBC TV, *Face the Music*, where Joseph Cooper (the chair of the show) played a single note on the piano, which Joyce Grenfell (a panel member) correctly identified as the beginning of Debussy's piano prelude *La Fille aux Cheveux de Lin* (2008). A single note!

The questions Coles poses after relating the episode are exactly the ones I would do too: How did she know? Was it a lucky guess or a photographic (or, rather, phonographic) memory on Grenfell's part? Or was there something special in the composition? If the latter were the case, what makes it a telltale signature, which at its best has the power to effectively express the whole? A better-known example is the few opening bars of the first movement of Beethoven's Fifth Symphony: Ba-ba-ba booom ... ba-ba-ba booom. They seem to come out of nowhere; yet, arresting and recognizable, these few bars work with such economy that the whole of the first movement of the symphony could be described as a development of that motif.

My starting point is that this part-whole relationship—especially when the relationship is embedded in the formal structure of strings of successive states, events, actions, or notes—has a clear and close analog in the sequence analysis as practiced in social sciences in general and as implemented in optimal matching in particular. In that vein, I explore a few parallels and intersections, musical and otherwise, among these analogs to find my bearings. While advances in recent years have mostly been in methodological and technical domains, not much reflection has been seen in the theoretical domain. What I attempt here is to break the hiatus by looking outside. Throughout these excursions, the main thrust is to appropriate the concept of *motif* and its various usages in a range of extracurricular settings.

In the following section, I frame motif as a special type of subsequence and elaborate the rationale for, and the issues involved in, doing so. Three sources to

S.-K. Han (✉)

Department of Sociology, Seoul National University, Seoul, Korea

e-mail: shinkaphan@snu.ac.kr

borrow from—musical composition, molecular (or computational) biology, and social network analysis—are examined next. While disparate in substantive context, these sources provide cases that are homologous in terms of their formal structure. I search for the points of contact that will allow appropriation theoretically as well as methodologically. Finally, I conclude by gathering these threads. With a discussion of the caveats, I suggest how they can be fruitfully repurposed to advance the technical fronts and solidify the substantive bases of sequence analysis.

Local Components in Global Comparisons

The operational core of sequence analysis is to align, usually through optimal matching, two or more sequences (strings or vectors) and measure the extent to which they differ.¹ One looks for patterns shared by multiple sequences, which shed light on the structure of those sequences, and possibly what those shared patterns might do—the functions and meanings—within them. In the identification of patterns through sequence comparison, however, there are two strategies. One is *global* multiple alignment, the goal of which is to align complete sequences, and the other is *local* multiple alignment, where the aim is to locate relatively short patterns, i.e., subsequences, shared by otherwise dissimilar sequences (Lawrence et al. 1993). Thus far, a large majority of the work in social sciences follows the standard procedure of global multiple alignment (Abbott and Tsay 1990), that is, to compare and sort whole sequences. Yet, though mostly in the disciplines outside of social sciences (such as in studies of DNA), there has been continued and active interest in looking at parts of the sequences for regions of similarity or common subsequences.

This global/local distinction, however, is neither an equivalent of the level of analysis problem nor a parallel to the holism-reductionism debate one often finds in social sciences. In both strategies, the theoretical focus is on the properties in connections between the elements arrayed, such as “narrative order, sequential dependency, interlocked contingencies” (Abbott 1995) or “molecular structures and biological properties” (Lawrence et al. 1993). Both preserve the essential properties of sequence. And, more often than not, the two trek the same path in tracing the process of enchainment and unfolding. The key distinction between the two is not in *what*, but in *how*—i.e., in analytical focus. Even then, as in the case of motif discussed below, they intersect and overlap with each other.

¹ In typical practices, the result of this operation in the form of dis/similarity matrix is used for clustering, which then is presented with a variety of visualization techniques, as in Han and Moen (1999).

Table 2.1 A typology of subsequence to isolate motif

Recurrent/prevalent in form	Thematic/central in substance		
	Yes	(1)	No
	No	(3)	(4)

Motif as a Special Type of Subsequence

Motif, in its general usage, refers to “a distinctive, significant, or salient theme or idea” or “a recurrent or prevalent characteristic” (*Oxford English Dictionary*, at “motif”). Note the pairing of the two features, one referring to the substantive aspect and the other the formal aspect of sequence-subsequence relationship. These features also key in to the two principal questions in sequence analysis: the “generating” (what produces temporal regularity?) and “pattern” questions (is there temporal regularity?), respectively (Abbott 1990; Stovel 2001).

Depending on the case at hand, however, the relative emphasis shifts between the two. In literature and literary criticism, for instance, the former is what matters most: The motif is to be elaborated, but not necessarily to be repeated.² The ‘recurrence,’ when used in these disciplines, is usually between and across works as in folklore studies (Propp 1968). In contrast, clearly apparent in music is the latter emphasis, where motif refers to a short, usually recurrent, melodic or rhythmic unit (Schenker 1980). Ravel’s *Boléro*, for instance, is a more exacting, ‘ostinato-based,’ case of such (Kamien 2010). Recurrence, not necessarily in the exactly identical form, is crucial there. In needlework and lacemaking as well as in art and architecture, it can be both, where motif may refer to a single or recurring form, shape, or color in the design or pattern (Jones 1987).

In these diverse locales, two features, either alone or in tandem, define and characterize motif as a special type of subsequence—a part that can represent the whole (*paris pro toto*). While they are not independent, they are not adjunct to each other either. When both of them are present, as in (1) of Table 2.1, it can be seen as an effective shorthand for the entire sequence and stand in for the basis for their comparisons. Even with only one of them present, as in (2) or (3), it could, though not as tidily as in (1), serve the same purpose. In these cases, and in theory, the distinction between global and local multiple alignments becomes practically moot, for the motif as a subsequence represents the complete sequence.

Presence of a motif with regard to substance, hence, suggests that there is a part, a subsequence, that contains and connotes the whole more directly—motif *of* sequence. The rest may not be as significant materially. Similarly, presence of a motif with regard to form means that there is a pattern that is being repeated—motif *in* sequence. With repetition comes redundancy, which may be dispensable. The part, of course, cannot totally contain the whole. But it always partially contains the whole

² Of interest in this context is its usage in chess, where it means an element of a move in the consideration of why the piece moves and how it supports the fulfillment of the problem stipulation. This particular usage is related to the word’s usage in French, in which motif also refers to *motive* or *purpose*.

(Kosko 1994). The more it contains in less, the better it is as a motif. If such a motif could be found, one may infer that it may be possible, so to speak, to “separate the wheat from the chaff” with proper handling. That is, instead of analyzing the entire array, one may be able to selectively focus on a part without much loss of information. Were it to be the case, thus, we gain technically in efficiency and effectiveness, for we can target more narrowly. We also gain substantively in relevance and validity, for we can concentrate on the part that matters.

On the Cutting Board

Conceptualized as a special type of subsequence, motif can be configured within the existing operational framework. As discussed below, however, doing so requires recalibrating it to focus on subsequence identification.

If s_i^k denotes the state i is in at time k , the sequence, S_i , can be represented as a vector:

$$S_i = (s_i^1, s_i^2, s_i^3, \dots, s_i^k, s_i^{k+1}, \dots, s_i^n),$$

where n , the dimension of S_i , denotes its length or the number of elements in it. S_i is aligned to the other sequences. In this alignment, the underlying premise is that these elements are not arrayed at random. Instead, some form of association (e.g., imitative, generative, etc.) between the elements is presumed, which provides the basis for a patterned regularity over time.

The alignment, though, can be done either globally or locally, comparing the sequences in whole or in part. And, at times, the distinction between the two is blurred. Take, for example, a speech that closely follows Dale Carnegie’s dictum that one should tell the audience what she is going to say, say it, and then tell them what she has said. Any of the three parts shown in Fig. 2.1-(a) can stand in for the whole, and consequently the results from the global and local multiple alignments will be identical to each other. This is a blissful case where, as in (1) of Table 2.1, a motif is found that satisfies both of the conditions.

The standard procedures for global multiple alignment are closely followed by Han and Moen in their study of the temporal patterning of retirement (1999). They obtain career pathway types by comparing respondents’ work careers and use that as a main explanatory variable. But in linking what happened before retirement to what happened during and after, they redefine the sequence boundary and analytically exploit the discrepancy between the two. In sorting career pathway types, they compare entire lengths of employment histories, thus adopting the global multiple alignment strategy. Yet when they bring in timing and type of retirement and post-retirement employment (the unshaded part in Fig. 2.1-(b)), the employment sequence (the shaded part) becomes a subsequence embedded in a lengthier sequence. In other words, the boundary of the sequence has been drawn twice, first to delineate the work career and then to extend it to retirement and beyond, thus



Fig. 2.1 Various sequence structures obviating the distinction between global and local alignments

turning the whole delineated at first into a part of the augmented whole later. In this setting, if we could tip the balance between the shaded (*explanans*) and unshaded (*explanandum*) parts, so much the better.

Further out along this line of reasoning lies the concept of motif. If a subsequence could be identified, which preserves and contains sequential dependency and narrative order, it could serve as an analytical catalyst. A motif as such not only obviates the distinction between global and local multiple alignments, but also provides a way to economize the process. If we let the length of a sequence be denoted by ϵ , the sequence of dimension n has $n = \epsilon$. Thus, ϵ may take up any value within the range, $[1, n]$. In theory, that is, ϵ of a motif as a subsequence can be of a value as small as 1 and as large as n . In practice though, the smaller the value of ϵ , the better; and the larger difference, $n - \epsilon$, the bigger the gain in efficiency.³ Presented in Fig. 2.1-(c) is a hypothetical motif with its length (ϵ) less than one tenth of n .

That is, of course, if we can identify it at a search cost less than the efficiency gain accrued. The search for such a subsequence poses a difficult analytical challenge of its own. The question, in short, is how, or more specifically, *where* to cut. It involves locating the beginning and ending points of the subsequence—and hence, deciding the location and length of the subsequence—to use as a motif.

Diagrammed in Fig. 2.1-(d) is an exceptional case, in which a theory provides explicit guidance. In organizational research, “imprinting” is a concept defined as a process whereby, during a brief period of susceptibility, an entity develops characteristics that reflect prominent features of the environment, and these characteristics continue to persist despite significant environmental changes in subsequent periods. As to why organizations and industries that were founded in the same period were so similar even today, for example, Stinchcombe argues that external environmental forces powerfully shaped firms’ initial structures during the founding period, and these structures persisted in the long run, well beyond the time of founding (Stinchcombe 1965; Baron et al. 1999).⁴ This, however, is the exception that only proves

³ The gain in calculation load, which is quadratic, could be as large as $n^2 - \epsilon^2$ (Abbott and Tsay 1990).

⁴ Although Stinchcombe did not specifically use the term “imprinting,” the term soon became associated with this essay (Lounsbury and Ventresca 2002).

the rule; such clearly specified theoretical guidelines for locating subsequences would be few and far between.

In Table 2.1, the two dimensions used to classify subsequences are not equally matched in terms of feasibility. Formal recurrence is far easier to detect than substantive centrality. One does not determine the other; however, they are not entirely independent of each other either. Thus, hopefully, a better understanding of one might lead to insights on the other. With that in mind, one may look first at the formal aspect.

Resources to Mobilize

If the discussions in the preceding section are to hold water at all, we need tools to delineate a motif, a special type of subsequence. Instead of forging them anew, I look near and far to borrow. This is, after all, the strategy followed in the early period of adopting sequence analysis into sociological research (Abbott and Tsay 2000; Abbott 1995), and I am merely refreshing the process here. And I am mindful of the issues such an enterprise is fraught with, as Abbott himself noted in the following: “Specialists in these various areas may find me superficial towards their own interests even as they find me unduly concerned with those of others. These seem to me to be the inevitable costs of such a survey” (1995, p. 129). The goal here, though, is not to replicate the original materials to the letter, but to adapt them for our own, very practical use.

I set the scope of prospecting wide. Locating the structurally parallel locales, I collect appropriate analogues, harvest suitable components, and glean apposite insights and inspirations from a range of disciplines. Of course, the applicability and efficacy of these tools will depend on the setting as seen in Table 2.1 and Fig. 2.1. Yet these disparate resources, when carefully put together, might be repurposed for the problem of identifying a motif.

In the literature from sociology proper, the studies that focus on subsequence are rather limited in numbers and sources. A quick round-up will net mostly the works by Abbott: brief considerations of common subsequences in Abbott and Tsay (1990) and Abbott (1997), a discussion of “turning points” as a particular case in point in Abbott (1997), and a more explicit and elaborate treatment of it, using a Gibbs sampler, in Abbott and Barman (1983).⁵ Much of these, however, are directed toward the theoretical and substantive elaboration of “turning points,” which provides little bearing on the problem at hand.

There are, on the other hand, robust and sophisticated algorithms available, such as *TraMineR* (Gabadinho et al. 2011), to handle technical issues of subsequence identification and transition sequences. Yet their developmental tracks have been oriented mainly toward methodological and empirical purposes. Their theoretical bearings, including one on motif, have been mostly unexplored thus far. It is, in

⁵ To some extent, Hollister’s *localized OM* is based on the similar logic, i.e., on giving differential weights to different parts of the sequence (2009).

large part, due to the absence of analytical framework to articulate the two sides, which is what this chapter is after.

Looking beyond the disciplinary boundary for opportunities to borrow (and chances to steal!) yields a more interesting ensemble. I select three wide-ranging areas for such a survey below: musical composition, molecular (or computational) biology, and social network analysis.

Note for Note

I start with music. First and foremost, in its formal structure, it is a type-case of narrative that unfolds itself over time in a sequential manner (Newcomb 1987). The best example is probably the sonata form, in which a single movement is divided into three main sections: the exposition (establishing the first and second themes in contrasting keys), development (modulating in structure, tone and rhythms) and recapitulation (returning to the main themes), sometimes followed by a coda. Also, as in the earlier examples that opened this chapter, it allows a quick intuitive grasp of the notion of motif.

The focal point is musical plagiarism. To establish it in court, the plaintiff must prove that the defendant had a reasonable possibility of access to her earlier work and demonstrate that there are substantial similarities between hers and the defendant's. It is the latter problem that presents a challenge here. In theory, the outline is clear:

There must be sufficient objective similarity between the infringing work and the copyright work, or a substantial part thereof, for the former to be properly described, not necessarily as identical with, but as a reproduction or adaptation of the latter.⁶

In practice, however, it is difficult to legally define what constitutes “sufficient objective similarity.” And those difficulties keep breeding peculiarities and inconsistencies in court decisions, providing grounds for continuing legal disputes on one hand and creating needs for innovative approaches to music and law on the other.

While musical expressions involve multiple layers (e.g., rhythm, harmony, phrasing, instrumentation, and style), judging whether the two pieces share unique musical components has been done largely in terms of melodic similarities and between no more than a few measures thus far (Cronin 1998).⁷ A typical case is *Hein v. Harris* shown in Fig. 2.2.⁸ In that case, Judge Learned Hand found from his bar-by-bar analysis that thirteen of the first seventeen bars of the two melodies were “substantially the same” and concluded that Howard must have copied Hein’s song.

⁶ Francis Day Hunter Ltd v Bron, Chap. 587 (1963).

⁷ We are leaving aside the issues of lyrics (e.g., Johnney Cash v Gordon Jenkins) and the recent phenomena of sampling (Vanilla Ice v Queen and David Bowie).

⁸ The composer of the defending work was Joseph E. Howard. The suit—Hein v Harris 175 F. 875 (C.C.S.D.N.Y. 1910)—was filed against his publisher, Harris. See Music Copyright Infringement Resource at USC Gould School of Law (mcir.usc.edu).



Silvio Hein, "Maria Cahill's Arab Love Song"



Joseph E. Howard, "I Think I Hear a Woodpecker"

Fig. 2.2 Comparison of two melodies (Dark note heads are for unisons between the two melodies)

Another case found George Harrison liable for copyright infringement.⁹ The court's tone is almost apologetic in determining that Harrison "subconsciously" misappropriated the "musical essence" of Ronald Mack's "He's So Fine" in his "My Sweet Lord." The court relied heavily on the fact that the melodic kernels of plaintiff's popular number were used in the same order and repetitive sequence.

In both cases, the court's analysis seems useful and the findings essentially accurate.¹⁰ The issue of musical similarities, however, is far from settled. Some, for instance, argue that such note-for-note comparison ("by the eye") is fundamentally incomplete given the inherent complexities of music and thus should be complemented by aural perception ("by the ear") (Cason and Müllensiefen 2012). Still at issue too is the substantial part doctrine. Once a claimant is successful in demonstrating a sufficient degree of similarity between the disputed works, she has to establish whether the section reproduced represents a "substantial part" of the claimant's work (Baker 1992). Note that both invoke the part-whole relationship as a principal aspect of sequence representation as discussed above. As such, these considerations prompt further questions concerning the two cases above. For the former, why the first seventeen bars, and not, say, the first eight measures? And for the latter, what are, and how does one delineate, kernels and motifs? These are the questions familiar to those who do sequence analysis in other disciplines.

Lastly, mating musical composition and computational methods engenders an interesting crossbreed—algorithmic music. It starts with the question about the very beginning: Where does music come from? David Cope, for one, believes that all music is essentially "inspired plagiarism" (2005). The great composers absorbed the music that had gone before them and their brains "recombined" melodies and phrases in distinctive, sometimes traceable, ways—a process he calls "inductive association." He contends furthermore that such a process can be programmed. He describes this computational process in his book, *Experiments in Musical Intelligence*

⁹ Bright Tunes Music v Harrisongs Music 420 F. Supp. 177 (S.D.N.Y. 1976).

¹⁰ Of the existing proposals to bring in computational methods to the issue, the main part is still based on string matching algorithms that represent music as sequences of notes (Robine et al. 2007).

(1996) and presents *Emmy*, the algorithm he developed. When fed with enough of a composer's work, *Emmy* could deconstruct it, identify signature elements, and recombine them in new ways. And it actually did, producing works, including *Virtual Mozart* and *Virtual Rachmaninoff*.¹¹ Whatever this implies for the question of what music means, this logic of recombinancy and the assumptions underlying it can have a profound implication for us on the ways in which sequence is seen.

The ideas and tools developed to deal with the plagiarism in music seem familiar and readily adaptable for our purpose. They might be especially useful in conceptualizing the issue of motif and the problems—both theoretical and practical—it entails (e.g., the contrast between “by the eye” and “by the ear,” “substantial part doctrine,” and “recombinancy”).

A Gene is Made of DNA Sequences

As songs consist of sequences of notes, genes are made of DNA sequences. In theory, thus, what is formally true for the former is also applicable to the latter and vice versa. While, instead of melodies, protein or nucleic acid sequences are searched for shared patterns, the general outline of the approach is the same.

Upon that basis of commonality, each build their own substantive applications with distinct disciplinary orientations: In molecular biology, it is to shed light not only on molecular structure, but also on biochemical functions and evolutionary development (Lawrence et al. 1993). The early use of sequence comparison in molecular biology has largely been to detect and characterize the homology, or correspondence, between two or more related sequences, leading to evolutionary inferences. Of late, though, there has been a shift of focus in research. DNA sequence data are becoming available at a rapidly increasing rate and are now offering new ways of looking at genetic processes, including genetic diseases. After all, “[P]roteins and nucleic acids are macromolecules central to the biochemical activity of all living things, including chemical regulation and genetic determination and transmission” (Kruskal 1999, p. 3). In other words, these sequences hold the information for the construction and functioning of these organisms.

The challenge is how to read them, or perhaps more to the point, how to read more of them faster. Seizing on the fact that these sequences contain many kinds of motifs—i.e. re-occurring patterns, associated with specific biological functions, much research has been devoted to computer algorithms for discovering such motifs in sequences. These recent streams of research differ from the previous ones in three ways. One, they range much more freely across substantive domains and analytical levels, e.g., from biochemistry and neurobiology to ecology and engineering (Kashtan et al. 2004; Milo et al. 2002). They search for structural design principles across fields where complex networks constitute the structural base. In

¹¹ For those interested in how Emmy's compositional abilities fare over large numbers of compositions, he collected 5000 MIDI files of computer-created Bach-style chorales and placed at <http://artsites.ucsc.edu/faculty/cope/5000.html> for download.

			Feed-forward loop			Bi-fan			Bi-parallel		
Network	Nodes	Edges	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z_{score}	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z_{score}	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z_{score}
Gene regulation (transcription)											
<i>E. coli</i>	424	519	40	7 ± 3	10	203	47 ± 12	13			
<i>S. cerevisiae</i>	685	1,052	70	11 ± 4	14	1,812	300 ± 40	41			
Neurons											
<i>C. elegans</i>	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs											
Little Rock	92	984							7,295	$2,220 \pm 210$	25
Chesapeake	31	67							26	5 ± 2	8
Electronic circuits (forward logic chips)											
a15850	10,383	14,240	424	2 ± 2	285	1,040	1 ± 1	1,200	480	2 ± 1	335
a38584	20,717	34,204	413	10 ± 3	120	1,739	6 ± 2	800	711	9 ± 2	320

Fig. 2.3 Some network motifs found in various networks (Excerpted from Table 2.1 Network Motifs Found in Biological and Technological Networks in Milo et al. (2002, p. 826).

these works, as a result, the following three phrases are often used interchangeably: sequence motif, structural motif, and network motif. The case in point is the table in Fig. 2.3 below taken from Milo et al. (2002), which allows a fascinating comparative perspective.

Two, at the operational core of their works is counting and sampling (a small set of) subsequences and identifying those that occur at numbers that are significantly higher than those expected. Interestingly, this computational turn seems to be a return to the root of sequence analysis, i.e., string algorithms, with a “big data” twist (Sankoff and Kruskal 1999; Gusfield 1997). In this sea of DNA, the practical question is, how do we search for instances of a motif? Consequently, much of the effort in the field of late has been in accelerating the speed and expanding the scale of the search (Frith et al. 2008; Grochow and Kellis 2007).

Three, and most importantly, they go beyond morphology. They link structures with functions. In these works, network motifs are demonstrated to play key information processing roles in biological regulation networks (Shen-Orr et al. 2002). These network motifs have recently been found in diverse organisms from bacteria to humans, suggesting that they serve as basic building blocks of transcription networks (Alon 2007; Kashtan et al. 2004). Of particular interest is the “regulator gene,” involved in controlling the expression of one or more other genes.

Possibly connecting the instrumentation for sampling and counting subsequences and the analysis to specify their functions is the issue of noncoding DNA sequences.

They refer to portions of a genome sequence for which no known biochemical function has been identified, hence the label, “junk DNA” (Orgel and Crick 1980). In the human genome, for instance, more than 98 % of the DNA is believed to be noncoding. If true, it could mean that it is not necessary to examine the entire length of the sequence. Increasing evidence, however, is indicating that there are discernible patterns in this noncoding DNA (Flam 1994) and it might be influencing the behavior of the coding DNA (Biémont and Vieira 2006).

The technical, especially computational, advances this field has made are certainly of interest for sequence analysts in general and particularly for those with big data implementation issues. Of more subtle, yet radical, import is the discussion of subsequences as building blocks of genes and their functions. This might as well be—if not, should be—one of the next steps for the social scientists.

Network as a Sequence, Sequence as a Network

Sequences can be conceived and represented as digraphs (or directed graphs) (Harrow et al. 1965). For sequences, however, the arcs in them are restricted to be only in one direction. A type-case is chronological sequences, in which the arcs must be directed from time t to time $t + k$ ($k > 0$). Even when there is no intrinsically determined way to designate origin and destination points, one chosen direction, whether from left to right or from top to bottom, is adhered to. As such, they form vectors, i.e., one-dimensional arrays. Strings of relations can readily be adapted to the sequence framework. Networks as such are digraphs—and hence, sequences—as well. They differ, however, in that they are typically two- or three-dimensional arrays. Establishing a formal analogy between networks and sequences thus requires a dimensional shift. With that as a caveat, one may look to social network analysis (SNA) to pan for materials.¹²

In the literature, there has long been a strand that looks at the distribution of parts (or smaller structures) to make sense of the nature of the whole (the larger structure). Triad census, based on all sixteen types shown in Fig. 2.4, is an earlier example of analysis using subgraphs (Holland and Leinhardt 1970). Watts (2004) finds the work on network motifs by molecular biologists, such as (Shen-Orr et al. 2002) and Milo et al. (2002), “identical in spirit” to this literature. There seems to be far more than spiritual similarity: It is analogous in defining network motifs as topologically distinct subgraphs whose frequencies in a network can be used to characterize its overall structure. It is also technically parallel—systematically enumerating all the motifs comprising three and four nodes in a number of networks first, and then comparing the resulting counts with those from random networks.

¹² In this volume, Bison too exploits this linkage, in a direct, albeit quite different, manner, by plotting sequences as networks. Such a translation, he argues, provides an alternative way to observe and measure new structural features of complex narratives (Bison 2012; Abell 2004; Bearman and Stovel 2000).

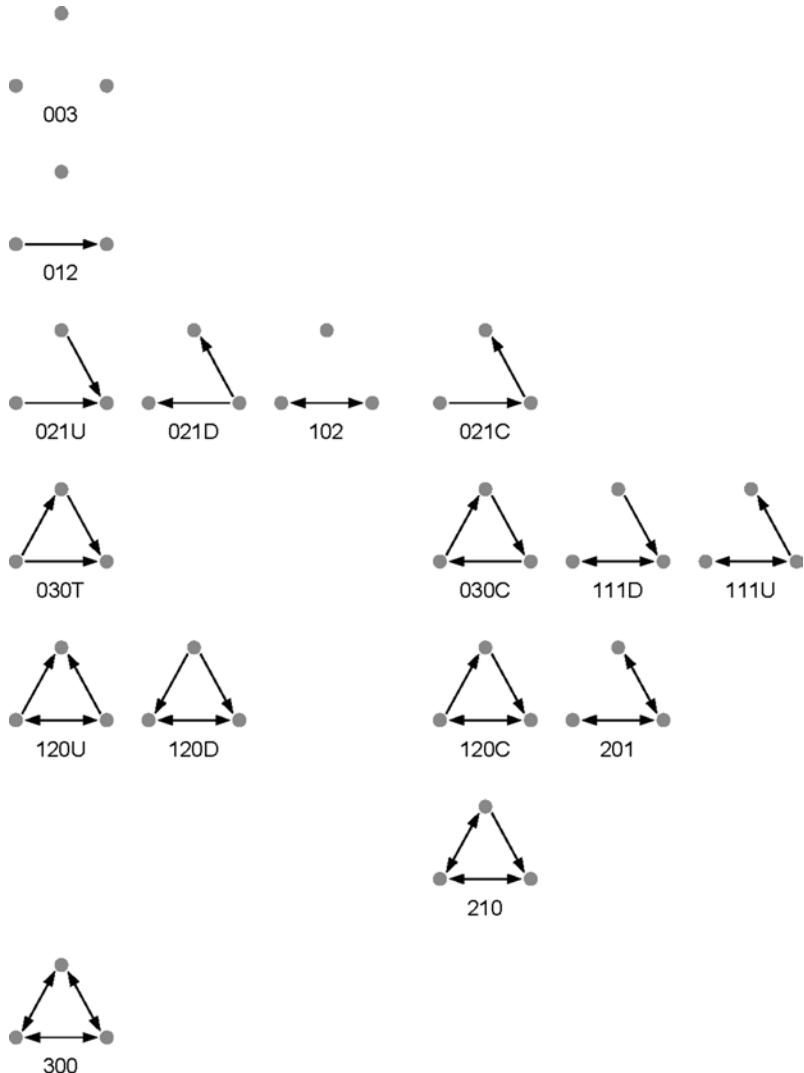


Fig. 2.4 Sixteen triad types (All sixteen triad types arranged vertically by number of choices made and divided horizontally into those with no intransitivities and those with at least one. See Fig. 2.1 in Holland Leinhardt (1970, p. 496) and Fig. 2.4 in Han (2003, p. 268)

Of more recent vintage along this line of research are the exponential-family random graph models (ERGMs) of networks. The basic stochastic model can be expressed by the following general form (Wasserman and Robins 2005):

$$Pr(Y = y) = \left(\frac{1}{k} \right) \exp \left\{ \sum_A \eta_A g_A(y) \right\},$$

where Y is a network realization and y is the observed network. The summation is over all configurations A . k is a normalizing factor. η_A is the parameter and $g_A(y)$ is the network statistic corresponding to configuration A .

The purpose of ERGMs is to describe parsimoniously the local selection forces that shape the global structure of a network. That is, a social network is thought of as being built of these local patterns of ties, called “network configurations,” which correspond to the parameters in the model (Lusher et al. 2012). In this framework, for instance, one may ask, does a given network structure occur due to processes of homophily, reciprocity, transitivity, cyclicity, or a combination of these?¹³ And as such, the formulation also takes into account the inferential potential of the sequence analysis (King 2013). The efforts at substantive as well as technical developments of late are mostly directed toward inclusion of higher-order local structures, such as k -stars, k -triangles, and independent two-paths and their alternating versions.¹⁴

In this framework, we are brought back to reconsider the part-whole relationship discussed earlier in formulating the concept of motif. Yet the two sides are engaged not just formally, but in theoretical and substantive ways as well. Such a dual engagement is precisely what will allow us to consider the two dimensions in Table 2.1 simultaneously: recurrent/prevalent in form and thematic/central in substance.

Concluding Remarks

I explored a few avenues that intersect our main topic—the core program of “sequence analysis,” in which sequences are strings of successive states, events, or actions, and for which optimal matching serves as the standard operational framework. In those excursions, the main thrust was in appropriating and repurposing the concept of motif, defined as a distinctive and recurring set of structural elements, and its various usages from a range of extracurricular settings. This focus on motif, of course, is not effective everywhere. But, where it works, it could generate interesting and important leads.

Although certainly not exhaustive in any way, a few leads are found scattered in diverse settings in varying shapes. And much has been gained, hopefully, culling usable bits and pieces to advance the technical front on one hand, and solidifying the substantive embedding of sequence analysis on the other. Those diverse settings all deal with arrays of element. The elements arrayed may be informational or corporeal. And the arraying may be linear/temporal or multidimensional/spatial. Yet the analytical issues—especially, in their structural forms—are analogous, which allows borrowing from one another. The key is to see the sequences as built from,

¹³ Currently, they are implemented in *statnet* (<http://statnet.csde.washington.edu/>) and *PNet* (<http://sna.unimelb.edu.au/PNet>).

¹⁴ Appendix A. Table of Model Terms in (Morris et al. 2008) provides quick reference for what terms are appropriate to a particular model.

with, and by component blocks, chunks of elemental, basic units that form a minimal substantive footing, akin to White's *social molecules* (1992).

These examples do resemble one another, particularly in highlighting the common subsequence problem. They all see the potential for motif. And at a deeper level, they see that sequences could be generated endogenously and recursively, i.e., that certain subsequences, stages, events or specific episodes could have an influence on the further enchainment of the sequential elements, as in the phrase "defining moments" or "critical events" (Blanchard et al. 2012). Yet the fact that music and narrative, for instance, both involve a succession of events in a regular order does not necessarily mean that music has a special affinity to narrative (Maus 1991). In general, claims that two kinds of object are close in terms of formal structure are risky: it is too easy to find and describe shared structures across many different domains (Kruskal 1999). That is, while it is useful to exploit the analogies between them, it is imperative to pay attention to the significant empirical differences and theoretical distinctions between them as well to avoid superficial transfer (Biemann 2011).

Further Issues to Consider

There are a few issues to consider to implement the idea of motif in practice. One of the key issues is that of bounding the motif. One twist here is that in so doing, we have to find a halfway stop between the parts and the whole—i.e., at a subsequence level. The problem has implications on several levels.

First and foremost, the underlying presumption on the nature of sequence structure is that, empirically, the elements are *not* arrayed randomly. The patterned regularities we seek are taken as the results of that non-randomness (such as "enchainment, order, convergence" (Abbott 1995)). In turn, theoretically, we presume that there are underlying processes that produce these regularities. Within this overall framing, motif specifies that those theoretical and empirical keys are to be found in the constituent components of sequences. As discussed earlier, however, it does not explicitly specify how big, or how long, those components are (Elzinga and Wang 2012). At another level are the related issues of granularity¹⁵, gap, and nestedness, which are particularly difficult ones in temporal dimensions. Thus far, the answers to these questions have been dictated largely by the exigencies of available data without much articulation.

While these issues may seem to be of more empirical/methodological nature, they touch upon the fundamental issue of how to break things down and how to put them back up. That is, if we are to understand sequences as social narratives, we have to identify not only their temporal structures, but the interdependent processes in (of?) them as well. With motif, in particular, we have to ask: What are the funda-

¹⁵ Granularity in general is the extent to which a system is broken down into small parts, either the system itself or its description or observation.

mental building blocks? How do they combine to form larger structures? Do these structures which share the same building blocks also share the same combinations of these blocks?

Looking Backward, Looking Forward

In the invitation letter for LaCOSA conference, the organizers write: “We currently lack a broader and systematic debate that takes stock of the advances and limits of sequence analysis, that encourages a careful standardization of the approach beyond diverging orientations, and that opens and explores new methodological paths and combinations.” For that, then, let’s ask again what we do sequence analysis for. In it, the first problem is always to figure out if the patterns are there. By ‘patterns,’ we mean regularities in sequence, as in sequence types, and we look for them by comparing sequences. At various stages of this classification exercise, we continuously engage in data reduction—either by necessity or for convenience. That much is clear from the technical point of view. Imperative as those demands are, we should also keep our eyes on the prize, i.e., theorizing intertemporal dynamics. That is, after, and at times simultaneously with, the classification, we use that ‘reduced-form data’ to do explaining-modeling-theorizing about the processes of unfolding, the mechanisms of entailment, and the structures of temporal space. And that is what we should keep our eyes on. As Abbott forcefully puts it, “The proof of the classificatory pudding comes in the explanatory eating” (1990, p. 15).

In exploring this avenue, it might be helpful to take lessons from a kindred experience. In an essay aptly titled “Structural Analysis: From Method and Metaphor to Theory and Substance,” Wellman (1988, pp. 19–20) poignantly describes the predicament of social network analysis as of 1988:

These misconceptions have arisen because too many analysts and practitioners have (mis) used “structural analysis” as a mixed bag of terms and techniques. Some have hardened it into a method, whereas others have softened it into a metaphor. Many have limited the power of the approach by treating all units as if they had the same resources, all ties as if they were symmetrical, and the contents of all ties as if they were equivalent.

Yet, structural analysis does not derive its power from the partial application of this concept or that measure. It is a comprehensive paradigmatic way of taking social structure seriously by studying directly how patterns of ties allocate resources in a social system. Thus, its strength lies in its integrated application of theoretical concepts, ways of collecting and analyzing data, and a growing, cumulating body of substantive findings.

There are quite a few meta-theoretical parallels and practical similarities between social network analysis then and sequence analysis now: We are very much at that juncture, where the direction for us too is *from* method and metaphor *to* theory and substance. To pose new intellectual questions, collect new types of evidence, and provide new ways to describe and analyze social structures are what sequence analysis has to achieve. And, for that, thinking about sequence in terms of motif, and looking for the motif in social narrative, is a turn that might show a new path.

Acknowledgement This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-330-2012S1A3A2033451). I thank Sang-Jic Lee, who provided help in the manuscript preparation, and Yun-joo Sung, who prepared the scores in Fig. 2.2.

References

- Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods*, 16(4), 129–147.
- Abbott, A. (1990). A primer on sequence methods. *Organizational Science*, 1(4), 375–392.
- Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21, 93–113.
- Abbott, A. (1997). On the concept of turning point. *Comparative Social Research*, 16, 85–105.
- Abbott, A., & Barman, E. (1997). Sequence comparison via alignment and Gibbs sampling. *Sociological Methodology*, 27, 47–87.
- Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods and Research*, 29(1), 3–33.
- Abell, P. (2004). Narrative explanation: An alternative to variable-centered explanation? *Annual Review of Sociology*, 30, 287–310.
- Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8(6), 450–461.
- Baker, M. (1992). La(w)-A note to follow so: Have we forgotten the federal rules of evidence in music plagiarism cases? *Southern California Law Review*, 65, 1583–1637.
- Baron, J. N., Burton, M. D., & Hannan, M. T. (1999). Engineering bureaucracy: The genesis of formal policies, positions, and structures in high-technology firms. *Journal of Law, Economics, and Organization*, 15(1), 1–41.
- Bearman, P. S., & Stovel, K. (2000). Becoming a Nazi: A model for narrative networks. *Poetics*, 27(2), 69–90.
- Biemann, T. (2011). A transition-oriented approach to optimal matching. *Sociological Methodology*, 41, 195–221.
- Biémont, C., & Vieira, C. (2006). Genetics: Junk DNA as an evolutionary force. *Nature*, 443(7111), 521–524.
- Bison, I. (2012). *Sequence analysis and network analysis: An attempt to represent and study sequences by using NetDraw*. In Lausanne Conference on Sequence Analysis (LaCOSA).
- Blanchard, P., Buhlmann, F., & Gauthier, J.-A. (2012). *Sequence analysis in 2012*. In Lausanne Conference on Sequence Analysis (LaCOSA).
- Cason, R. J., & Müllensiefen, D. (2012). Singing from the same sheet: Computational melodic similarity measurement and copyright law. *International Review of Law, Computers & Technology*, 26(1), 25–36.
- Coles, R. (2008). Got it licked. The Guardian. www.guardian.co.uk/music/2008/jul/22/popandrock.classicalmusicandopera Accessed 8 May 2013.
- Cope, D. (1996). *Experiments in musical intelligence*. Middleton: A-R Editions.
- Cope, D. (2005). *Computer models of musical creativity*. Boston: The MIT Press.
- Cronin, C. (1998). Concepts of melodic similarity in music-copyright infringement suits. *Computing in Musicology*, 11, 187–209.
- Elzinga, C., & Wang, H. (2012). Versatile string kernels. In *Lausanne Conference on Sequence Analysis* (LaCOSA).
- Flam, F. (1994). Hints of a language in junk DNA. *Science*, 266(5189), 1320. PMID: 7973718.
- Frith, M. C., Saunders, N. F. W., Kobe, B., & Bailey, T. L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Computational Biology*, 4(5), e1000071.

- Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software*, 40(4), 1–37.
- Grochow, J. A., & Kellis, M. (2007). Network motif discovery using subgraph enumeration and symmetry-breaking. In T. Speed & H. Huang (Eds.), *Proceedings of the 11th annual international conference on research in computational molecular biology (RECOMB'07)* (pp. 92–106). Berlin: Springer-Verlag.
- Gusfield, D. (1997). *Algorithms on strings, trees and sequences: Computer science and computational biology*. New York: Cambridge University Press.
- Han, S.-K. (2003). Tribal regimes in academia: A comparative analysis of market structure across disciplines. *Social Networks*, 25, 251–280.
- Han, S.-K. & Moen, P. (1999). Clocking out: Temporal patterning of retirement. *American Journal of Sociology*, 105(1), 191–236.
- Harary, F., Norman, R. Z., & Cartwright, D. (1965). *Structural models: An introduction to the theory of directed graphs*. New York: Wiley.
- Holland, P. W., & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology*, 76(3), 492–513.
- Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods and Research*, 38(2), 235–264.
- Jones, O. (1987). *The grammar of ornament: All 100 color plates from the Folio edition of the great Victorian sourcebook of historic design*. New York: Dover Publications.
- Kamien, R. (2010). *Music: An appreciation*. New York: McGraw-Hill.
- Kashtan, N., Itzkovitz, S., Milo, R., & Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11), 1746–1758.
- King, T. (2013). A framework for analysing social sequences. *Quality and Quantity*, 47(1), 167–191.
- Kosko, B. (1994). *Fuzzy thinking: The new science of fuzzy logic*. New York: Flamingo.
- Kruskal, J. B. (1999). An overview of sequence comparison. In D. Sankoff & J. Kruskal (Eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (pp. 1–44). Stanford: CSLI Publications.
- Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262(5131), 208–214.
- Lounsbury, M., & Ventresca, M. J. (2002). Social structures and organizations revisited. *Research in the Sociology of Organizations*, 19, 3–26.
- Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2012). *Exponential random graph models for social networks: Theory, methods, and applications*. New York: Cambridge University Press.
- Maus, F. E. (1991). Music as narrative. *Indiana Theory Review*, 12, 1–34.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298(5594), 824–827.
- Morris, M., Handcock, M. S., & Hunter, D. R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software*, 24(4), 1548–7660.
- Newcomb, A. (1987). Schumann and late eighteenth-century narrative strategies. *Nineteenth-Century Music*, 11(2), 164–174.
- Orgel, L. E., & Crick, F. H. (1980). Selfish DNA: The ultimate parasite. *Nature*, 284(5757), 604–607.
- Propp, V. (1968). *Morphology of the folktale*. Austin: University of Texas Press.
- Robine, M., Hanna, P., Ferraro, P., & Allali, J. (2007). *Adaption of string matching algorithms for identification of near-duplicate music documents*. In Proceedings of the International SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN'07) (pp. 37–43). Amsterdam, The Netherlands.
- Sankoff, D., & Kruskal, J. (Eds.). (1999). *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Stanford: CSLI Publications.
- Schenker, H. (1980). *Harmony*. Chicago: University of Chicago Press.

- Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1), 64–68.
- Stinchcombe, A. L. (1965). Social structure and organizations. In J. G. March (Ed.), *Handbook of Organizations* (pp. 142–193). Chicago: Rand McNally & Company.
- Stovel, K. (2001). Local sequential patterns: The structure of lynching in the deep south, 1882–1930. *Social Forces*, 79(3), 843–880.
- Wasserman, S., & Robins, G. (2005). An introduction to random graphs, dependence graphs, and p^* . In P. J. Carrington, J. Scott & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 148–161). New York: Cambridge University Press.
- Watts, D. J. (2004). The “New” science of networks. *Annual Review of Sociology*, 30(1), 243–270.
- Wellman, B. (1988). Structural analysis: From method and metaphor to theory and substance. In B. Wellman & S. D. Berkowitz (Eds.), *Social structures: A network approach* (pp. 19–61). New York: Cambridge University Press.
- White, H. C. (1992). *Identity and control: A structural theory of social action*. Princeton: Princeton University Press.

Chapter 3

Using Optimal Matching Analysis in Sociology: Cost Setting and Sociology of Time

Laurent Lesnard

Optimal Matching Analysis and Biology

As with any new statistical method in the social sciences (Desrosières 1993), the use of Optimal Matching Analysis (OMA) was much debated, especially with regard to the sociological interpretation of transformations and the role of costs in findings (Levine 2000; Wu 2000). OMA was widely used in biology, where the three operations (insertion, deletion, and substitution) are supposedly reflected in biochemical processes. Yet OMA was not created by biologists; it originated from research in computer science conducted by Richard Hamming and Vladimir Levenshtein in the 1950s and '60s (Hamming 1950; Levenshtein 1966). In OMA terms, Hamming's distance only uses substitution operations, whereas the first distance put forward by Levenshtein uses all three operations, each with a unit cost (see Table 3.1). Thus, the operations used in OMA have nothing to do with biochemical transformations; many of the latter, such as the transposition of sub-sequences, would be missing (Abbott 2000). While it is tempting to attribute the method's success in biology to these similarities, it is, in fact, through the interpretation and determination of costs that biologists were able to adapt OMA to their data.

The applications of OMA in biology are quite different from its use in the social sciences since in the former case the goal is not to create typologies, but more often to match a group of sequences with unknown properties to other sequences with known properties (Durbin et al. 1998). Substitution costs are determined on the basis of phylogenetic hypotheses and statistical results, while insertion and deletion costs are set more arbitrarily. Indeed, substitution costs must reflect the probability of change and are empirically estimated from sample sequences with established or probable phylogenetic links. For example, Point Accepted Mutation (PAM) matrices are estimated on the basis of proteins that are experimentally or hypothetically related phylogenetically (Dayhoff et al. 1978).

L. Lesnard (✉)
Sciences Po, Paris, France
e-mail: laurent.lesnard@sciencespo.fr

Table 3.1 Hamming and Levenshtein Distances (Lesnard 2010)

Distance	Transformations used	Insertion and deletion
	Substitution	
Hamming	Yes (cost=1)	No
Levenshtein I	Yes (cost=1)	Yes (cost=1)
Levenshtein II	No	Yes (cost=1)

The Importance of Costs

The main lesson that can be drawn from the use of OMA in biology is the importance of focusing on the interpretation and determination of costs rather than seeking an interpretation in the operations themselves. The interpretation of substitution costs is key in this respect since they need to reflect the proximity between the different states in which the sequences unfold. In biology, this proximity between states is interpreted in phylogenetic terms and the costs are therefore derived from samples based on phylogenetic knowledge and hypotheses. In the social sciences, this probability must express the degree of sociological proximity between states. But what is meant by sociological proximity?

It is precisely this issue that other critiques have targeted. In fact, the first researchers to use OMA in sociology often expressed doubts as to their choice of costs and the sensitivity of their results to these choices. As Katherine Stovel and her colleagues stated, “the assignment of transformation costs haunts all optimal matching analyses” (Stovel et al. 1996, p. 394). If the results are too sensitive to cost, it means that the costs, and not the data, are producing the results (Wu 2000). Conversely, if the results are not sensitive to costs, how could OMA be valid (Levine 2000)?

The Empirical Consequences of Costs

Before delving into the interpretation and determination of costs, it is worth noting that in instances where the analysed data shows temporal regularities, costs only have a minor effect on the results. This is hardly surprising. In terms of classification analysis, these are instances of strong clusters or stable groups; that is, groups of very similar individuals that will end up being grouped together no matter what distances and classification methods are used (Lebart et al. 2006, p. 254). Classification specialists generally seek to identify these strong clusters. By definition, a group of virtually identical sequences needs very few transformations to make them identical, and any assigned cost will have very little influence on this group. The greater the differences between the compared sequences, the greater the importance of costs.

When the data include several strong clusters, costs are important at two levels. First, costs play a role in cases that do not fit easily into stable groups. Depending on their nature, costs assigned these atypical sequences to one or another of the strong clusters, and, if need be, to a more heterogeneous group including rare sequences.

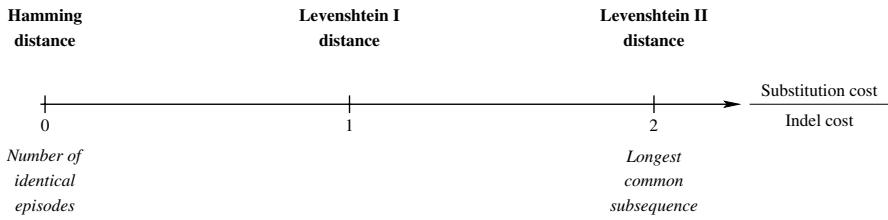


Fig. 3.1 Patterns corresponding to costs in Optimal Matching Analysis (Lesnard 2010)

Second, costs also have an impact on the distance between the strong clusters and, in the case of an ascending hierarchical classification, affect the way in which they might be aggregated. Costs therefore play a role, albeit not always a clear one since they often involve variations in the number of sequences assigned to each stable group by the classification or partitioning algorithm.

Determining costs first requires understanding and interpreting them with regard to the processed data and purpose of the analysis. In the social sciences, the order of states is not tied to a biochemical combination as it is in biology; rather, it is tied to time. When a state is inserted in, or deleted from, a sequence to make it identical to another, a time shift occurs between these two sequences. The more these two operations are used, the more the elements compared will be distant in the original sequences. Insertion and deletion operations therefore warp the position of the elements of the sequence. On the other hand, substitution operations respect the position of sequence elements, the trade-off being that states are replaced. Once insertion and deletion operations have been used, substitution operations do not introduce additional shifts; however, they only recognise the shifted positions of the states to be substituted.

In sum, insertions and deletions shift the sequences in order to identify identically coded subsequences, while substitutions do not shift the sequences but replace states with others. In the first case, the time structure of the sequences is altered, and in the second the states of which they are composed are changed (Lesnard 2010). When the substitution cost is double or more the insertion-deletion cost, substitution operations are no longer used since it is equivalent, or less costly, to insert the desired state and delete the other than it is to substitute them (Kruskal 1983). Conversely, when substitution costs are very low compared to insertion-deletion costs, the latter operations are no longer used. As a result, any cost combination falls within the distances between Levenshtein 2 and Hamming (see Fig. 3.1). In the first case (Levenshtein 2), the algorithm seeks to shift the sequences in order to compare identically coded, but not necessarily contemporaneous, subsequences. In the second, the algorithm only takes identical subsequences into account (that is, a series of identical states located in the same place in the sequences). The Levenshtein 1 distance is a sort of compromise between these two poles. The choice of costs thus depends on the time scale and its importance for the analysis. The greater the importance of respecting the time scale, the more imperative it is for costs to approach or reach the Hamming distance.

Multiple Substitution Costs

By definition the constituent states in a state space are considered to be different and to refer to situations that are not, from an analytical viewpoint, the same. Several substitution costs are generally used and defined—empirically or theoretically—to reflect the fact that some states are closer than others. Thus, in their study of inter-generational social mobility, Halpin and Chan define substitution costs according to a nomenclature of social classes and their theoretical distance (Halpin and Chan 1998). Frequencies of transitions between states, calculated on the basis of all episodes, have also been used to define substitution costs. This solution is equivalent to assigning substitution costs between states that decrease with higher levels of observed transitions. This practice has been challenged, however, because substituting a state with another is theoretically not a transition and substitution costs must reflect the proximity of states (Abbott and Hrycak 1990; Halpin 2010).

The use of only one substitution cost therefore presumes that no state is *a priori* closer than another. Using a matrix of substitution costs introduces information on states' proximity into the analysis. To set lower substitution costs for certain states is to posit that these states are close but different, allowing for the grouping of these sequences if they are sufficiently similar. In other words, if the differences between two sequences only concern certain states with low substitution costs, then the total minimum cost will be low and they will show close general proximity. In the context of sequence comparison in the social sciences, this close proximity between states can be interpreted to mean that even if a difference appears in two sequences at a given moment, the difference is almost negligible, and that if the differences are of this nature the algorithm should consider them unimportant.

Indeed, a low substitution cost is only of interest if it is much lower than the combined cost of a deletion and an insertion. A low substitution cost will cause the algorithm to not shift the sequences when looking for an identical subsequence, while a high cost will prompt a search for identical-but-shifted subsequences. Low substitution costs lead to the search for contemporaneous, identical sequences, thereby implying that the sequences are also close. *In other words, to claim that certain states are close is equivalent to positing that the presence of these states in two sequences is the marker of their proximity as a whole, and that they might belong to very similar social rhythms.* Conversely, if two sequences have two states with high substitution costs, the implication is that the sequences are distant and that it is more useful to look for identical-but-shifted sequences. This also holds true in biology, since substitution costs between two elements of a DNA sequence decrease the closer they are phylogenetically, a proximity that extends to all the sequences in which they appear.

In the time frame of OMA's use in the social sciences, substitution costs also relate to the whole sequence: Does the presence of two different states point to two very different social rhythms? With a single substitution cost, there is no reason to presume the answer to this question and the facts can speak for themselves: social rhythms appear in the typology. A matrix of substitution costs suggests the

existence, be it theoretical or empirical, of certain social rhythms—combinations of states that signal closer proximity. Typologies created with a matrix of substitution costs present two types of groups. The first includes sequences in which some parts are close but time-shifted, while the second includes sequences where identical or very similar subsequences are contemporaneous.

More precisely, since the sequences that are not part of a strong cluster are most affected by costs, the sequences that are most difficult to classify will be assigned to the strong cluster to which they come closest. A matrix of substitution costs will then adjust this allocation by emphasizing either the closest contemporaneous subsequences (low substitution costs), or shifted identical subsequences (high substitution costs). Is it legitimate to assign the sequences to strong clusters in different ways, according to the proximity defined in the matrix of substitution costs? Rare sequences are, by definition, difficult to classify. And given that the goal of taxonomy is to classify, there are only two possible solutions: either assign them to strong clusters, or relegate them all to an unclassifiable category.

There is a need here to distinguish between rare sequences with subsequences that approach strong clusters, and very rare sequences that are difficult to move toward a stable group. Empirically, what most often appears is a catchall category that cannot be interpreted. So the question is whether it is desirable to assign rare sequences that share several characteristics with strong clusters. If the goal is to find the most homogeneous groups (outside of the one that brings together the unclassifiable sequences), then all of the costs must be high in order to increase the distance between them and the rare sequences. If less homogeneous classes are acceptable, then a matrix of substitution costs allows for more flexibility in managing allocations to stable groups.

From this perspective, transitions provide partial information on social rhythms, and it therefore seems reasonable to use this information; indeed, the use of a low substitution cost in the social sciences implies that the difference between the two states does not indicate that the sequences belong to two different groups of rhythms. Using transitions to determine substitution costs is equivalent to substituting diachronic proximity for synchronic proximity, which recognizes that certain states link together more often than others over time, pointing to a greater proximity of these states. However, using transitions between states to set substitution costs raises the issue of circularity, because transitions are generally calculated from sequences that are then analysed with OMA. Insofar as this method is descriptive it is not really a problem, especially since transitions are aggregated while sequences are individuals.

Optimal Matching Methods and Social Theory of Time

A large portion of my research using OMA focuses on time-use surveys (Lesnard 2008, 2009, 2010; Lesnard and Kan 2011). These surveys feature a time diary in which respondents must describe activities over one or several days. The temporal

dimension of the sequences derived from the time-use surveys is therefore very strong. In this respect, the analysis of time use through OMA aims to go beyond the so-called “time budget” approach, which reduces time uses to durations, by reintroducing chronology into the study of daily lives. It is therefore necessary to limit time distortions and curb the use of insertion and deletion operations that emphasize identical sub-sequences—that is, in terms of time use, the time-consuming identical activities that take place at different times of the day—and focus on duration rather than timing.

This is also true for other sequences that include a time dimension. When shifts are emphasized, the similarities highlighted are a series of identical states, meaning the duration of identical subsequences is emphasized to the detriment of timing. Sequences in the social sciences most often have a time dimension, and the goal of OMA is precisely to underscore possible temporal regularities. The risk in intensively using insertions and deletions is to obtain results that are highly dependent on similarities in duration to the detriment of similarities in timing. It is perfectly legitimate to value duration over timing, but in this case solutions that are simpler and faster than OMA should be favoured, such as a cluster analysis directly and exclusively applied to duration.

Thus the analysis of time-use surveys using OMA reveals more general problems with the use of this type of method on sequences in the social sciences; these are especially important in the case of time-use surveys. Indeed, in the more specific case of research on the work hours of individuals and couples, where the point of using OMA is to identify different types of distributions of work hours, the warping that any insertion or deletion introduces into the subject of study itself is particularly problematic (Lesnard 2008). Consequently, only substitution operations should be used. The state space is highly simplified: two for individual work hours (work and nonwork), four for couples (nobody works, the two spouses work, one of the spouses works but the other does not, and vice-versa). The issue I faced was whether to use one cost—Hamming’s distance—or several. The Hamming distance effectively amounts to counting the number of discordant pairs in two sequences; any gap, even if it is minimal, increases the distance. Even though insertion-deletion operations are undesirable, I felt it was necessary to somewhat mitigate the radicalism of this parameterisation.

Toward this end, I developed a variant of this distance: the dynamic Hamming distance (DHD) (Lesnard 2010). Like the Hamming distance, DHD only uses substitution operations in order to best preserve the timing of the sequences. Its specificity is in the use of costs, which depend on transitions between states and which change over time. The distance between states varies over time and decreases when the transitions before and after the considered date are strong between the state and the one for which it is to be substituted.

For example, in the 1998–1999 time-use survey, the transition from no work to work accounted for 22 % of the working population between 8:00 and 8:10, but only 3 % between 10:40 and 10:50. With DHD the two states would be considered close at 8:00 and far at 10:40. For the vast majority of employed people, the early morning is the beginning of the workday, so it would not be very costly to make the two

sequences that contain different states identical, given that differences at that time of day are not important. Mid-morning, however, such differences indicate different working rhythms that will be reflected in the distances between their respective sequences. DHD thus takes limited gaps into account without resorting to insertions and deletions that would shift the sequences.

The transitions between the different possible states at any given moment illuminate certain aspects of social rhythms. However, the transition matrixes only provide macrosocial information by date without links. With OMA, these matrixes can be expressed in terms of distances between individual sequences, allowing for the creation of a taxonomy. In other words, by integrating transitions between states in OMA, DHD individualizes them and connects them to one another in order to better identify the different social rhythms that underlie the aggregate figures of transition matrixes.

An arithmetically identical number of years (or minutes) may refer to substantial social differences. This is a classical result from the sociology of time pioneered by Hubert. Using his research on the links between time, magic, and religion, Hubert showed that the calendar is underpinned not by a quantitative time, but rather a qualitative time made of discontinuous and heterogeneous episodes (Hubert 1905). The linearity of calendars and clocks in no way implies that time is linear. By considering sequences in their entirety, OMA, and DHD in particular, help place the various elements into their context and thus restore this heterogeneity.

Sociology of Time

The process of deriving substitution and transition costs and allowing them to vary over time also needs to be linked to the sociology of time program initiated by Hubert and Durkheim, and further developed by Norbert Elias and Pierre Bourdieu. Building on Hubert's seminal work, Durkheim outlined an agenda to develop the sociology of knowledge. For Durkheim, categories of understanding are not transcendental (that is to say innate, constitutive of human nature) but rather derive from the organization of life in society (Durkheim 1912). In an analysis of ethnographic writings and materials on the Aboriginals Durkheim showed that time in these societies is a religious symbol that separates the sacred from the profane, forming two homogeneous, alternating periods of time that punctuate the Aboriginal calendar. This alternation also reflects the ebb and flow of community life; during sacred times the group gathers, whereas profane times are more isolating, focused on individual and family survival. The homogeneity of, and conceptual difference between, the two times thus refer to distinct collective practices and temporalities, or relations to time; in other words, the time category of understanding is interdependent with the Aboriginals' binary social rhythm.

One of the consequences of this first cornerstone of Durkheim's sociology of time is that a "calendar expresses the rhythm of collective activity while ensuring its regularity" (Durkheim 1912, p. 15). The history of calendars and other systems for

determining time is also that of social rhythms and of temporality. Indeed, as Pierre Bourdieu put it, time is a symbolic system that is structured—by the collective rhythm, embodied by religion in less differentiated societies—and structuring—as a temporality or form of temporal knowledge (Bourdieu 2001, p. 201). Time, and more generally all symbolic systems, follows “logical conformism” (Bourdieu 2001, p. 204), meaning that it both accepts and validates the social order. A system for determining time, as a symbol, cannot be used if its use is not known and recognized by the whole group. In less differentiated societies that consist of small autonomous communities with a limited division of labour and somewhat interchangeable members, given that identity is not very individualised but rather dominated by the group, social order is based on respect for the group and its rhythm through religion. In these societies, to follow the collective rhythm is to alternate between the sacred and the profane, and between collective time and more individual time.

Calendars and, in general, any system to determine time such as the clock, were not intended as time-measuring instruments, but as a means to mark time. “The institution of calendars is not intended solely, and probably not even primarily, to measure time as a quantity. It does not derive from the idea of a purely quantitative time, but rather the idea of a qualitative time that is made of discontinuous and heterogeneous parts and that constantly turns on itself” (Hubert 1905). Calendars crystallize and stabilize collective rhythms and activities: “Units of time are not units of measurement but of a rhythm where oscillation between alternatives periodically leads back to the similar” (Hubert 1905).

It is the tools developed in relation to the social symbol of time that bear the traces of shared transformations in the organization of society and time; that is, of collective rhythms and temporalities. To paraphrase Hubert, the history of calendars reflects the evolution of the “code of the qualities of time”¹, the direction and degree of diversification of collective life, and also the “level of synthesis” as defined by Elias (1991) of the symbol of time.

In this respect, the appearance of bell clocks in European towns in the Middle Ages was a key step in the process of social and temporal differentiation. The different rhythms of urban collective life in the Middle Ages² created synchrony issues throughout the day that were resolved by using multiple acoustic and optical signals: bands, flags, pennants, and especially bells. At first canonical hours—some marked by ringing church bells—were used as time references in several areas of collective life; for example, the market’s opening coincided with the bell sounding the end of mass, while the end of the workday for day labourers was marked by the bell for the compline.

¹ “The calendar is the periodic order of rites. Its history also teaches us that it is the code of the qualities of time. The first calendars were almanacs that recorded, day by day, magico-religious forecasts and prescriptions” (Hubert 1905).

² Meetings of bourgeois councils or courts to deal with the affairs of the town, market gatherings, the beginning and end of work for day labourers, the opening and closing of city gates, and more unusual events such as gatherings to respond to threats of fire or of war (Landes 2000, p. 76; Dohrn-van Rossum 1992, p. 206).

Beginning in the fourteenth century, the development of cities created the need for a wider variety of time signals that church bells alone could no longer handle. Cities acquired bells to convene meetings, indicate payment due dates for interest and taxes, assemble neighbourhoods, set market hours, etc. This inflation of urban signals culminated with work bells that set working hours for salaried workers. The multiplication of urban signals also reflected the increasing division of labour since a number of them were no longer aimed at the community as a whole, but rather at particular groups (employees, members of various councils, travellers, merchants, etc.).

The increasing need to synchronize certain groups fed into a complex system of acoustic and optical signals in big cities that probably reached its climax at the beginning of the fourteenth century. Growth in the number of acoustic signals could, in large part, be attributed to the coordination of the needs of certain groups, as opposed to the urban communities as a whole. As long as collective life was punctuated by a series of events involving the entire community, a single signal could ensure coordination, as was the case in the sounding of canonical hours. The proliferation of bells testified to the division of labour, and therefore the pluralisation of time. In the towns of the Middle Ages, time was no longer alternating between the sacred and the profane, but was differentiated with the development of functional interdependence. However, when the bells started sounding the temporal rhythms for certain segments of the population in addition to those for the community as a whole, they became less effective at finely synchronising subgroups of the population. Today, it is hard to imagine the background noise that the bell signals created on a daily basis: "Everyday life was temporally structured through and through by bell signals, with hardly a day resembling another. Apart from signals for proclamations, prohibitions, and ordinances, the inhabitants of the city also received a wealth of acoustic information about important public civic events" (Dohrn-van Rossum 1992, p. 217).

The introduction of bell clocks in the fourteenth century unified and simplified the various acoustic signals. The clock bells did not chime for anyone in particular. Rather, different groups or segments of the population could use the signal to synchronize. Because clocks can only ring for *equal* hours without introducing too many technical complications, the system of canonical hours, and attendant hold of religion over time, collapsed, liberating other collective times from their ties to religious time (Landes 2000, p. 81).

Bell clocks substituted a regular and neutral tempo for the profusion of specific and muddled signals. However, hours did not mark any particular event; it was up to the local community's members to associate them to collective or private activities they wished, or were required, to attend. Before bell clocks were introduced, the relationship to time was theoretically still largely an external constraint that elicited poor anticipation. Everyone knew that it was time for a lunch break when they heard the work bell ring a second time. One only needed to recognize the sound of different bells to situate oneself in time. With bell clocks, hours had to be linked to events; time constraints were no longer explicitly marked by a particular bell, but

rather implicitly signalled. Each hour chime only evoked the constraints that one chose to associate with it.

Specific time constraints therefore had to be internalized by each individual. The introduction of the bell clock marked the beginning of the internalization of time constraints, or at least accelerated the process. With the clock striking hours, people had to get into the habit of adjusting their behaviour in response to the signal. The regularity of hours also allowed for a greater anticipation of the day's events, whereas the previous system was characterized by very low predictability. Thus the new system provided a greater opportunity to regulate one's behaviour, that is, to self-regulate. *Paradoxically the homogeneity of the movement of the clock's hands was precisely the sign of the heterogeneity of time*: autonomous social areas emerged along with their particular collective rhythms.

In the everyday life of individuals involved in a number of fields, time at the individual level alternated between the different times in various social fields (Halbwachs 1947, p. 167; Bourdieu 2003, p. 202). This meant that the rhythm of daily life was determined in large part by the rhythms associated with positions held in different social fields. Sequences of participation in social fields could come into conflict due to inconsistencies in the rhythms associated with successively held positions.

Over the longer run, life courses can also be analysed as alternating participation in the principal social fields, which are the economic field and the family field (Bourdieu 1994, pp. 135–145). In contemporary societies, adult status is achieved through economic independence from one's original family and the formation of a so-called childbearing family. It therefore involves two forms of integration, in the productive and in the married and parental worlds, and is generally marked by events such as the end of studies, the first job, leaving the parental home, the first relationship and/or the first marriage, and the birth of a first child (Modell et al. 1976). The study of early adulthood actually comes down to analysing access to the economic fields (directly or indirectly through marriage) and family field (with a new status, through the formation of a so-called childbearing family), as well as their interlinks.

A social field's time has two dimensions since it is both the rhythm of collective life for the group of agents involved, and a relationship to time and to the future in the field (temporality). Using all the transitions as substitution costs is a way to break down the average rhythm they represent into homogeneous collective sub-rhythms, and thereby describe in detail the collective life of the field under consideration. With DHD the structure of the state space (distance between the states) is no longer static, but bends to the rhythm of collective life. In theory this adaptation of Hamming's distance goes further since the temporality of one field flows from the field's repeated exposure to the collective rhythm. The use of transitions allows for collective rhythms to be described in ways that make sense in terms of a field's temporality. The differences between states, which are variable over time, have a meaning for the agents in the given field. To work or not at 9:00 a.m. is not socially very significant, while the opposite would be true for 11:00 p.m. Similarly, to transition from unemployment to employment at age 20 or 25 does not have the same significance as a first entry into the labour market at age 30.

This is what makes OMA, and DHD in particular, a type of descriptive method that is especially well suited for sequences in the social sciences. The mean is also descriptive and could be used to describe these sequences, but it does not take into account the heterogeneity of time and therefore imposes the linear framework of the clock on time; as a result, the collective rhythms disappear. DHD is a method for describing sequences that incorporates the heterogeneity of time in its very design, drawing on the sociological theories of Durkheim, Elias, and Bourdieu. It includes key elements from each of these authors. DHD centres on Durkheim's dual concept of time as it was reworked in the research of Elias and Bourdieu on social differentiation and the concept of fields.

Strictly speaking, the use of DHD implies the adoption of this general theoretical framework. This means considering the state space—a little-discussed cornerstone of OMA—in this framework. Discussion on the use of OMA in the social sciences has focused on transformation and cost issues, and has relatively ignored the state space which, unlike in biology and computer science, is not given, but rather is a matter of choice. Field theory can help guide this choice; the state space must represent the simplified structure of a field. The different states should be able to capture the collective rhythm of the field, or at least one of its aspects.

More generally, the inclusion of DHD in the theoretical basis of the sociology of time illustrates the links that can be made between theory and optimal matching methods. In the same way that regression models cast the functioning of society in terms of a “general linear reality” that is at odds with most sociological theories (Abbott 1988), the choice of costs for OMA is never methodologically neutral. In particular, the use of insertions and deletions focuses the analysis on duration and therefore ignores similarities in timing, but also more generally the heterogeneity of time. This may be perfectly legitimate, even if in this case it would be easier to use a simple cluster analysis on the durations. But it must be justified. Although the effect of costs on the results is often subtle, the use of OMA in sociology must be made in a rigorous methodological framework so that it no longer appears as the black box disparaged by some of its critics.

References

- Abbott, A. (1988). Transcending general linear reality. *Sociological Theory*, 6(2), 169–186.
- Abbott, A. (2000). Reply to Levine and Wu. *Sociological Methods and Research*, 29, 65–76.
- Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence analysis: An optimal matching analysis of musicians' careers. *American Journal of Sociology*, 96(1), 144–185.
- Bourdieu, P. (1994). *Raisons pratiques: sur la théorie de l'action*. Paris: Seuil.
- Bourdieu, P. (2001). *Langage et pouvoir symbolique*. Paris: Seuil.
- Bourdieu, P. (2003). *Méditations pascaliennes* (revised edition, 1st ed. 1997). Paris: Seuil.
- Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5(Suppl. 3), 345–352.
- Desrosières, A. (1993). *La politique des grands nombres. Histoire de la raison statistique*. Paris: La Découverte.
- Dohrn-van Rossum, G. (1992). *Die Geschichte der Stunde*. Munich: C. Hanser.

- Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.
- Durkheim, É. (1912). *Les formes élémentaires de la vie religieuse*. Paris: Alcan.
- Elias, N. (1991). *The symbol theory*. London: Sage.
- Halbwachs, M. (1947). La mémoire collective et le temps. *Cahiers Internationaux de Sociologie*, 2, 3–31.
- Halpin, B. (2010). Optimal matching analysis and life course data: The importance of duration. *Sociological Methods & Research*, 38(3), 365–388.
- Halpin, B., & Chan, T. W. (1998). Class careers as sequences: An optimal matching analysis of work-life histories. *European Sociological Review*, 14(2), 111–130.
- Hamming, R. W. (1950). Error-detecting and error-correcting codes. *Bell System Technical Journal*, 29(2), 147–160.
- Hubert, H. (1905). Étude sommaire de la représentation du temps dans la religion et la magie. *Annuaire de l'École Pratique des Hautes Études, section des sciences religieuses*, 1–39. doi:10.3406/ephe.1904.19635.
- Kruskal, J. B. (1983). An overview of sequence comparison. In D. Sankoff & J. B. Kruskal (Eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (pp. 1–44). Reading: Addison-Wesley.
- Landes, D. S. (2000). *Revolution in time. Clocks and the making of the modern world*. Cambridge: The Belknap Press of Harvard University Press.
- Lebart, L., Piron, M., & Morineau, A. (2006). *Statistique exploratoire multidimensionnelle: visualisation et inférence en fouilles de données*. Paris: Dunod.
- Lesnard, L. (2008). Off-scheduling within dual-earner couples: An unequal and negative externality for family time. *American Journal of Sociology*, 114(2), 447–490.
- Lesnard, L. (2009). *La famille désarticulée. Les nouvelles contraintes de l'emploi du temps*. Paris: PUF.
- Lesnard, L. (2010). Cost setting in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3), 389–419.
- Lesnard, L., & Kan, M. Y. (2011). Investigating scheduling of work: A two-stage optimal matching analysis of workdays and workweeks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 349–368. doi:10.1111/j.1467-985X.2010.00670.x.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.
- Levine, J. H. (2000). But what have you done for us lately? Commentary on Abbot and Tsay. *Sociological Methods and Research*, 29(1), 34–40.
- Modell, J., Furstenberg, F. F., & Hershberg, T. (1976). Social change and transitions to adulthood in historical perspective. *Journal of Family History*, 1(1), 7–32.
- Stovel, K., Savage, M., & Bearman, P. (1996). Ascription into achievement: Models of career systems at Lloyds Bank, 1890–1970. *American Journal of Sociology*, 107(2), 358–399.
- Wu, L. L. (2000). Some comments on “Sequences analysis and optimal matching methods in sociology: review and prospects”. *Sociological Research and Methods*, 29(1), 41–64.

Chapter 4

Distance, Similarity and Sequence Comparison

Cees H. Elzinga

Introduction

Over the last decades, sequence analysis has developed from a fiercely debated trick from computational biology—and I was one of the discussants too—into a broadly accepted toolbox for those interested in classifying all sorts of career data. Recently, with the introduction of ANOVA-like techniques to explain distances in terms of one or more categorical covariates (Bonetti et al. 2013; Studer et al. 2011), sequence analysis tools can be used to test hypotheses about causal relations.

This very book is crammed with state-of-the-art applications of sequence analysis and all chapters have in common that they start by somehow constructing distances and/or similarities from sequence data. Therefore, it seems justified that this paper does not deal with data, not even with simulated data, but instead revisits the fundamental concepts used—distance and similarity—in order to stress the importance of the axiomatic foundations of these concepts. From these axiomatic foundations, we will try to clarify some principles and common misunderstandings pertaining to transforming and normalizing our numerical basis and make some remarks on the relation between the concepts of similarity and distance. None of the ideas in this chapter is new, nor are their interrelations. The purpose of this paper is not to develop new maths or methodology. Instead, its purpose is to make the mathematical concepts accessible and intelligible to social scientists. Therefore, the tone is informal, definitions are sometimes replaced by small graphs and proofs are omitted. On the other hand, since the subject matter is formal and abstract, it is inevitable that I use some formulas and inequalities: I tried to restrict myself to a level of abstractness that allows me to formulate some proposals that are directly applicable to problems of sequence comparison.

We start out in the next two sections to discuss the formal basis of distance and similarity. In Sect. 4, we will discuss the transformations that, respecting the axiomatic basis, may be applied to distance and similarity. In Sect. 5, we discuss the

C. H. Elzinga (✉)
VU University Amsterdam, Amsterdam, The Netherlands
e-mail: c.h.elzinga@vu.nl

relation between the two kinds of measures and deal with the issue of normalization. In Sect. 6, we concisely discuss the fundamental difference between the two concepts as, in practice, partitioning a set of sequences on the basis of similarity may not yield the same partitioning when one would use a distance measure instead. Finally, we summarize in Sect. 7.

Distance

In this section, we will deal with the concept of “distance” and some of the abstract properties of the various measures that we use to evaluate distance. All of us know distances as numbers that refer to the relative location of objects in some space. In this chapter, we do not deal with objects of our everyday lives, located in the physical space that surrounds us, but with quite different objects—sequences—that have no physical location. Hence, we need a definition of “space” that is quite general and that allows us to evaluate distances between sequences according to the same principles that we use in considering distances in our everyday lives.

Axioms of Distance

We shall say that a space consists of a set of objects that has some structure, i.e. a set of rules that govern the relations between the objects. Such relations could be, for example, relations of order or adjacency. In the present context, our space will consist of a set of sequences and the structure will be determined by some distance measure that is defined on all pairs of sequences from that set. However, it is not at all clear what a distance between objects is like, not even in our everyday life. This is illustrated by the two panels of Fig. 4.1. In the left panel, we define the distance between locations a and b as the length of the straight line between these locations. Using the coordinate vectors $\mathbf{a} = (a_1, a_2)$ and $\mathbf{b} = (b_1, b_2)$, the Pythagorean formula yields

$$d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \quad (4.1)$$

In the right panel however, there is an obstacle, say a building, between a and b . For you, bound to walk the streets, the above calculation of distance is not very relevant so you would rather use

$$d(a, b) = |a_1 - b_1| + |a_2 - b_2|, \quad (4.2)$$

calculating the length of the walking route from a to b . Another, frequently used way of calculating distance is according to the rules of spherical geometry in which shortest distances are not straight lines but curves on the surface of a slightly flat-

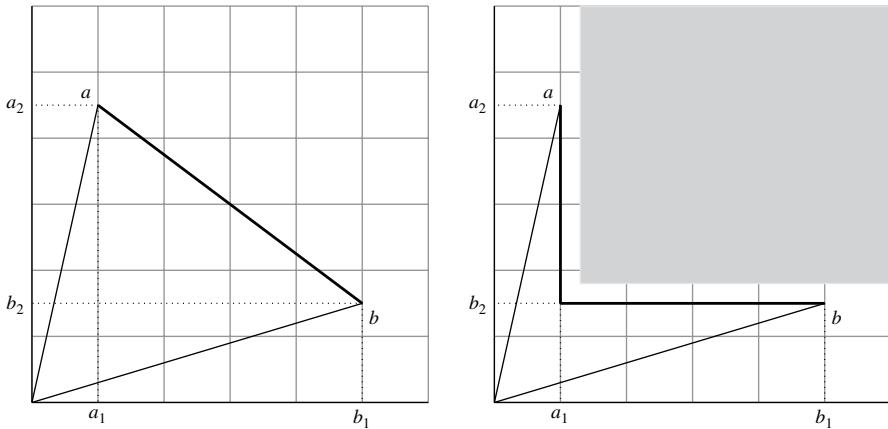


Fig. 4.1 In the left panel, the distance $d(a, b)$ is calculated according to the Pythagorean formula $d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$ and in the right panel, distance is calculated according to $d(a, b) = |a_1 - b_1| + |a_2 - b_2|$

tended sphere. This is what navigators of ships and planes do when they have to cross long distances, as for example between the ports of Boston and Rotterdam.

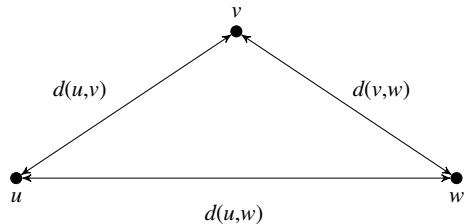
So, even in our daily lives, distances are calculated in different ways, depending on their intended use and the properties of the space. Therefore, instead of listing all possible measures of distance, we characterize a function on pairs of elements of a set $X = \{u, v, w, x, \dots\}$ as a distance measure through a few very general properties, also called “axioms”. However, these axioms still correspond to our intuitive “living” of distances among cities, buildings and objects in kitchens and offices. Below, we first list the axioms and then comment on them. We write $d(u, v)$ for the distance between objects u and v and say that d is a distance if, for all triples u, v, w from the set $X = \{u, v, w, x, \dots\}$, it is true that

- $D1 : d(u, v) = 0$ if and only if $u = v$,
- $D2 : d(u, v) > 0$ if and only if $u \neq v$,
- $D3 : d(u, v) = d(v, u)$,
- $D4 : d(u, w) \leq d(u, v) + d(v, w)$.

A function d that satisfies all four of the above axioms is called a *metric* or, equivalently, a *metric distance* and the pair (X, d) is called a *metric space*. Today, the concept of a metric space is over a hundred years old as it was first coined by Fréchet (1906), at the beginning of the twentieth century, in the context of metricizing the constructive Euclidean geometry.

Axiom D1 says that no two distinct objects can be on the same location and Axiom D2 says that distinct objects must be in different locations. From Axioms D1 and D2, it is immediate that distances cannot be negative. Axiom D3 states that

Fig. 4.2 An illustration of the triangle inequality (Axiom D4 in the main text)



distances are symmetric: u is as remote from v as v is remote from u . Clearly, the first three axioms correspond to our intuitions about space and distance. The fourth axiom is called the “triangular inequality”, since it pertains to three objects, and it expresses our intuition that “a detour always takes more time”. Axiom D4 says that, in order that some measure is to be called a “distance”, it should always result in the conclusion that going directly from u to w yields a distance $d(u, w)$ that does not exceed the distance $d(u, v)$ that results from first visiting v plus the distance $d(v, w)$ that results from subsequently traveling from v to w . This is illustrated in Fig. 4.2. A slightly different interpretation is that when two objects (u and w) are close to a third object (v), they must be close to each other, i.e. $d(u, w)$ must be small.

So we see that the triangular inequality fits within our intuitions about space and distance too. Furthermore, given the objects, the triangular inequality acts as a sort of boundary on the numerical values of the metric: $d(u, w)$ is bounded by all pairs $(d(u, x), d(x, w))$, i.e. Axiom D4 must hold for all triples of objects of X . So, Axiom D4 ensures that the metric space is “smooth in all directions” in the sense that if we know the properties of a subspace of a metric, i.e. d on a subset of X , we can be sure that in overlapping sets, these properties will be quite similar. Therefore, Axiom D4 is a very important axiom: if it is not satisfied, it is very risky to generalize from our measurements since we are not sure that the measurements would not be wildly different, had we observed only slightly different objects. So we should avoid the use of “proximities” or “dissimilarities” that are not proper metrics in the sense that they do not satisfy all four axioms D1–D4.

Clearly, the axioms D1–D4 do not prescribe a specific way of measuring distances; the axiom system only limits our freedom of choice of a particular method of gauging distances. In practice, constructing measures that satisfy the first three axioms rarely appears to be a problem; however, constructing measures that also satisfy the triangular inequality is a bit harder. We will come back to this problem in the subsection on normalizing a metric.

OM-Metrics

In all of the chapters of this book, some variant of Optimal Matching (OM) is used to determine distances between sequences. Here we concisely discuss those variants

of OM that generate proper metrics in the sense that the numbers produced satisfy the axioms D1–D4. Formally, let $x = x_1 \dots x_n$ and $y = y_1 \dots y_n$ denote two n -long state sequences over the alphabet of states $\Sigma = \{\lambda, a, b, \dots\}$ with λ denoting the empty state and let $e = e_1 \dots e_k$ denote a series of admissible sequence edits such that $e(x) = e_k(e_{k-1}(\dots e_2(e_1(x))\dots)) = y$. For any pair of sequences, there may exist many distinct series of edits that transform x into y and we write $E(x, y)$ to denote the set of such edit-series. Furthermore, to each edit e_i , a nonnegative cost or weight $c(e_i)$ is assigned and the cost of an edit-series $C(e)$ equals the sum of the costs of the edits involved: $C(e) = \sum_i c(e_i)$. The OM-distance $d_{OM}(x, y)$ between a pair of sequences x and y is the minimum of the costs of the edit-series in $E(x, y)$:

$$d_{OM}(x, y) = \min\{C(e) : e \in E(x, y)\}. \quad (4.3)$$

Let us now denote the edit-costs with respect to the characters or states of the alphabet $\Sigma = \{\lambda, a, b, c, \dots\}$ as a symmetric array \mathbb{C} over all pairs of states. In this array, $\mathbb{C}(a, b)$ denotes the cost of substituting a for b , $\mathbb{C}(\lambda, a)$ denotes the cost of deleting a (and substitute it for the empty state λ), and $\mathbb{C}(a, \lambda)$ denotes the cost of inserting state a . With this notation, the “standard” cost matrix is of the form

	λ	a	b	c	\dots
λ	0	1	1	1	\dots
a	1	0	2	2	\dots
b	1	2	0	2	\dots
c	1	2	2	0	\dots
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots

A proof that d_{OM} is a proper metric, provided that the cost-matrix itself is a metric can be found in e.g. Yujian and Bo (2007). The condition that the array \mathbb{C} constitutes a metric means that \mathbb{C} , understood as a function on pairs of states, satisfies the axioms D1–D4. The reader easily verifies that the axioms indeed hold for the standard cost function as shown above. However, it is not difficult to construct a cost-matrix that violates the triangular inequality D4 as demonstrated in the array below;

	λ	a	b	c
λ	0	1	1	1
a	1	0	1.5	4
b	1	1.5	0	2
c	1	4	2	0

In this array, we have that $4 = \mathbb{C}(a, c) > \mathbb{C}(a, b) + \mathbb{C}(b, c) = 1.5 + 2 = 3.5$, a violation of axiom D4. So, if we do not properly set the edit-costs, the OM-algorithm

will generate numbers that are not proper distances. The condition of a metric cost matrix also implies that OM-variants that use a dynamic, data-driven cost-matrix (Halpin 1950; Hollister 2009; Rohwer and Pötter 1999) will not automatically generate a proper metric over the pertaining sequences.

Rather than directly using the metric properties of a sequence-space, Massoni et al. (2009) tried to utilize the topological structure of an OM-generated sequences space through exploring Kohonen-maps. Perhaps this will turn out to be a seminal approach.

Subsequence-Based Metrics

Once we can represent sequences as vectors in a vector-space, we can use a whole family of proper distance metrics of which the examples in Eqs. 4.1 and 4.2 are just special cases. Let us suppose that we can represent sequences x and y as vectors, i.e. as coordinate arrays¹ $\mathbf{x} = (x_1, x_2, \dots)$ and $\mathbf{y} = (y_1, y_2, \dots)$, it is easy to calculate distances through the Pythagorean formula

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (4.4)$$

$$= \mathbf{x}'\mathbf{x} + \mathbf{y}'\mathbf{y} - 2\mathbf{x}'\mathbf{y}, \quad (4.5)$$

wherein $\mathbf{x}'\mathbf{y} = \sum_i x_i y_i$. But how to construct sequence representing vectors? Well, we can do this anyway we like as long as the procedure results in equally long arrays of numbers, one for each sequence. We may select any number of distinct, quantifiable features of the sequences that we consider as relevant, say their length L , their number of distinct states N and the number of spells S , and use the values of these features as the coordinate values of the vectors. In this example we have three features $L(x)$, $N(x)$ and $S(x)$ which can be used to represent each sequence x as a 3-dimensional array

$$\mathbf{x} = (L(x), N(x), S(x)).$$

For example, for the toy-sequence $x = aababbca$, this would yield $\mathbf{x} = (8, 3, 6)$. Obviously, most choices of coordinate-systems will not lead to relevant, useable representations. However, in 2003, I argued (Elzinga 2003, 2005) that using the subsequences as features would generate meaningful results and that idea has been successfully applied in several contexts, e.g. in Berghammer (2010), in Fasang (2010), and in Manzoni et al. (2010). Therefore, we begin with elaborating on the

¹ Please note that I write x, y for sequences, x_1, y_1 for the states of the sequences, \mathbf{x}, \mathbf{y} for the representing vectors and $\mathbf{x}_1, \mathbf{y}_1$ for the coordinates of the vectors.

concept of “subsequence”. For a more formal treatment, the reader is referred to e.g. Crochemore et al. (2007) or Elzinga et al. (2013).

Consider the toy-sequence $x = x_1x_2x_3x_4 = abac$ over the state alphabet $\Sigma = \{\lambda, a, b, c\}$. We may take any nonnegative number of states from x and we will then be left with a subsequence of x : a subsequence u of states that have the same order in x and we will write $u \sqsubseteq x$ to denote such fact. For example, when we take out the a 's from x , we will be left with $u = bc$, one of the five 2-long subsequences of x . At most, we can take away all states from x and we will then be left with the empty sequence λ . We might also take the smallest nonnegative number of states from x , zero states, and we would be left with x itself and hence we conclude that $x \sqsubseteq x$. The reader easily verifies that x has 13 distinct subsequences, including λ and x itself.

Now we will use the concept of subsequence to construct a vector-representation \mathbf{x} for the sequence x . We do this by defining coordinates that correspond to all possible sequences that can be constructed from the alphabet Σ by setting those coordinates to 1 that correspond to sequences that occur as a subsequence in $x = abac$

$$\begin{array}{ccccccccccccc} u : & \lambda & a & b & c & aa & ab & \dots & cc & aaa & \dots & aba & \dots \\ r(u) : & 0 & 1 & 2 & 3 & 4 & 5 & \dots & 12 & 13 & \dots & 16 & \dots \\ x_{r(u)} : & 1 & 1 & 1 & 1 & 1 & 1 & \dots & 0 & 0 & \dots & 1 & \dots \end{array}$$

Formally, from Σ , we construct the set Σ^* of all sequences that are constructible from Σ and we fix the order of the elements of Σ^* , say in lexicographical order. Then we map the ordered sequences to the nonnegative integers \mathbb{Z}^* , i.e. each sequence $u \in \Sigma^*$ is mapped to an integer $r(u) \in \mathbb{Z}^*$ and we use these integers to index the coordinates of the vectors. So, for each sequence x , we construct a binary vector $\mathbf{x} = (x_1, x_2, \dots)$ such that

$$x_{r(u)} = \begin{cases} 1 & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise.} \end{cases} \quad (4.6)$$

This construction characterizes strings by their subsequences and the resulting vectors are also called “feature vectors”, the subsequences being treated as features of the sequence. However, since the number of sequences that can be constructed from even a small alphabet is countably infinite, i.e. as big as the size of the set of non-negative integers, actually constructing the vectors and calculating their products as required in Eq. 4.5 is not feasible. Therefore, one needs special methods, called “kernels” (see e.g. Schölkopf and Smola 2002; Shawe-Taylor and Cristianini 2004) to calculate the quantities appearing in Eq. 4.5.

Once the principle of assigning feature vectors to sequences is understood, it is easy to generalize and construct representations that are far more sophisticated than the simple binary vector as sketched above. For example, we may want to account for the length, the duration or the embedding frequency or a combination of such properties. This accomplished by a generalized representation:

$$x_{r(u)} = \begin{cases} f(u, x) & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise} \end{cases}. \quad (4.7)$$

Provided a suitable kernel function can be found, any specification of the function f in the above representation may be used (e.g. Elzinga and Wang 2011).

Axioms of Similarity

The concept of metric distance is an old, well established concept in all sciences and there is no debate about its definition or usefulness. However, although the concept of similarity is widely used in many branches of science and engineering, especially in biological taxonomy, in chemistry and in psychology, a widely accepted definition is still lacking and there is an abundance of authors (see e.g. Batagelj and Bren 1995; Gower 1971; Gower and Legendre 1986; Holliday et al. 2002; Wang 2006) proposing quite different, application-specific quantifications of similarity.

Because of the widely different applications of similarity, it is not very sensible to propose another similarity measure for comparing sequences in the social sciences. Rather, it is much more interesting to formulate a set of intuitive axioms that quantifications of similarity should adhere to, irrespective of the application. Therefore, we will discuss a proposal recently made in Chen et al. (2009) and generalized in Elzinga et al. (2008) and in subsequent subsections discuss normalization and some relations between a similarity and a distance. Finally, we will mention some similarities that could be used in combination with some of the well-known distance metrics for sequences like OM and subsequence-based distance.

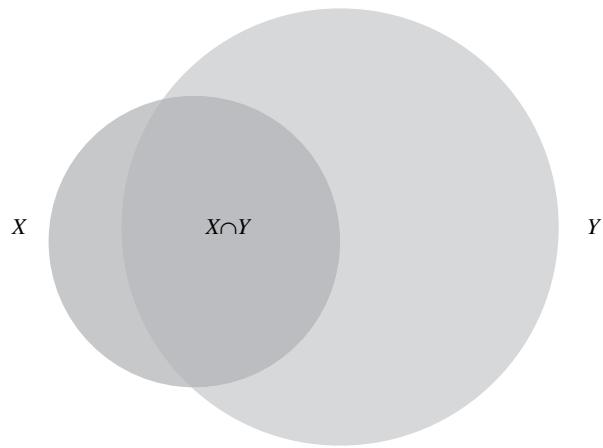
Similarity

Intuitively, two objects are similar if they share one or more features or properties and similarity seems to increase with an increasing number of such shared features. What relevant features are and whether or not all relevant features are equally important depends on the application area and the purpose of the comparison. Let X denote the set of features or properties possessed by some object x and let Y denote the set of features of object y . When we interpret similarity $s(x, y)$ as the amount of common, shared features, we are inclined to imagine

$$s(x, y) = |X \cap Y|, \quad (4.8)$$

i.e. we interpret similarity as the size $|X \cap Y|$ of the intersection $X \cap Y$ of the feature sets X and Y . With this interpretation of similarity, the next axioms directly follow from Eq. 4.8:

Fig. 4.3 When we assign to sequences x and y associated sets of features X and Y , $s(x, y)$ could be interpreted as the (weighted) number of common, shared features, i.e. as if $s(x, y) = |X \cap Y|$



$$\begin{aligned} S1 : & s(x, y) \geq 0, \\ S2 : & \min\{s(x, x), s(y, y)\} \geq s(x, y), \\ S3 : & s(x, y) = s(y, x), \end{aligned}$$

$S1$ follows because a number of common features, a count, cannot be negative. $S2$ follows because there can be no more shared features than possessed by either of the objects and $S3$ follows from the symmetry of intersection $X \cap Y = Y \cap X$. These principles are illustrated in Fig. 4.3. The reader notes that the first three axioms are highly similar to the distance axioms $D1 - D3$. Only, $D1$ says that the distance $d(x, x)$ is minimal while $S2$ states that the similarity $s(x, x)$ is maximal. So, it seems that distance and similarity are opposite counterparts. In one of the next subsections, we scrutinize the relation between distance and similarity.

What is lacking from the similarity axioms is an axiom that bounds the similarity function like the triangle inequality $D4$ bounds the distances. However, when we look at Fig. 4.4, the axiom

$$S4 : s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$$

seems inevitable and we will call this inequality the “covering inequality” because when $s(x, z) = 0$, $s(x, y) + s(y, z)$ cannot exceed $s(y, y)$. Again, the reader notes that the direction of the covering inequality is opposite to that of the triangle inequality. Finally, if we operationally define object equality $x = y$ if and only if $|X| = |Y|$, i.e. when we define object equality through equality of feature sets, we must have that

$$S5 : s(x, x) = s(y, y) = s(x, y) \text{ if and only if } x = y$$

and this axiom connects similarity to object-equivalence just like $D1$ connects distance to object-equivalence. The reader notes that, just like the distance axioms do not prescribe how to measure, how to establish distances in space, the similarity

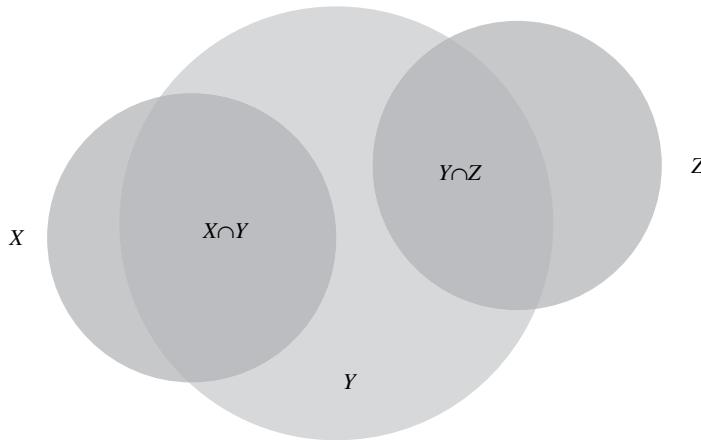


Fig. 4.4 Similarity interpreted as the size of common subsets of features and self-similarity $s(x, x) = |X|$. In the Venn-diagrams below, we have that $s(x, y), s(y, z) > 0$ but $s(x, z) = 0$ violates $s(x, y) + s(y, z) \leq s(x, z)$ but the general rule $s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$ will always hold

axioms do not prescribe how to establish similarity. Also, it is important to note that $s(x, y) = X \cap Y$ is an interpretation of s but there is no need to actually construct a measure s through comparisons of feature sets. Actual measuring systems should only adhere to the axioms in order that they yield metric distance or similarity.

On the Wrong Track

In several papers (e.g. in Bras et al. (2010) and in Elzinga and Liefbroer (2007)) we used a subsequence-based vector-representation of life course sequences and proposed to use

$$s(x, y) = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x} \cdot \mathbf{y}'\mathbf{y}}} \quad (4.9)$$

as a similarity measure. This $s(x, y)$ satisfies the similarity axioms S1–S3 and S5 and it has the nice, additional property that its numerical value is easy to interpret since $0 \leq s(x, y) \leq 1$ (actually, $s(x, y)$ evaluates the cosine of the angle between the representing vectors \mathbf{x} and \mathbf{y}). Unfortunately, as a general measure of similarity between vectors, $s(x, y)$ does not satisfy the covering inequality S4. This is easily demonstrated with a toy example²: suppose that we have vectors $\mathbf{x} = (0, 1)$, $\mathbf{y} = (1, 0)$ and $\mathbf{z} = (1, 1)$. Then the matrix of similarities according to Eq. 4.9 is given by

² This example was suggested to me by Matthias Studer through personal communication.

	x	y	z
x	1		
y	0	1	.
z	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	1

According to the covering inequality, we should have that

$$s(x, y) + s(z, z) \geq s(x, z) + s(z, y)$$

but we observe that

$$0 + 1 < \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2} \approx 1.41,$$

a clear violation of the covering inequality. Of course, this is only a toy-example. But the example demonstrates that we cannot be certain that $s(x, y)$ calculated as defined by Eq. 4.9 for real-life vector-products satisfies S4.

Back on Track Again

We began our thinking about similarity from the intuition that similarity should be proportional to the size of the set of common features. This invites to define

$$s(x, y) = |X \cap Y| \quad (4.10)$$

wherein X and Y denote the sets of features of the objects x and y . Indeed, it is true that this $s(x, y)$ satisfies the similarity axioms S1-S5 (for a proof, see Chen et al. (2009) or Elzinga et al. (2008)). Let us now look back at the vector representation of Eq. 4.6, which we repeat here for convenience:

$$\mathbf{x}_{r(u)} = \begin{cases} 1 & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise} \end{cases}. \quad (4.11)$$

When we say that the subsequences of x constitute the feature set X of x , then we have that

$$\mathbf{x}'\mathbf{x} = \sum_i x_i^2 = |X|, \quad (4.12)$$

$$\mathbf{x}'\mathbf{y} = \sum_i x_i y_i = |X \cap Y| \quad (4.13)$$

and hence that the vector-product itself satisfies the similarity axioms S1-S5. So, we could set $s(x, y) = \mathbf{x}'\mathbf{y}$ in order to obtain a proper similarity for a vector-space. However, we then have two practical problems. The first is that the numerical value of this $s(x, y)$ is hard to interpret: if it would equal 712340762134, would that mean that the objects x and y are very much alike? We simply wouldn't know unless we knew the value of s of all pairs of objects. The second problem is that we may encounter very many pairs of objects x and y for which $s(x, x) \neq s(y, y)$, simply because $|X| \neq |Y|$. This would be counterintuitive since we are inclined to think that all objects are equally similar to themselves.

To remedy these problems of interpretation, we would like to see that s is tightly bounded, preferably by 0 and 1, and that $s(x, x) = s(y, y)$ for all pairs of objects. But this was exactly the purpose of the definition of $s(x, y)$ in Eq. 4.9, the “cosine-similarity”. Apparently, not all normalizations of a proper similarity generate a proper normalized similarity. Therefore, in the next sections, we will turn our attention to properly transforming and normalizing distances and similarities.

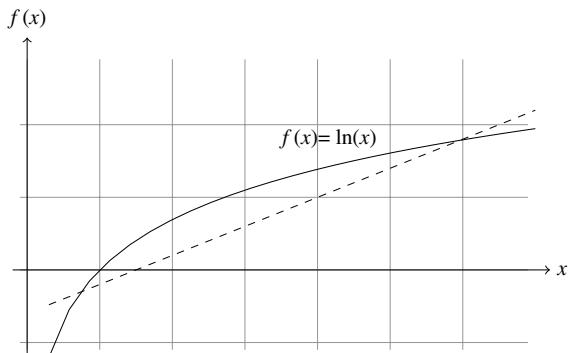
Units and Transformations

Distances are not dimensionless numbers. Instead, distances are expressed in terms of some standard unit of length. When we consider distances in our physical environment, we always mention the unit of length that the numbers refer to: kilometers, nautical miles, lightyears or Ångströms and what not.

When we calculate distance between sequences, we also have a unit of distance although we tend not to mention it. However, even the simple Hamming-distance (Hamming 2010) has a unit: it counts the number of positions in which two sequences have unequal states. Hence, Hamming's unit of distance is “position”. Similarly, the unit of OM-distances, when indel cost is set to 1 and substitution cost is set to 2, refers to the number of “edits”. When the OM-distance $d_{OM}(x, y) = 10$, it means that the minimum number of edits to change x into y equals 10. So, comparison of distances is straightforward, even if the pertaining pairs of sequences are defined over different state alphabets.

However, there are at least two kinds of problems that warrant choosing a different scale for our distances. First, when the distances are very big numbers, it may be advisable to chose a bigger unit of distance. This will facilitate the interpretation of e.g. averages and standard deviations within clusters or the averages and standard deviations of distances to centroids or medoids of clusters or to particular “prototypes”. Second, certain sequences may be extremely remote from other sequences in the data and such extreme data may heavily influence the subsequent clustering or discrepancy analysis of the sequences. In such cases, we would be well-advised to apply a compressive but order-preserving transformation, e.g. a logarithmic transformation, to the distances as calculated before further analysis. Hence, we have to talk about admissible transformations $f(\cdot)$ of the form $d'(x, y) = f(d(x, y))$. Of course, admissible transformations are those that, if d is a metric, ensure that $d' = f(d)$ is a distance metric too, i.e. ensure that d' too satisfies the four metric axioms.

Fig. 4.5 Plot of the concave function $f(x) = \ln(x)$. If there are two intersections of $f(x)$ with the same straight line, then between the intersections, the graph of $f(x)$ is above the straight line



All admissible transformations f satisfy a few simple properties:

- A1 : $f(0) = 0$,
- A2 : $a < b$ if and only if $f(a) < f(b)$,
- A3 : $f(a+b) \leq f(a) + f(b)$,

for all real numbers a and b . Clearly, A1 ensures that the interpretation of 0-distance (object-equivalence) is retained, A2 ensures that the order of distances is retained (monotonicity) and properties A2 and A3 (sub-additivity) together ensure that the triangle inequality is retained after transformation of the distances. If we require that f is continuous, such an f must be *concave*. A graphical account of what “concave” means is given in Fig. 4.5 where we use $f(x) = \ln(x)$ as an example. To see if a function is concave, draw its graph and a straight line intersecting it at two locations. If the graph is not below any such straight line between the points of intersection, we say that the function is concave. The reverse case, when the graph is nowhere above the straight line between the points of intersection, we say that the function is *convex*. A limiting case is the linear function: it is both concave and convex. Concave functions all have the property that the growth of the function values decreases everywhere: in Fig. 4.5, the increase of $f(x)$ always diminishes with increasing x .

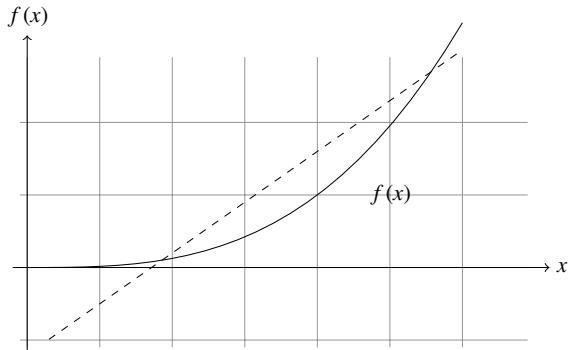
So, examples of admissible transformations are (see also Batagelj and Bren 1995)

$$\begin{aligned} f(x) &= ax && \text{with } a > 0, \\ f(x) &= \log_c(x+1) && \\ f(x) &= x^p && \text{with } 0 < p \leq 1, \\ f(x) &= a^{-x} - 1 && \text{with } a > 0 \end{aligned}$$

Why do we only allow for concave transformations? Why do we not accept all order-preserving transformations? The reason is that we want the transformed distances to satisfy the triangular inequality too:

$$f(d(u, w)) \leq f(d(u, v)) + f(d(v, w)).$$

Fig. 4.6 Plot of a convex function $f(x)$. If there are two intersections of $f(x)$ with the same straight line, then between the intersections, the graph of $f(x)$ is below the straight line



This can only be attained when f is non-decreasing in such a way that the left side of the inequality does not grow faster than the sum in the right side. For example, suppose that

$$d(u, w) = 10, \quad d(u, v) = 9 \text{ and } d(v, w) = 2$$

and that we would set $f(x) = x^2$. This would yield

$$f(d(u, w)) = 100 > f(d(u, v)) + f(d(v, w))$$

and thus we would violate the triangular inequality. The covering inequality S4 is the opposite of the triangular inequality, hence admissible transformations of similarities should have the opposite quality of concavity: convexity. A graphical definition of that property is provided in Fig. 4.6. Precisely: if f is a convex function such that $f(0) \geq 0$ and $f(x) < f(y)$ whenever $x < y$, then f is an admissible similarity transformation, i.e. if $s(x, y)$ is a similarity, then $s'(x, y) = f(s(x, y))$ is a similarity³ too. Examples of admissible similarity transformations are

$$f(x) = \alpha x^p + \beta \text{ with } \alpha > 0, \beta \geq 0 \text{ and } p \geq 1, \quad (4.14)$$

$$f(x) = e^x. \quad (4.15)$$

Normalization

A weight-difference of 10 kg between male adults of roughly the same age and length may not be very significant but that very same weight difference is a matter of life and death when it pertains to two 2-year old children. Similarly, a difference

³ Chen et al. (2009) use the expression “similarity metric” for any s that satisfies the axioms S1-S5. This is well-defendable since a similarity s “metricizes” the sequence space, just like a distance. However, we prefer to call such an s a “similarity” since the noun “metric” has been associated with “distance” for over a century now.

of 2 years of unemployment between careers of over 30 years on the labor market may be insignificant, but the same difference between the labor market careers of two 18-year old youngsters could have a dramatic differential effect on their future careers.

So, in many applications of measurement, differences or distances between pairs of objects are weighted according to the properties of the separate objects. As soon as we start considering differences, distances or proximities relative to the properties or features of the objects involved, we will almost automatically do two different things: we will tend to use relative, *dimensionless* or unit-free measures and we will create a *bounded* scale. Bounded by the maximum and/or the minimum of the difference, distance or proximity that could have been obtained, given the properties of the pertaining objects. For example, the weight difference between two male adults probably cannot exceed an upper bound of 250 kg. So, expressing the actual, observed difference relative to this maximum, will convey useful information about an observed weight difference.

The above intuitions about dimensionlessness and boundedness are captured by the notion of normalization. We will say that a measure M is “normalized” precisely when it satisfies the next two properties:

1. M is tightly bounded, i.e. $a \leq M \leq b$ for some numbers $a < b$,
2. M is dimensionless or, equivalently, unit-free.

Some authors require only boundedness and do not demand the dimensionlessness. When a measure is bounded, we know that its maximum and minimum values are fixed, independent of the (pairs of) objects it is applied to. We say that the bounds are “tight” when the maximum and minimum values, the boundaries, will be actually attained if M is applied to some suitable, real (pair of) object(s). A good example of a tightly bounded measure is Pearson’s correlation coefficient r : we know that $-1 \leq r \leq 1$ and that either of these boundaries will indeed be attained when the one variable is a linear transform of the other variable. The value 1.6 is a boundary of r too since $r < 1.6$ but 1.6 is not a tight boundary: $1 < 1.6$ is the smallest, the tightest upper boundary.

When a measure is “dimensionless”, we know that the numerical value of the measure does not refer to a unit or, equivalently, is not affected by admissible transformations of the scale it derives from. Again, Pearson’s r is a good example since linear transformations of either variable will not affect the numerical value of the measure of association $M = r$.

These two properties, boundedness and dimensionlessness, are valuable properties: boundedness implies that we can interpret the actual value of the measure with respect to its boundaries ($r = 0.87$ is a “high” correlation since close to the upper bound of 1) and we can compare different instances, different values of the measure, independently of sample space or the scales of the pertaining variables ($r = 0.87$ denotes a stronger degree of linear association than $r = 0.43$, regardless of the scales involved).

Therefore, normalization of a measure will greatly enhance its applicability and practical usefulness and thus, it makes sense to report normalized versions of measures, in particular of measures of distance and similarity. However, we should

be aware of the fact that a normalizing transform may not be order-preserving: a weight-difference of 10 kg for 2-year old kids is very serious, much more serious than a 12 kg weight-difference between two adults and this will be expressed in the normalized versions of the weight differences. So, a normalized distance or similarity does *not* preserve the order of the original distances since it weights relative to the properties of the pertaining objects. When these properties differ, the same weight will be normalized to different values, depending on the pertaining pairs of objects.

As distance and similarity are nonnegative, it is practical to normalize such that our measures will be tightly bounded by the closed interval [0,1]. A tight upper bound of 1 will invite to interpret the actual values as fractions of the maximum attainable upper bound. Normalization of distance or similarity should not lead to a loss of one of the metric properties. So, we demand that a normalized distance D and a normalized similarity S satisfy the axioms as stated below:

Distance	Similarity
D'1 $D(x, x) = 0$	S'1 $S(x, x) = 1$
D'2 $0 \leq D(x, y) < 1$	S'2 $0 < S(x, y) \leq 1$
D'3 $D(x, y) = D(y, x)$	S'3 $S(x, y) = S(y, x)$
D'4 $D(x, z) \leq D(x, y) + D(y, z)$	S'4 $S(x, z) + 1 \geq S(x, y) + S(y, z)$

Clearly, the two axiom systems are complementary so we expect that normalized distances can be obtained from normalized similarities and vice versa. Indeed, we will make some remarks on such conversions later. First, we will turn to the most important problem, the actual normalization of distance and similarity, such that the results adhere to the axioms D'1-D'4 or S'1-S'4 as stated above. In passing, we will specifically deal with OM-based and subsequence-based metrics.

Normalized Distance

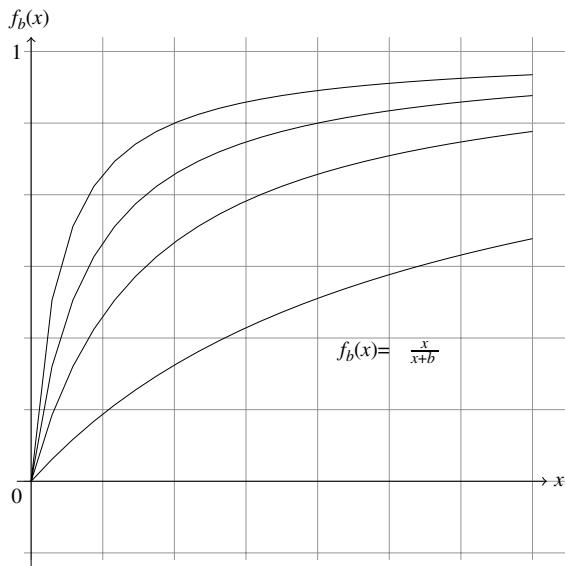
Here, we discuss two most simple versions of distance-normalizations as proposed by Chen et al. (2009) and apply them to OM- and subsequence-based metrics.

The first method relies on a simple “bounding” transform:

$$f_b(x) = \frac{x}{x+b} \text{ for } x \geq 0 \text{ and } b > 0 \quad (4.16)$$

Clearly, when $x=0$, $f(x) = 0$ and for increasing values of x , $f(x)$ will tend to 1 since whatever the value of b , it will become almost irrelevant for a big enough x . In Fig. 4.7, we show curves for four different values of b . A transformation like f_b has two shortcomings: when applied to a distance, the resulting transformed distance does not satisfy the triangle inequality and it does not yield a dimensionless quantity since it does not relate to the properties of the individual objects. To remedy these

Fig. 4.7 Plots of normalizing transform for $b = \{0.4, 0.8, 1.6, 5.4\}$



shortcomings, we replace the bounding constant b by a quantity that pertains to the individual objects involved and that ensures that the resulting quantity does satisfy the triangle inequality:

$$0 \leq D_r(u, v) = \frac{d(u, v)}{\underbrace{\{d(u, v) + d(u, r) + d(r, v)\}}_b / 2} \leq 1 \quad (4.17)$$

is a normalized distance for any reference object r . Often, it is most convenient to set $r = \lambda$, i.e. to take the empty sequence as the reference object. That 1 is the smallest upper boundary of D_r derives from the fact that d is a distance and thus satisfies the triangle inequality: since $d(u, v) \leq d(u, r) + d(r, v)$, we must have that $(d(u, r) + d(r, v) + d(u, v))/2 \geq d(u, v)$ and thus that $D_r(u, v) \leq 1$. Whatever the reference object r , $D_r(u, v)$ will depend on it. But $D_r(u, v)$ will not only depend on r and the comparison of u and v , but also on the comparison of u with r and the comparison of v with r . This implies that D_r is *not* order-preserving. For example, if $d(u, v) = d(u', v')$ but u' and v' are much more remote from r than u and v , $D_r(u, v) > D_r(u', v')$ (The reader is invited to graphically display this example). Hence, “raw” distances $d(u, v)$ are weighed by the lengths of the sequences involved. At first sight this may seem to be an unfortunate state of affairs. However, let us apply Eq. 4.17 to standard-cost OM-distance d_{OM} and set $r = \lambda$, i.e. take the empty sequence as the reference object. Then the above normalisation comes down to

$$D_\lambda(u, v) = \frac{d_{OM}(u, v)}{(d_{OM}(u, v) + |u| + |v|)/2} \quad (4.18)$$

wherein $|u|$ denotes the length of the sequence u . Now suppose that $d_{OM}(u, v) = 5$ and $|u| = 7 = |v|$. When we have to use at least 5 edits to turn a short u into a short v , $d(u, v)$ is a big distance. But if $|u'| = 50 = |v'|$, $d(u', v') = 5$ is only a small distance and we will see that $D_\lambda(u, v) > D_\lambda(u', v')$. So indeed, “raw” distances are weighed by the lengths of the sequences involved and this attractive property results in a normalized distance that is *not* order-preserving. The same effect will occur when we apply Eq. 4.17 to subsequence-based distances d_v :

$$D_\lambda(u, v) = \frac{d_v(u, v)}{(d_v(u, v) + \sqrt{\mathbf{u}'\mathbf{u} + \mathbf{v}'\mathbf{v}})/2}. \quad (4.19)$$

In both cases, the original distances of the sequences involved are weighted according to their lengths—in the OM-metric, the length of u relative to λ equals $|u|$ and in the subsequence-metric it equals $\sqrt{\mathbf{u}'\mathbf{u}}$. In actual application, it is relevant to stress that Eq. 4.18 is correct only when the standard cost matrix is applied. If not, we have to calculate $d_{OM}(u, \lambda)$ as the sum $\delta(u)$ of the deletion costs of all states of u and thus our general normalizer for OM-distances becomes

$$D_\lambda(u, v) = \frac{d_{OM}(u, v)}{(d_{OM}(u, v) + \delta(u) + \delta(v))/2}, \quad (4.20)$$

a normalization already proposed by Yujian and Bo (2007). A second way to normalize distances is through

$$D_r(u, v) = \frac{d(u, v) - \min\{d(u, r), d(r, v)\} - \max\{d(u, r), d(r, v)\}}{2 \cdot \max\{d(u, r), d(r, v)\}} \quad (4.21)$$

Again, this D_r is not order-preserving, due to the same effects as explained above and it can be adapted in an analogues way to OM- or vector-based distances. The details of applying this normalization to OM or vector-based distance are left to the reader. A simple normalization that does not depend upon a reference object r but satisfies the axioms D'1-D'4, is attained by using the exponential transform as in

$$D(u, v) = 1 - e^{-d(u, v)}. \quad (4.22)$$

Consequently, the latter normalization will *not* weigh distance according to lengths of the sequences involved, it is order-preserving and $D(u, v) = D(u', v')$ whenever $d(u, v) = d(u', v')$. It is not a proper normalization since the result is not unit-free. I find it difficult to imagine a sensible application of it. For OM, normalizations have been proposed, e.g. in Gabadinho et al. (2011), that are quite similar to the normalizers that we have seen so far:

$$D(u, v) = \frac{d_{OM}(u, v)}{|u| + |v|} \text{ and } D(u, v) = \frac{d_{OM}(u, v)}{\max\{|u|, |v|\}} \quad (4.23)$$

but these normalising transformations do not yield proper distances in the sense that they not adhere to the triangular inequality. Therefore, the use of such normalizers should be avoided.

Similarities and Their Normalization

So far, we discussed general properties of similarities but we did not discuss how to construct them in actual practice. This is especially relevant for OM-based analysis, since the OM-algorithm, provided with a metric cost matrix, generates distances. So, these distances must serve as the basis for the construction of edit-based similarity. Therefore it is relevant to mention two different ways to construct a similarity, given some distance metric d . Both methods require a reference object r ; once this has been set, we have that

$$s_1(u, v) = d(u, r) + d(v, r) - d(u, v) \quad (4.24)$$

and

$$s_2(u, v) = \min\{d(u, r), d(v, r)\} - d(u, v). \quad (4.25)$$

and s_1 and s_2 both satisfy the similarity axioms S1–S5. Both constructions lead to objects that are remote from r being more similar than objects that are close to r . We mention some details pertaining to s_1 . First, we deal with the general OM-variant of it, setting $r = \lambda$. This yields

$$s_1(u, v) = \delta(u) + \delta(v) - d_{OM}(u, v) \quad (4.26)$$

wherein $\delta(u)$ again denotes the sum of all deletion costs of the characters of u . Interestingly, Yujian and Bo (2007) proposed a special case of the above similarity and then proved that it satisfies the covering inequality S4 but they did not recognize that inequality as a general property of similarity measures. For vector-based distances and $r = \lambda$, Eq. 4.24 yields

$$s_1(u, v) = \sqrt{\mathbf{u}'\mathbf{u}} + \sqrt{\mathbf{v}'\mathbf{v}} - d(u, v). \quad (4.27)$$

Similar details for s_2 are left to the reader. There is a well-known similarity coefficient for sets, first proposed by Rogers and Tanimoto (1960): for sets X and Y , the quantity

$$0 \leq T(X, Y) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \leq 1. \quad (4.28)$$

is a normalized similarity measure. The coefficient is widely known as the Tanimoto-coefficient (see e.g. Duda et al. 2001) and several authors have come up with

generalizations and variants (see e.g. Tversky 1977). Lipkus (1999) proved that the Tanimoto-coefficient satisfies the covering inequality S4 and used this to prove that $D(X, Y) = 1 - T(X, Y)$ is a distance metric over sets. A more general formulation of the Tanimoto-coefficient yields a normalizer for any similarity and thus for the coefficients specified in Eqs. 4.24 and 4.25:

$$S_1(u, v) = \frac{s(u, v)}{s(u, u) + s(v, v) - s(u, v)} \quad (4.29)$$

An alternative normalizer is

$$S_2(u, v) = \frac{s(u, v)}{\max\{s(u, u), s(v, v)\}} \quad (4.30)$$

More and more general normalizers are discussed in Chen et al. (2009). Detailing S_1 for the case of a vector-representation yields the normalizer

$$S_1(u, v) = \frac{\mathbf{u}'\mathbf{v}}{\mathbf{u}'\mathbf{u} + \mathbf{v}'\mathbf{v} - \mathbf{u}'\mathbf{v}}. \quad (4.31)$$

Finally, it is always possible to generate a normalized similarity S from a normalized distance D , since whichever of S or D is given, the other can be easily obtained through the equation $S(u, v) = 1 - D(u, v)$.

Similarity or Distance: does it Matter?

A final issue to be dealt with relates to the generally held belief that distance and similarity are interchangeable. For example, Martin et al. (2008) write:

In the context of the efforts to evaluate similarity and change in life course transitions, optimal matching analysis (OMA) has been recommended to complement classical statistical methods to make use of the holistic information encoded in biographical status sequences.

Similar neglecting of the difference between distance and similarity is abundant (e.g. in Gauthier et al. 2010) and not only in the social sciences. The belief seems to be that distance and similarity are opposite or inverse in the sense that similar sequences should be close, that dissimilar sequences should be remote, that remote sequences should be dissimilar, etc. However, distances and similarities are just vehicles that are used to create partitions of big sets of sequences and it is not at all clear that partitioning on the basis of distance will lead to the same partitioning as partitioning on the basis of similarities.

Interestingly, Emms and Franco-Peña (2013) investigated precisely this problem, confined to edit-based distances and similarities. Remarkably, one of their conclusions is that partitions based on hierarchical clustering of distances can always

Table 4.1 d denotes an arbitrary distance metric and s an arbitrary similarity measure and S and D denote their normalized versions. The table shows how to transform a row quantity to a column quantity, either by providing a simple formula, or by referring to equation numbers from this chapter

	d	s	D	S
d	concave f	4.24, 4.25	4.17, 4.21, 4.22	e^{-d}
s		convex f		4.29, 4.30
D	concave f			$1-D$
S		convex f	1-S	

be replicated by similarity-based hierarchical clustering but not vice versa: some similarity-based clusterings will not be replicable through distance-based hierarchical clustering. This suggests that similarity is a more general, more encompassing concept than distance. Indeed, this is reflected in the axiom systems: the negation of the axioms of a distance yields a similarity but the negation of the axioms of a similarity does not necessarily yield a distance.

Summary

In this chapter, I focussed on the concepts of distance and similarity and their intricate relations. I tried to explain the importance of the axiomatic foundation of the concepts and the importance of regularity-axioms like the triangle inequality and the covering inequality. I also tried to explain the principles of admissible transformations and, more importantly, the principle and consequences of normalization. In passing, I provided for some ready-to-use implementations for OM- and vector-based distances. To help the reader find her way through the forest of intricate relations and formulae, I constructed Table 4.1: it shows how to transform a “row”-measure into a “column”-measure. Some of the cells of Table 4.1 are empty; not because such transformations are impossible but because I could not think of a sensible application. The reader should be aware that *none* of the transformations I described is unique in the sense that they would be the only solutions to the pertaining transformation-problem; we only know that these solutions respect the axioms of distance or similarity and there might be a wealth of other solutions.

In the last section, I made a few remarks on the fact that similarity is a more encompassing, more general concept than distance. So we have to be precise in the questions we ask and the concepts we use. Once a structuring concept—distance or similarity—is chosen, the next question is to either embed in an OM-space or a vector-space. Finally, the issue to be dealt with is that of a cost-matrix or, in case of a vector-representation, the features to incorporate and how to weigh and compare them.

References

- Batagelj, V., & Bren, M. (1995). Comparing resemblance measures. *Journal of Classification*, 12, 73–90.
- Berghammer, C. (2010). Family life trajectories and religiosity in Austria. *European Sociological Review*, 26, 1–18.
- Bonetti, M., Piccarreta, R., & Salford, G. (2013). Parametric and nonparametric analysis of life courses: An application to family formation patterns. *Demography*, 50, 881–902.
- Bras, H., Liefbroer, A. C., & Elzinga, C. H. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography*, 47(4), 1013–1034.
- Chen, S., Ma, B., & Zhang, K. (2009). On the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24–25), 2365–2376.
- Crochemore, M., Hancart, C., & Lecroq, T. (2007). *Algorithms on strings*. Cambridge: Cambridge University Press.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.
- Elzinga, C. H. (2003). Sequence similarity—A non-aligning technique. *Sociological Methods and Research*, 31(4), 3–29.
- Elzinga, C. H. (2005). Combinatorial representation of token sequences. *Journal of Classification*, 22(1), 87–118.
- Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization and differentiation of family life trajectories. *European Journal of Population*, 23(3–4), 225–250.
- Elzinga, C. H., & Wang, H. (2013). Versatile string kernels. *Theoretical Computer Science*, 495, 50–65.
- Elzinga, C. H., Rahmann, S., & Wang, H. (2008). Algorithms for subsequence combinatorics. *Theoretical Computer Science*, 409(3), 394–404.
- Elzinga, C. H., Wang, H., Lin, Z., & Kumar, Y. (2011). Concordance and concensus. *Information Sciences*, 181, 2529–2549.
- Emms, M., & Franco-Peña, H.-H. (2013). On the expressivity of alignment-based distance and similarity measures on sequences and trees in inducing orderings. In P. Latorre Carmona, J.J. Sánchez, & A.L. Fred (Eds.), *Mathematical methodologies in pattern recognition and machine learning. ICPRAM 2012 international conference on pattern recognition applications and methods*, vol. 30 of *Springer proceedings in mathematics & statistics*, (pp. 1–18). New York: Springer.
- Fasang, A. E. (2010). Retirement: Institutional pathways and individual trajectories in Britain and Germany. *Sociological Research Online*, 15(2), 1.
- Fréchet, M. R. (1906). Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo*, 22(1), 1–72.
- Gabadinho, A., Ritschard, G., Müller, N., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Gauthier, J.-A., Widmer, E., Bucher, P., & Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology*, 40(1), 1–38.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857–871.
- Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3, 5–48.
- Halpin, B. (2010). Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods & Research*, 38(3), 365–388.
- Hamming, R. W. (1950). Error-detecting and error-correcting codes. *Bell System Technical Journal*, 26(2), 147–160.
- Han, S.-K., & Moen, P. (1999). Clocking out: Temporal patterning of retirement. *American Journal of Sociology*, 105(1), 191–236.

- Holliday, J. D., Hu, C.-Y., & Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D-fragment bit-strings. *Combinatorial Chemistry and High-Throughput Screening*, 5, 155–166.
- Hollister, M. N. (2009). Is optimal matching sub-optimal? *Sociological Methods & Research*, 38, 235–264.
- Lipkus, A. H. (1999). A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry*, 26, 263–265.
- Manzoni, A., Vermunt, J. K., Luijkx, R., & Muffels, R. (2010). Memory bias in retrospective collected employment careers: A model-based approach to correct for measurement error. *Sociological Methodology*, 40(1), 39–73.
- Martin, P., Schoon, I., & Ross, A. (2008). Beyond transitions: Applying optimal matching analysis to life course research. *International Journal of Social Research Methodology*, 11(3), 179–199.
- Massoni, S., Olteanu, M., & Roussel, P. (2009). Career-path analysis using optimal matching and self-organizing maps. In J.C. Príncipe & R. Miikkulainen (Eds.), *Advances in self-organizing maps. Lecture notes in computer science 5629*. (pp. 154–612). New York: Springer.
- Rogers, D. H., & Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, 132, 1115–1118.
- Rohwer, G., & Pötter, U. (1999). *TDA User's Manual*. Bochum: Ruhr-Universität.
- Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels. Support vector machines, regularization optimization, and beyond*. Cambridge: MIT Press.
- Shawe-Taylor, J., & Christianini, N. (2004). *Kernel methods for pattern recognition*. Cambridge: Cambridge University Press.
- Studer, M., Ritschard, G., Gabadinho, A., & Muller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, 40(3), 471–510.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352.
- Wang, H. (2006). Nearest neighbors by neighborhood counting. *IEEE Transactions on Pattern Learning and Machine Intelligence*, 28(6), 1–12.
- Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6), 1091–1095.

Chapter 5

Three Narratives of Sequence Analysis

Brendan Halpin

Three Narratives of Sequence Analysis

At the heart of sequence analysis lies the calculation of similarity between pairs of sequences. The Optimal Matching algorithm (OM) is clearly the most popular measure in the recent literature. However, despite its success, it has limitations, particularly when applied to lifecourse data. This paper addresses some of these problems, and considers a number of alternative ways of defining similarity between lifecourse trajectories. Similarity measures can be compared in terms of three related dimensions: one, their algorithmic structure, two, their practical results, and three, the extent to which they can yield sociologically interpretable distances between sequences, their “narrative” of similarity. OM’s narrative defines similarity as edit-distance, and this is attractive for sequences that are naturally discrete in time, but does not map without problem onto sociologically meaningful intuitions of similarity, particularly when time is continuous.

The alternatives considered below include two relatively minor modifications of OM, a group of methods that focus on enumerating common subsequences, and a time-warping method. Each of the alternative algorithms implicitly defines similarity in a different way, providing a different narrative and different intuitions of similarity. The two modifications of OM attempt to address problems arising from OM’s focus on context-free comparisons between tokens, but for technical reasons fail to produce valid distance data. The subsequence-oriented approach defines similarity in terms of the extent to which two sequences go through the same states in the same order, which is radically different from the principles underlying OM, and empirically provides very different distances. Finally, the time-warping method provides distances defined in terms of local expansion and compression of the time axis, which yields a very different way of looking at similarity of longitudinal data. The implementation of the time-warping algorithm is structurally very similar to OM,

B. Halpin (✉)
Department of Sociology, University of Limerick, Limerick, Ireland
e-mail: brendan.halpin@ul.ie

though it produces distances that are in many cases intermediate between OM and the subsequence approach.

Thus, in terms of tools for comparing sequences, we have three distinct narratives: similarity in terms of edit distance (potentially with modifications), similarity in terms of going through the same states in the same order, and similarity in terms of warping the time axis. At one level measures are algorithms, with deterministic behavior set by their rules and parameterisation, but we can think of “narrative” as referring to our understanding, perhaps heuristic, of the relationship between the algorithms and interpretable, meaningful comparisons made between sequences.

Method and Meaning

For sequence analysis to be useful to sociology, it has to make sociological sense. We can reduce this requirement in the current context to a demand that inter-sequence distances can be made to correspond systematically to sociological ideas of similarity, leaving aside the important but less distinctive issue of what use (clustering, comparison with ideal types, etc.) is made of the similarity data. This topic has not received a great deal of systematic treatment, though proponents of sequence-analytic measures will implicitly or explicitly assert that their measure is meaningful.

Sociological meaningfulness was one of the multiple foci of criticism of sequence analysis posed by Wu (2000) and Levine (2000) (in response to Abbott and Tsay 2000), who still represent the most sustained intellectual attack on the approach even now, some 14 years later. Other criticisms tend to focus on the advantages of one distance measure compared to others, while taking the utility of sequence analysis as such for granted (e.g., Lesnard 2010; Hollister 2009; Halpin 2010).

Wu and Levine raised many concerns about the black-box nature of the process, and the inability of its proponents to satisfactorily elucidate the link between the operation of the OM algorithm and theoretically relevant differences between trajectories. They saw its utility in molecular biology as arising from a close relationship between its elementary operations and processes of change in DNA recombination, with this mapping being entirely invalid for social processes. However, OM’s algorithm does not intentionally (or even particularly closely) model molecular biology processes; rather it is driven by computational tractability, and can be applied to token strings in a variety of contexts, including DNA and social applications. This view of the algorithm as modelling the data generation process lead them to compare it unfavourably to the alternative narrative commonly applied to longitudinal social data, hazard rate modelling or “event history analysis” (EHA). EHA presents a very attractive mental (as well as statistical) model of lifecourse processes, with causality operating in continuous time, taking account of current state, history and duration dependence, and can indeed model the data generating process. Moreover, as a statistical model it can estimate the effects of other measured variables and test precise hypotheses about them. However, sequence analysis (which typically has other goals) is not a model of data generation, but rather (at the level of the distance measure) a way of defining pairwise similarity for (typically) descriptive and exploratory purposes. Wu in particular read the algorithm as competing with

models like EHA, which lead him to unnecessarily strong objections to the logic of the elementary operations. In particular he misread the substitution operation as a kind of transition, and objected to “impossible” transitions (such as from married to never-married). While he may have been mistaken, parameterisation of OM worries lots of people—rightly, because parameterisation has a stark effect on the resulting distances. However, the problems are not insurmountable, particularly if substitution costs are thought of as simply describing differences between the categories of the state variable. Sequence analysis therefore consists of a mapping of the state-space distances onto the sequence domain, yielding a set of sequence-space distances. Viewed this way substitution costs are less intimidating, and in many simple ways the state-space patterns will influence the sequence-space patterns, though given the extra complexity inherent in sequences, this is not a deterministic relationship.

As we have seen, the competing narrative of event history analysis is very attractive, though it does a very different task to SA. Other statistical models that provide competing narratives include latent class analysis (LCA) (e.g., Barban and Billari 2012), which focus on the search for classes rather than modelling the data generation process, or latent growth curve models (LGCM) (e.g., Lovaglio and Mezzanica 2013) which focus on the multiple-individual time-series nature of the data.

While one may discount some of Wu and Levine’s objections, and note that sequence analysis can serve different functions from EHA, LCA, LGCM and other statistical models, it is well to note that their fundamental injunction holds true: for sequence analysis to be worth the candle we need to be able to justify the narrative it gives us about similarities between sequences through sociological state spaces.

The Narrative of Sequence Analysis: Mapping State-Space onto Sequence Space

Sequence analysis is concerned with identifying similarity between sequences. We do this by reference to information about the state space through which the sequences move, at the minimum in terms of binary match or mismatch between states (as in subsequence approaches), but usually with more detailed ideas of distances within the state space (as when we use the “substitution cost matrix” in OM). In a broad sense the narrative of sequence analysis is the mapping of information about the state space onto the sequence space. How the mapping occurs is obviously critical, but varies, sometimes dramatically, from one algorithm to another.

The simplest mapping is the Hamming distance, where the inter-sequence distance is the sum of the state distances at each timepoint (Hamming 1950). While the correspondence between the state space and sequence space is thus utterly clear, Hamming is blind to similarity that is displaced in time. The other available algorithms allow time-dislocation in different ways, at the expense of a more complicated relationship between the state space and the sequence space. OM uses deletion and insertion of elements in the sequence to allow time-dislocation by “alignment”, but otherwise replicates the Hamming distance’s mapping between state and sequence (though in the context of OM this is described as “substitution”).

Time-warping methods achieve the dislocation by warping the time axis, and then doing a Hamming-style comparison. Subsequence-based methods depart entirely from the Hamming-style comparison, and achieve a radical time-dislocation by focusing on the order of states rather than when they occur. Thus the available algorithms will have different consequences and give different ways to think about the similarity.

OM's edit-distance narrative leads to differences after alignment being understood as substitutions (an edit operation). However, viewing the substitution cost matrix as a state-space distance matrix has advantages, stressing the commonality with other contexts where substitution is a less attractive metaphor, such as Hamming distance and time-warping. Viewing substitution costs as state-space distances also demystifies the derivation of substitution costs: state-space distances are just statements about differences between the categories of the state-space variable. These differences can be derived from theory or intuition, based on external data or even derived from the sequences' transition rates. It might also avoid common misunderstandings about substitution, sometimes seen as directly modelling the empirical sequence-generating process, or leading to worries about the misapplication of molecular-biology models to sociology, and about "impossible transitions" etc.

The paper proceeds by looking at this set of measures, benchmarked against the simplest of them all, the Hamming distance, and compares how they perform with real lifecourse-history data. All methods but Hamming allow some sort of time dislocation, be it by insertion and deletion, compression and expansion, or counting of subsequences wherever they appear. The measures' performance in a typical sequence-analysis workflow is compared: cluster analysis to generate a descriptive data-driven typology. All methods except the subsequence methods give results that are not very different from Hamming, suggesting that time-dislocation is important only a small part of the time; the subsequence method produces quite different results. The next section considers the shape of the sequence space in so far as multi-dimensional scaling can reveal it; this again reveals that the subsequence method is different, but also that for OM and the time-warping method there is little evidence of natural clustering among the trajectories. The following section bypasses the relative instability of cluster analysis by analysing correlations between measures: this shows again that the subsequence method is very different, and that the structure of distances in the state space is more important, much of the time, than the nature of the measure. In the final section, the time-warping measure is explored in more depth, showing that as its parameters are varied, it constitutes a bridge between OM-style measures and order-based measures like the subsequence algorithm.

Lifecourse Data and Token Sequences

Optimal matching, or the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), and related approaches to defining distance or similarity between sequences, work with sequences that pass through a finite, discrete state space, such as strings

of bytes representing characters (in the case of Levenshtein 1966, bits or other tokens representing “words”). This maps well onto domains like fuzzy text search in computer science, or molecular biology where sequences are linear macro-molecules with repeating elements drawn from a small set (e.g., DNA, which can be represented as a sequence drawn from a set of four elements or “bases”, labelled C, A, G and T). It also works for other naturally discrete state spaces such as a coding of utterances in conversations, steps in a dance, voting behaviour at successive elections, and so on. OM’s elementary operations of insertion, deletion and substitution implicitly require sequences to consist of discrete, successive tokens, which can be operated on as independent atomic units. In contrast, life courses operate in continuous time (though in practice always measured with rounding, for example, in whole months). Two approaches are commonly used to represent continuous time/discrete space trajectories as token sequences: either whole spells in a given state are represented as single tokens, ignoring duration, or time units are represented as tokens so that spells are represented by runs of consecutive tokens, representing duration as repetition. The latter approach is used widely and, on the whole, it works well. However, if the distance measure implicitly treats tokens as strictly distinct from their neighbours, this is sub-optimal. Where the average spell length is longer than one token, successive tokens are highly likely to have the same value, simply because no transition has occurred. In a true token string, while successive tokens will be correlated and often identical, each token is an independent realisation. For instance, if our sequence is vote at successive elections, party loyalty and patterns of flow between parties will mean that successive observations are related, but each vote is a distinct event. If however, the sequence is an annually-collected report of the most recent vote, it can only change when there has been an election: in this sequence for most $t / t+1$ transitions there is a very different sort of dependence. A more common example would be a trajectory through an occupational group state space: here most months will succeed the previous month without change of state, but only because no change of job has occurred, not that a new job in the same group has been found. This sort of dependence is not well picked up by OM and similar techniques, because we are dealing with spell data and they do not take account of that sort of structure.

That is, we often have reason to feel that the token-oriented view is inadequate for lifecourse data. Alternatives exist, but they have not been considered systematically together. In this paper I consider together a number of approaches that

- attempt to look at the context of tokens (the values of their neighbours, the length of the spells in which they are embedded)
- focus on spells, weighted by duration, rather than repeating tokens, or
- focus on the time dimension, defining similarity in terms of warping time.

I consider both the technical characteristics of the approaches, and the sort of socio-logical story we can use them to tell about similarity between life course sequences.

Modifying the Algorithm: “Localised” and “Duration-Adjusted” OM

One key problem with OM is that by defining distance in terms of token editing, it is blind to sequential context. Only the values of the pair of tokens, one in each sequence, are taken into account, and their environment is invisible. Intuitively it is attractive that edits could mean different things according to context. In lifecourse research, where transitions are relatively rare, and spells are typically rather longer than one unit, a deletion that alters a spell length is less important than one which removes a short spell entirely. Under OM, AAAB is as distant from AACB as from AABB (given $d(A, B) = d(A, C)$). More generally, it is reasonable that the importance of an edit can be affected by context, such that substituting B for A in WAX may not mean the same thing as in YAZ, while for OM the substitution cost is a function of A and B and nothing else.

This general problem motivated two independent approaches to modifying OM. While different in detail, both the modified algorithms proposed by Hollister (2009) and Halpin (2010) take context into account.¹ Unfortunately, neither approach preserves the metric character of OM distances, which limits their utility considerably. I consider them here insofar as they throw light on the relationship between the algorithm and the meaning of the resulting distance, rather than as measures to be recommended for general use.

Hollister’s “Localised OM” Algorithm (LOM)

Hollister’s motivation is specifically the issue of sequential context, that, in not taking account of a token’s neighbours, OM gives edits that should be substantively different the same weight. If one inserts an element in a string, that insertion should be less costly, the more the inserted element is similar to the two adjacent tokens. This will have the consequence that insertions which lengthen a spell (i.e., the two neighbours are identical to the inserted element) will be very cheap, but it is more general in that insertions between tokens similar but not identical to the inserted token also cost less. Thus, to insert element k between elements i and j the indel cost is:

$$\iota_{ijk} = \alpha \frac{d_{ik} + d_{jk}}{2} + \beta$$

where α and β are chosen by the analyst and d_{ij} is the distance between i and j in the state space (i.e., the ij substitution cost). To insert at an end of a sequence, $\iota = \alpha d_{ik} + \beta$. The β is a fixed cost for insertion, and α weights the distance, d , between the inserted element and each of its neighbours. (Note that while the OM

¹ The LOM and OMv measures are available as part of the SADI package of sequence analysis tools for Stata, at <http://teaching.sociology.ul.ie/sadi/>.

algorithm is usually described in terms of insertions, deletions, and substitutions, we can think of deletions as insertions in the other sequence, and of substitutions as insertions (or deletions) in both sequences).

Duration-Adjusted OM (OMv)

OMv, the duration-adjusted version of OM described in Halpin (2010), has a motivation that is superficially different from Hollister's LOM, but with strong underlying similarities. The key intuition is that operations on longer spells should cost less than operations on shorter spells. An insertion of a missing element in s_1 is equivalent to deleting the extra element in s_2 ; a substitution is equivalent to a paired deletion of the corresponding elements in the two sequences. Thus all operations can be considered deletions, and deletions shorten the spell in which they occur, obliterating it if it has only one element. OMv operates by weighting such operations less, the longer the spell in which they operate. A weight of $1/\sqrt{l}$ is used as the default, where l is the spell length. Thus deleting an entire 1-unit spell would cost more than shortening, for example, a 5-unit spell by 1 unit. However, deleting a whole spell still costs more the longer it is. LOM will have a similar effect of favouring the cheapening operations within spells, though only by considering the adjacent elements, not the whole spell. OMv is more general than LOM in looking to the entire spell rather than the two adjacent elements, but LOM is more general than OMv in looking at the distance to the adjacent tokens, rather than just whether they are part of the same spell. While LOM has a narrative of attention to local context, OMv has a narrative of paying attention to lifecourses as sequences of spells.

Both measures are implemented as relatively simple modifications of the Needleman–Wunsch algorithm, but it can be shown that, unlike OM, the dissimilarities they produce do not have the metric property. In the next section I discuss what this means and why it matters, and then go on to demonstrate that the measures are not metric.

Why Metricity Matters

It is important that dissimilarities be metric, if we are to use them with conventional cluster-analytic or multi-dimensional scaling tools, which rely on the dissimilarities having a coherent global relationship, such that they can be used to construct a (perhaps latent) space within which the observations can be arrayed. Even more, it is very hard to think of a non-metric dissimilarity as a distance. The metric property distills critical characteristics of distance as we experience it in everyday Euclidean 3-dimensional space, and allows us to think of certain non-Euclidean measures as distances. A dissimilarity $\Delta(x, y)$ can be considered as metric if it has the following characteristics:

1. $\Delta(x, y) = 0 \Leftrightarrow x$ and y are the same
2. $\Delta(x, y) \geq 0$
3. $\Delta(x, y) = \Delta(y, x)$
4. $\Delta(x, y) \leq \Delta(x, z) + \Delta(z, y)$: the triangle inequality

In other words: distance is zero if and only if the entities are identical (or are in the same location); negative distances are impossible; distances are symmetric; and the distance from x to y cannot be greater than the distance from x to any other point plus the distance from that point to y —there are no short cuts! These conditions are naturally fulfilled by distances in Euclidean space, but they are also satisfied for a large body of measures in non-Euclidean space. At an intuitive level it is easy to see that dissimilarities that satisfy these conditions are like distances, and that those that do not may cause difficulties. For many dissimilarities the first three are easy to satisfy (though if the measures are travel times in a city with one-way streets, requirement 3 is not satisfied). The triangle inequality often causes difficulties.

It is inherent in the OM algorithm that the triangle inequality is satisfied, since it calculates the dissimilarity as the cheapest additive concatenation or summation of single-edit distances. First, note that as long as the state-space distances are metric, all one-operation distances in OM are metric. That is, the distance from ABA to ACA depends on the substitution cost or state-space distance, $d(B, C)$, and for there to exist a state D such that $\delta(ABA, ACA) > \delta(ABA, ADA) + \delta(ACA, ADA)$ would imply that the distance in the original state space from B to C is greater than the sum of $B \rightarrow D$ and $D \rightarrow C$; i.e., substitution gives metric distances between sequences unless the space described by the state-space distance matrix is non-metric. Since insertions and deletions have a single fixed cost, distances like $\delta(ABC, AC)$ will also be metric.

That OM's minimising concatenations will generate metric distances can be demonstrated by considering a route, $\delta'_1 2$, from s_1 to s_2 , made by a sequence of elementary OM operations without minimising the cost. If we find a sequence s_3 such that there exist routes such that $\delta_{13} + \delta_{23} < \delta'_{12}$ we have shown that δ'_{12} is not the cheapest route, since the $s_1 \rightarrow s_3 \rightarrow s_2$ route is also valid and is cheaper. Thus OM necessarily produces metric routes, and if we find the *cheapest* route from s_1 to s_2 , there can be no third sequence such that the triangle inequality is not satisfied.

In a general sense, we can consider non-metric dissimilarities to arise where the measures are not coherent enough to imply a space, a structured relationship independent of the observations, within which the observations can be located. Some measures are very good at identifying close matches, but where the dissimilarity is great, differences between the values of the measure are uninformative. Some measures may compare x and z using one set of shared characteristics, and z and y using another set, such that x and z are similar, and z and y are similar, but if x and y do not share characteristics they may be judged as very dissimilar. Elzinga (2006) has a valuable discussion on measures and metric properties.

Localised and Duration-Adjusted OM are not Metric

It is easy to demonstrate that LOM is not metric. If we consider the following three sequences:

- $s_1 = \text{BBBBAB}$
- $s_2 = \text{CCCAAC}$
- $s_3 = \text{BBBACC}$,

given the following cost structure:

- $\iota = 0.5 \frac{d_{ik} + d_{jk}}{2} + 0.5$
- $d_{ij} = 1, i \neq j$
- $d_{ii} = 0, i = j$, the following dissimilarities are generated:

Pair	LOM		OM	
	$d = 1, \alpha = \beta = 0.5$	$\iota = 1.0$	$\iota = 0.75$	$\iota = 0.75$
δ_{12}		6	6	5.5
δ_{13}		2.5	3	2.5
δ_{23}		3	3	3

The dissimilarity between s_1 and s_2 is greater than the sum of the dissimilarity between s_1 and s_3 , and s_3 and s_2 : $d(1,3) + d(3,2) = 2.5 + 3 < d(1,2) = 6$; the triangle inequality does not hold. The sequence s_3 becomes s_1 with a discounted insertion of a B between two Bs plus two other operations, which puts the dissimilarity below the level OM gives with the corresponding *indel* cost of 1.0. If we reduce OM's *indel* cost to reduce δ_{13}^{OM} to match, it also reduces δ_{12}^{OM} , preserving the triangle inequality. Thus under LOM, because there is a discounted edit in δ_{13}^{LOM} that is not triggered in the δ_{12}^{LOM} or δ_{23}^{LOM} comparisons, the measure is not metric.

Like LOM, OMv also produces non-metric dissimilarities. Effectively, sequences with long spells will be judged more similar to all other sequences, because they attract discounted operations. Two sequences with many short spells may be quite dissimilar from each other but each may be judged quite similar to the long-spell sequence, violating the triangle inequality. For instance, the OMv distance between BBBBAB and CCAAAC is 3. However, going through BBBBAA the distance is $0.41 + 2.45 = 2.86$. The fact that BBBBAA consists of a single spell means that its distance even from a spell with no shared elements such as CCAAAC is reduced.

To summarise, from a narrative point of view, OM gives a definition of distance in terms of edits in token strings, which has great virtues of simplicity and efficiency, and takes into account the spatial structure of the initial state space in a clean way. However, by working with a model of time as naturally discrete, and of elementary operations as blind to their sequential context, the distances do not

map unproblematically on to lifecourse history and other sociological data. The two attempts to adapt the algorithm to improve its narrative are sociologically attractive, but run into technical difficulties.

Combinatorial Measures: Similarity by Counting Subsequences

Token-editing approaches define distance as the cost of the edits necessary to turn one sequence into another, but another tradition in sequence analysis has a radically different approach—and, as we will see below, distinctly different results.

In the 1990s Dijkstra and Taris (1995) proposed a way of looking at similarity of longitudinal data that focused on order. Leaving aside for the moment concerns with duration, two sequences are more similar, the more they go through the same states in the same order. They proposed a way of calculating this similarity; by ignoring non-common elements and repeated elements their algorithm was tractable. However, as Abbott (1995) pointed out, this discards a lot of meaningful information, and for token sequences OM offers a more flexible and general measure. Nonetheless, there is something very attractive about this order-focused principle of similarity.

Elzinga proposed a more general approach which draws on this concept of similarity (Elzinga 2003, 2005, 2006). His method efficiently counts subsequences (order-preserving subsets of a sequence, not necessarily consecutive) and he proposed a variety of measures based on enumerating matching subsequences. If two sequences share a subsequence, they go through those states in the same order; Elzinga's method efficiently enumerates all such matching subsequences, and thus achieves in a more general way Dijkstra and Taris's goal.

All three of OM, the Dijkstra–Taris measure and Elzinga's measures are oriented to token-strings—treating the individual elements as atomic and capable of being considered out of context. Where duration is important, we can represent time by repeating tokens proportionally to spell length; as discussed above this has drawbacks, but further with Elzinga's method, long sequences are costly to process (since the number of subsequences is 2^l where l is the sequence length). However, he describes a number of variants of his measure which treat spells as tokens whose subsequence matches can be weighted by a function of the spell durations. This is more efficient than treating person–time-units as tokens: a five-year sequence will contain 60 person–months but perhaps only three or four spells. However, it must be noted while person–months are clearly not naturally discrete units, neither, strictly speaking, are spells. The observed duration has a substantial random component: the fact that a case stayed in a given state for a given duration does not make that duration a naturally discrete unit, when it is possible that a very minor change in circumstances would have created two shorter spells from one longer one.

Thus Elzinga's version of Dijkstra and Taris's principle of similarity provides another compelling narrative, which can be applied to lifecourses considered as

sequences of spells with durations. In so far as questions remain about its operation, they have to do with the fact that existing implementations count matches as binary, and cannot deal with partial similarity (as other methods do via state-space distance matrices), and whether the algorithms of enumeration (which involve different types of double counting, of subsequences of subsequences), and duration-weighting of spells as tokens, make for interpretable distances in lifecourse contexts. With reference to the issue of partial matches, it is worth noting that Elzinga and colleagues are preparing methods that deal with “soft matching” (Elzinga and Wang 2013; Elzinga and Studer 2013).

Implementation

Below, I present results using a duration-weighted, spell-oriented version of Elzinga’s “number of matching subsequences” (NMS) similarity measure, which I refer to as the X/t measure. My implementation in Stata differs from the algorithm described in Elzinga (2006), in two main ways. First, rather than use the efficient algorithm he describes for enumerating common subsequences of a pair of sequences (which is called $\frac{N(N-1)}{2}$ times), I enumerate by brute force the subsequences of each sequence (called only N times) and then efficiently count the matches ($\frac{N(N-1)}{2}$ times). Second, Elzinga proposes a number of ways of taking duration

into account, the most intuitively attractive of which is to weight by the sum of the spell-wise product of the durations of the subsequences. Thus, for a pair of subsequences, [A/5,B/4] and [A/1,B/2], Elzinga suggests the matching AB be weighted by $5 \times 1 + 4 \times 2 = 13$. However, as a consequence of the internal data structure, I cannot exactly replicate the duration weighting and instead weight by the product of the cumulated subsequence duration, $(5+4) \times (1+2) = 27$.

The distance measure is defined as:

$$\delta^{X/t} = \sqrt{SXX + SYY - 2 \times SXY}$$

where SXY is the sum of the product of the cumulated duration of each subsequence shared between sequences X and Y. SXX and SYY represent the same measure for X compared with X and Y with Y, respectively, that is, the sum of the square of the cumulated duration of each subsequence.

Consider three example sequences: $s_1 = (A/10, B/4, C/6)$, $s_2 = (A/10, B/7, C/3)$ and $s_3 = (B/9, A/5, B/6)$. All three have the same number of elements, and thus the same number of subsequences. However, since s_3 has a repeated element, it has a smaller number of *distinct* subsequences (the subsequence B appears twice, with a total duration of 15). The SXX measure (respectively 1104, 1116 and 1192) is the sum of the square of the cumulated duration in each distinct subsequence (so the BB subsequence in s_3 yields 15×15 rather than $9 \times 9 + 6 \times 6$). The resulting distances

are as follows: $D(s_1, s_2) = 6.0$, $D(s_1, s_3) = 42.0$ and $D(s_2, s_3) = 40.3$. The sequences s_1 and s_2 share the ABC structure so match quite closely, while s_3 is judged quite different, though it does match the other two with the AB structure. Because it spends more time in B, s_2 is marginally closer to s_3 than s_1 is.

Elzinga has implemented many of his proposed measures in his own software, CHESA. A number are also implemented in the R package for sequence analysis, TraMineR (Gabadinho et al. 2011), and this X/t implementation is available in SADI.

Time-Warping

The concept of time-warping provides a third type of narrative: trajectories may have different or varying speeds such that we can locally compress and expand time to maximise their similarity. The amount of distortion plus the amount of residual difference can be viewed as a measure of dissimilarity. While this gives a mechanism to “align” parts of sequences, to respond to full or partial matches at the same or different times, it does so by mechanisms that have a different surface logic from OM (time-warping rather than string editing), one which is much more appealing for cases where we think of time as essentially continuous (such as lifecourse data). While the surface logic is, and resulting dissimilarities can be, quite different from OM and related algorithms, it is interesting to note that it is implemented in a manner very similar to OM’s Needleman-Wunsch algorithm.

Time warping has been around as a term for quite a while: for instance, Abbott and Hrycak (1990) use the term to suggest using non-linear time scales (for instance the log of sequence time) to cope with domains where the rate of transition varies with time (e.g., labour market volatility in youth giving way to stability), and in a sense OM warps time by matching patterns at different time-points. However, in processing longitudinal data the term has always had a more specific meaning, which features in the seminal work on sequence analysis, *Time Warps, String Edits and Macromolecules* (Sankoff and Kruskal 1983).

Time-warping has been used widely in computer-science contexts. It was favoured for tasks like speech recognition, signature verification, and other machine learning tasks. Conceptually it is a continuous time approach, but in practice any time-series data processing must discretise, either by sampling the data at regular intervals (e.g., sound recording samples air pressure 41,000 times per second, unemployment figures are given on a weekly basis) or through periodic summaries (such as income per month, daily cumulated rainfall).

While not all time-warping dissimilarities are metric, a modified time-warping distance measure, the time-warp edit distance (TWED), has been proposed by Marteau (2007, 2008), who has shown it to be metric. I consider the measure here as a competitor to OM for lifecourse data. It is quite similar to OM in its operation, as it uses a state-space distance matrix (functionally equivalent to OM’s substitution matrix), and has operations analogous to substitution, insertion and deletion (though

the latter two are better thought of as compression and expansion, or even better as compress-A and compress-B). It has a stiffness parameter v and a gap-penalty, λ .

Formally, time warping is a family of algorithms that can be said to do “continuous time-series to time-series correction” while OM et al. do “string to string correction” (Marteau 2007). That is, conceptually time-warping uses continuous time, but it can be shown to work well in discrete time (Kruskal and Liberman 1983). Marteau shows that there is a low bound to the discrepancy caused by such discretisation for this measure. While TWED can accommodate any sort of state space, and is usually described in terms of \mathbb{R}^n , a space composed of possibly many real dimensions where distances between points can be calculated in Euclidean or other terms, there is no difficulty in mapping to a discrete state space where distances between states can be given in a lookup table, the state-space distance matrix. TWED is designed to accommodate irregular time-sampling, but is a little simpler to program when we have fixed time steps, as is the case considered here, and as is typically the case with lifecourse data.

In its internal operation, it differs strongly from OM in that the operations (i) consider consecutive pairs of tokens in all three operations, (ii) have a stiffness parameter that cumulates each time a comparison is made where time is realigned, and (iii) do not edit the content or order of the sequence (insert or delete) but align by altering the time dimension (though in reality there is some elision of elements).

TWED offers an alternative to OM that is very similar in terms of implementation, but quite different in its motivation. By virtue of its stretching and compressing operations, and its attention to successive pairs of tokens, it is likely to respect the spell structure of the trajectory better than OM. In this respect, and since it generates metric distances, it may well achieve what LOM and OMv attempted.

The TWED Algorithm

In practice, the implementation and application of TWED is very similar to OM. It takes as parameters a matrix of distances between the states and two parameters for stiffness and gap penalty that are broadly analogous to *indels* in that they facilitate or deter compression/expansion. Thus we are looking for full and partial similarity at the same or near location, where “partial” and “near” are affected by our parameterisation (as is true of OM).

Also similar to OM is the internal detail of the implementation. As is well known, OM can be described as a recursive algorithm such that the distance between sequence A (up to element p) and sequence B (up to element q) is given by:

$$\delta^{OM}(A^p, B^q) = \min \begin{cases} \delta^{OM}(A^{p-1}, B^q) + \iota \\ \delta^{OM}(A^{p-1}, B^{q-1}) + d(a_p, b_q) \\ \delta^{OM}(A^p, B^{q-1}) + \iota \end{cases}$$

where ι is the *indel* cost and $d(a_p, b_q)$ the substitution cost between element p of sequence A and element q of sequence B . This can be programmed efficiently in $p \times q$ operations. TWED can be expressed in an identical structure:

$$\delta^{TW}(A^p, B^q) = \min \begin{cases} \delta^{TW}(A^{p-1}, B^q) + d(a_p, a_{p-1}) & + \nu d(t_{a_p}, t_{a_{p-1}}) + \lambda \\ \delta^{TW}(A^{p-1}, B^{q-1}) + d(a_p, b_q) + d(a_{p-1}, b_{q-1}) + 2\nu d(t_{a_p}, t_{b_q}) \\ \delta^{TW}(A^p, B^{q-1}) + d(b_q, b_{q-1}) & + \nu d(t_{b_q}, t_{b_{q-1}}) + \lambda \end{cases}$$

On the simplifying assumption that observations are taken at fixed 1-unit intervals (i.e., $d(t_{i+1}, t_i) = 1$), this simplifies to:

$$\min \begin{cases} \delta^{TW}(A^{p-1}, B^q) + d(a_p, a_{p-1}) & + \nu & + \lambda \\ \delta^{TW}(A^{p-1}, B^{q-1}) + d(a_p, b_q) + d(a_{p-1}, b_{q-1}) + 2\nu |p - q| \\ \delta^{TW}(A^p, B^{q-1}) + d(b_q, b_{q-1}) & + \nu & + \lambda \end{cases}$$

The first row represents compressing sequence A , the third compressing sequence B (equivalently expanding sequence A), and the middle represents the residual-mismatch cost. While the structural analogy to OM is strong, the manner in which the costings work is different: the equivalent of *indel* operations take context into account: if we compress A the cost depends on the pair of values, $d(a_p, a_{p-1})$ as well as the parameters ν and λ (stiffness and gap penalty). Thus compression is cheaper within a spell in the same state, like OMv, and is cheaper where the previous value is similar, like LOM. The cost of the residual mismatch is also an incomplete analogy to substitution: we look at the mismatch at both t_i and t_{i-1} , and incur an extra penalty depending on how far apart the locations are in the unwarped sequences ($2\nu |p - q|$). Effectively we are paying a cumulating penalty for alignment, both in the compression/expansion operation and in the subsequent comparison. There is little experience with the effects of these parameters, but the impact of varying them is explored below (Section ‘‘Parameterising TWED and Similarity to Other Measures’’).

Thus, though the structure is very similar, we can expect the resulting distances to be different from OM, since on the one hand more context is taken into account in the elementary operations, and the penalty for alignment bears not only in the act of alignment (compression/expansion or insertion/deletion) but also in the comparison of segments warped or aligned out of their original locations.

Some Results

In what follows I compare OM, TWED and X/t using an example data set which consists of 6 years of monthly labour market data for women who have a birth at the end of the second year (derived from British Household Panel Study data). The

Table 5.1 The “linear” and “flat” state-space distance matrices

Linear matrix	FT	PT	UE	Non
Full-time employed	0	1	2	3
Part-time employed	1	0	1	2
Unemployed	2	1	0	1
Non-employed	3	2	1	0

Flat matrix	FT	PT	UE	Non
Full-time employed	0	1	1	1
Part-time employed	1	0	1	1
Unemployed	1	1	0	1
Non-employed	1	1	1	0

state space differentiates between full- and part-time employment, unemployment and non-employment.

Additionally to OM, TWED and X/t, I also present results for Hamming distance, which can be considered as a special case of OM where alignment through *indels* is suppressed. The Hamming distance makes for a very simple “narrative” where the mapping between state-space distances and sequence distances is a simple summing of the distance at each time point: full or partial similarity at the same time. This is an important comparison because where sequences tend to have long spells and differ largely in when exactly transitions occur (as is often the case with lifecourse data), the Hamming distance will be fairly low, and the amount it can further drop if alignment were allowed will be relatively small. It may be that with OM and other measures we gain only a little information (through better distances) at the expense of a much more complicated story about similarity.

We can imagine the four statuses being compared in terms of commitment to the labour market, from full-time employment to non-employment. This suggests a simple, attractive, state-space distance structure, with the four states on a single dimension (hence “linear”) with (for simplicity) equal intervals between them (see Table 5.1, upper panel). Other state space structures are of course possible. However, X/t does not use state-space distance information, it cannot be fairly compared using this structure. Therefore, I here compare X/t with Hamming, OM, and TWED, using another simple state-space cost structure, where all states are equally dissimilar (hence “flat”: see Table 5.1, lower panel). While this permits comparison, it means the other distance measures are a little handicapped by using a neutral state-space distance structure, rather than one that differentiates between states.

Patterns

For each measure, distances are calculated and grouped into clusters using Ward’s method. Eight clusters are chosen for convenience and ease of exposition. Results for the “flat” cost structure are presented in Figs. 5.1, 5.2, 5.3 and 5.4, which show

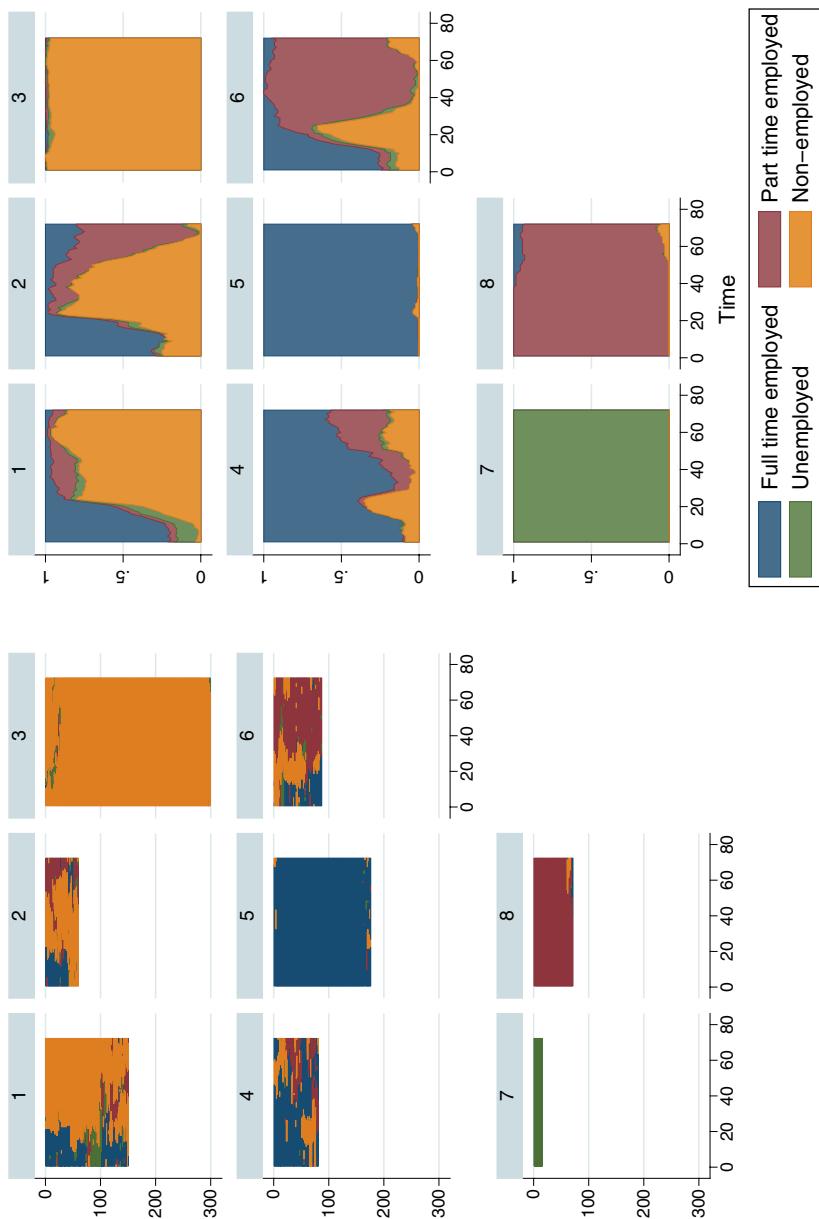


Fig. 5.1 Indexplot (l) and chronogram (r) for Hamming distance, flat state space

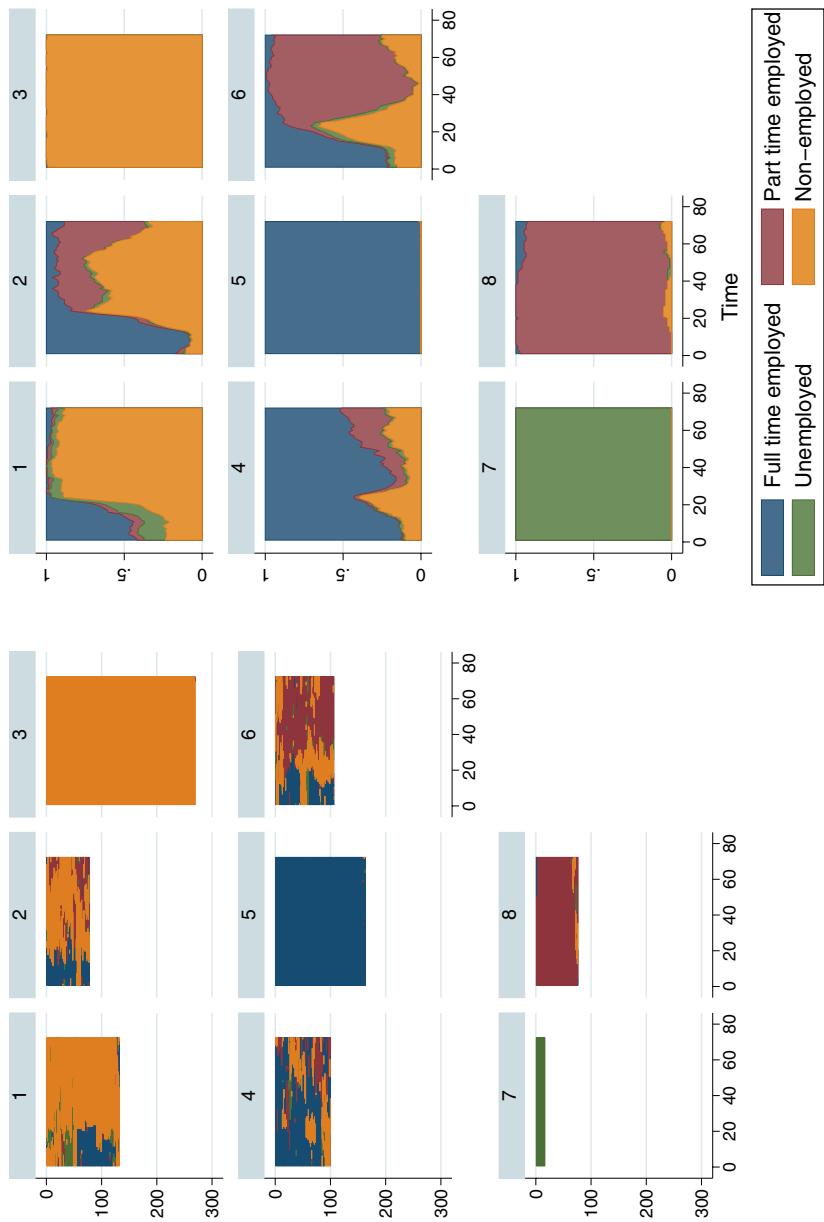


Fig. 5.2 Indexplot (I) and chronogram (r) for OM distance, flat state space

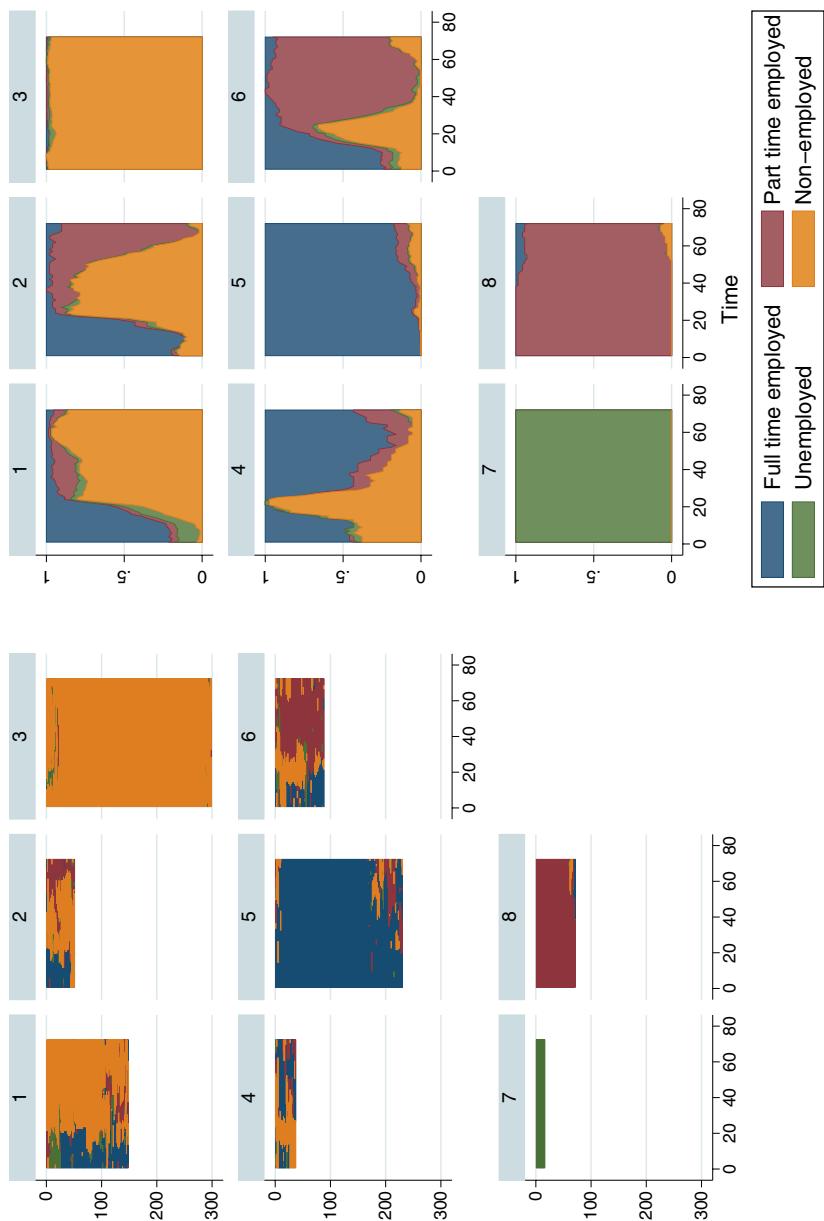


Fig. 5.3 Indexplot (I) and chronogram (t) for TWED distance, flat state space

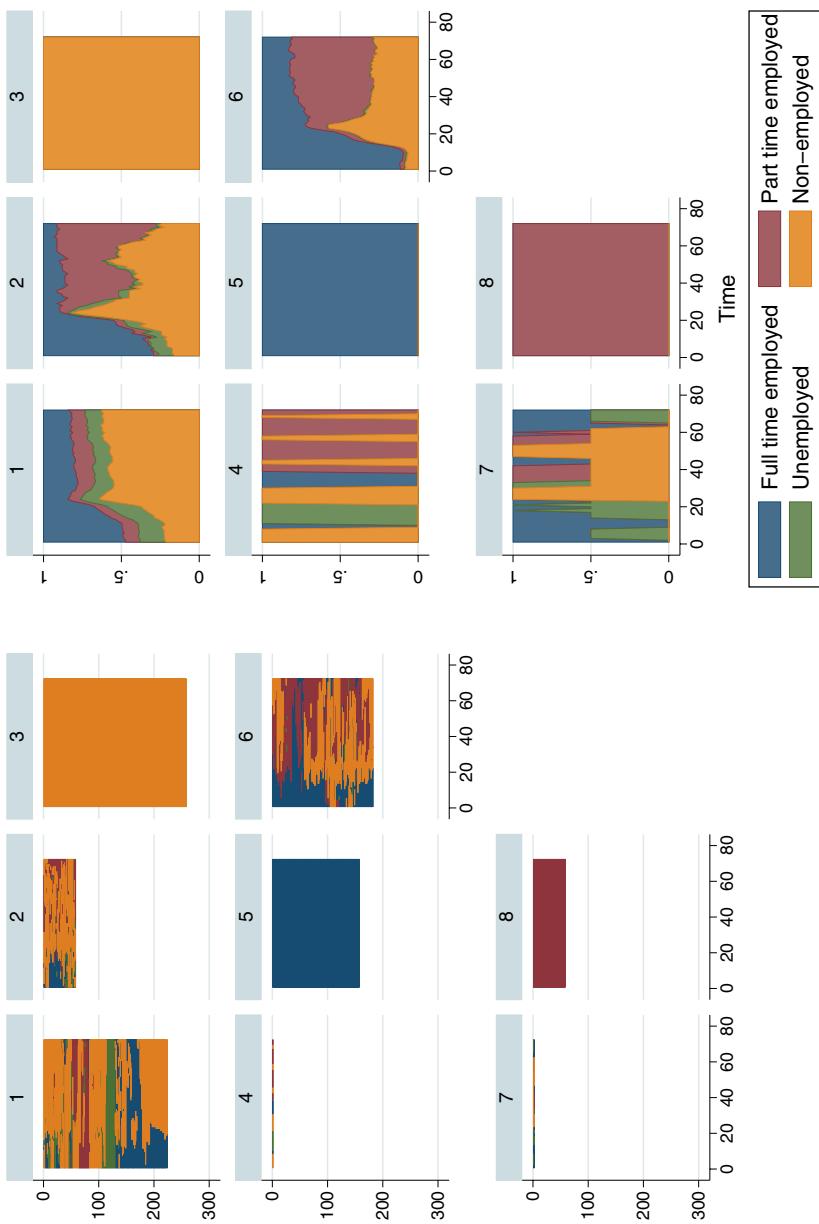


Fig. 5.4 Indexplot (I) and chronogram (r) for X/t distance

indexplots (sorted within clusters by the nested subcluster structure²) and state-distributions (or chronograms). Cluster solutions have been re-ordered to maximise agreement across measures, in so far as possible. While there are clearly differences in the assignment of sequences to clusters, there is a remarkable level of similarity (especially considering that cluster analysis can be relatively unstable). In terms of similarities, all three of Hamming, OM and TWED identify four clusters dominated by a single spell in a single state (clusters 3, 5, 7 and 8), while X/t also identifies three of these (3, 5 and 8, relegating the small set of 100 % unemployed sequences to complex cluster 1). It is also to be noted that for X/t the simple clusters are completely dominated by single-spell sequences, while for the others there is a tendency to also fold in sequences that are only nearly 100 % in a single spell.

For Hamming, OM, and TWED, the four other more complex clusters are quite congruent, though TWED is more insistent that cluster 4 contains sequences involving exit from the labour market around the birth, and then return, and is more inclined than Hamming or OM to add sequences where long spells of full-time employment are mixed with other statuses, to the cluster dominated by single-spell full-time trajectories. It is also worth noting that, compared with a similar analysis with the linear state space (not shown), similarities are stronger across Hamming, OM and TWED. This is possibly because the absence of the strong good or bad matches possible with the linear matrix means that there is less opportunity to benefit from time-displacement, causing OM and TWED to converge on Hamming. However, it is worth noting that the difference between Hamming, OM and TWED is much less than the difference within measures across state-space matrices: switching from the linear to the flat matrix makes a bigger difference than switching between algorithms. It should not be surprising that the state-space distance structure should matter: it creates quite different landscapes, one in which full-time work and non-employment are very distinct, while unemployment and part-time are more similar, and the other in which everything is equally distinct so nothing stands out.

Compared with the other three, Elzinga's measure produces radically different results. Single-spell sequences are naturally maximally distinct (or identical), and so are pulled out in pure simple clusters, without the admixture of sequences that are only close to 100 % in that state. There are three substantial clusters with moderate numbers of transitions, and two of these (2 and 6) correspond relatively well with patterns found by the other measures, though cluster 1 is very heterogeneous (and across the eight clusters, only 71 % of sequences fall in the corresponding OM cluster, 42 % if we exclude the three clusters of simple sequences). Finally there are two tiny clusters containing sequences with high numbers of transitions (and therefore complex sets of subsequences) which are very distinct from all other sequences. It is evident that the number of transitions is very important (clusters 1, 2 and 6 have respectively 2.9, 6.5 and 3.8 spells on average, and the two micro-clusters 11 and 12), and that ever being in the same state is important. Moreover, while duration is

² As will become apparent below, clustering does not recover stable natural clusters for OM and TWED. Thus cluster solutions are somewhat unstable and arbitrary. Nevertheless, the full hierarchical structure of the cluster analysis captures a lot of the information embedded in the distance matrix. Sorting index plots within cluster by the nested subcluster structure (i.e. presenting the dendrogram order), thus makes a lot more useful information available to the eye.

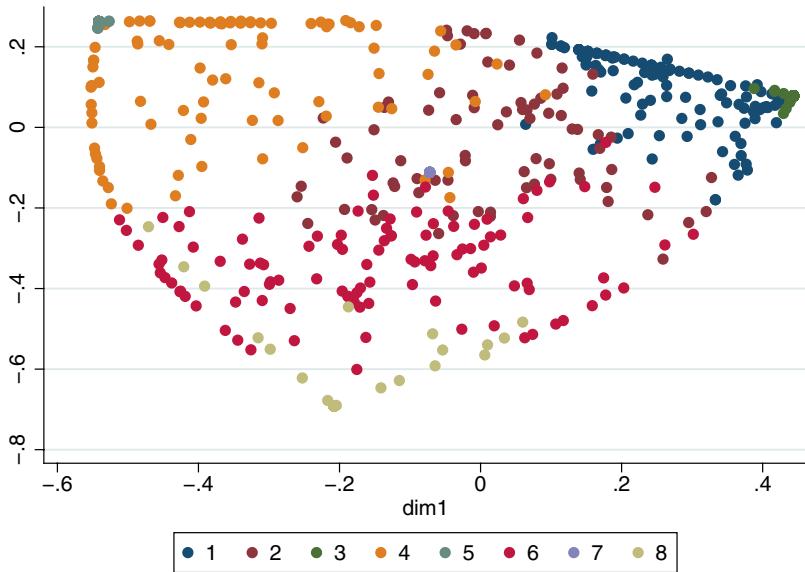


Fig. 5.5 Multidimensional scaling of OM distances, by cluster, flat state space

explicitly brought into consideration, this does not tie time down in the same way as the other measure: i.e., since according to all-common-subsequence measures, $xxABC$, $AwBwC$ and $ABCzz$ will all be equidistant because the measures focuses on order (if weighted by time) and not time, while OM and TWED penalise for temporal displacement. Where the substantive concern is with order, X/t has big advantages, but with life-course sequences such as these, where there is an explicit anchor time point (here the birth at the end of year two, but more typically the start of the trajectory) and a developmental time-scale of some sort (here, changing constraints on labour market activity as the child ages) measures that pick up “full or partial similarity at the same or similar time” produce more useful patterns.

Multidimensional Scaling of Inter-Sequence Distances

The foregoing cluster analysis reflects a typical work-flow with sequence analysis, where distances are converted into a data-driven classification of sequences. We can understand a little more about how the clustering generates the classification by looking at the space implied by the distance. Do the sequences have a lumpy distribution in this space (and hence generate robust clusters)? Can we understand something about the overall pattern of distances: which sequences are distant from each other, which more close? Does this structure emerge in a similar way across measures or are there systematic differences. In this section we look at multi-dimensional scaling of the OM, TWED and X/t distances.

Figure 5.5 graphs the first two dimensions for the OM distances, distinguishing the 8-cluster solution, using the flat state space. The sequences are located in an

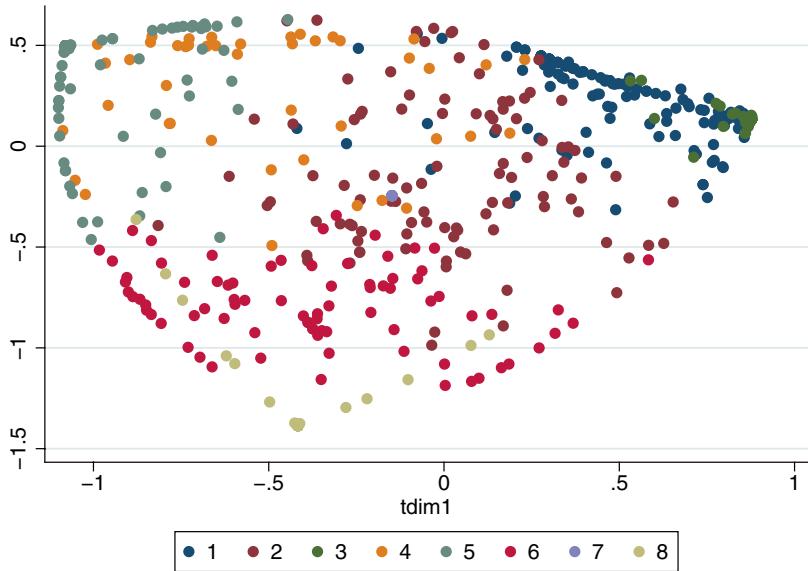


Fig. 5.6 Multidimensional scaling of TWED distances, by cluster, flat state space, $\lambda = \nu = 0.5$

approximately oval cloud, with clear poles marked by sequences that are 100 % in one state (more easily detected with more dimensions). Some sequences appear to form strings: inspection of the data shows these tend to contain one transition, where as the transition occurs earlier they move progressively closer to the pole concerning the second state. However, apart from these structures, the distribution through space is relatively even, such that there is no natural clustering. As is clear, the cluster solution that Ward's algorithm generates does not identify well defined groups. However, it does partition the space into distinct areas, albeit with rather arbitrary boundaries. In that much, we can consider it a useful data reduction strategy but not a successful discovery of natural clusters.

Figure 5.6 shows the corresponding graph for TWED. The detailed shape of the cloud is different, but the qualitative statements relating OM also apply here. The poles and strings are apparent, as is the absence of natural clustering.

The structure apparent in the MDS of the X/t distances is very different. Fig. 5.7 shows shows a single dense cluster and a spread of more remote points (in fact, a small number of extreme outliers have been excluded). The most remote points consist of sequences with high numbers of spells (the outliers have very many spells) and, as we move towards the cluster, sequences get simpler. In terms of the number of spells in each sequence (not shown), the very simplest sequences are in the top right, and as we move left the number of spells increase steadily.

In this measure, sequences with many spells are further from all other sequences, particularly those with fewer spells, but also from other high-spell sequences with

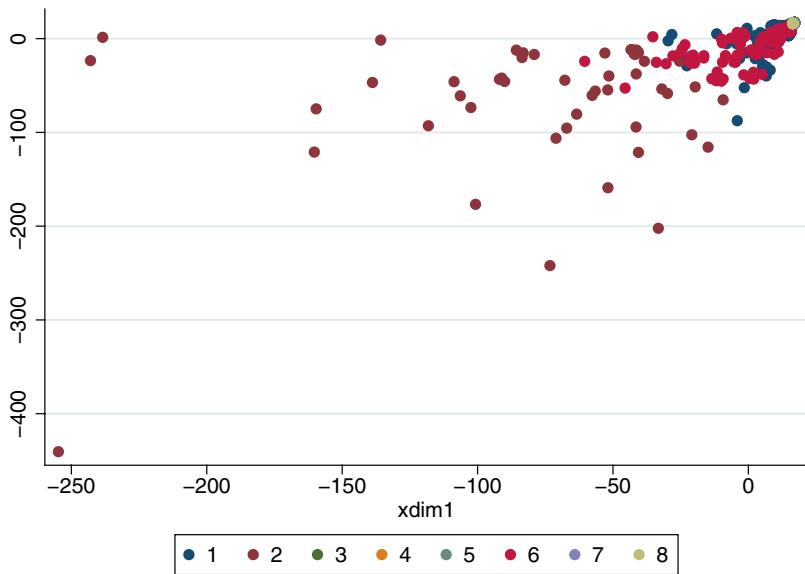


Fig. 5.7 Multidimensional scaling of X/t distances, by cluster (excluding outliers)

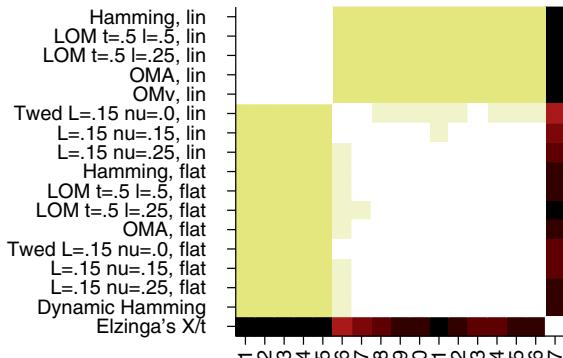
different patterns. Low-spell sequences, and *a fortiori* single-spell sequences have lower distances, but are still distinct from each other (when different). Because simple sequences are common, they will tend to form clear clusters despite their relatively low distances to other sequences—they form dense natural clusters in the space, clearly separated from each other albeit by small distances. Correspondingly, while high-spell sequences are distant from others, they are sparsely distributed in the space and therefore tend to make quite loose clusters, which can be affected by number of spells at least as much as substantive pattern.

The main conclusions we can draw from this analysis is that OM and TWED are relatively similar and do not show a natural cluster structure, while X/t gives a very different result, with strong natural distinctions between clusters of simple sequences, but lower power to distinguish complex sequences.

Correlation Analysis

As we have seen, cluster analysis is a somewhat unstable technique unless there are strong natural clusters in the data, giving results that are sensitive to small changes in parameterisation or sampling. However, the underlying distances are not the cause of this instability, so another approach is to compare the distances directly. That is, comparing cluster solutions across measures can be messy and risks being subjective, but we can get an alternative overview of the various measures by looking at correlations between the distance matrices. Focusing on a single number makes

Fig. 5.8 Correlations between a variety of measures, BS data. LOM and TWED are present with multiple parameterisations; in the case of LOM, t refers to the β “time” parameter, and l to the “local” α parameter; for TWED, L refers to the gap penalty, λ , and ν to the stiffness parameter, v . Results are shown for linear and flat state spaces



for a less rich but more tractable comparison, and it is clearly the case that the difference between a pair of measures has many more than one degree of freedom, and differences between two highly correlated measures may well be important. However, the extent to which measures agree about sequence pairs is important, and looking at their correlation is clearly informative.

Tractability, moreover, allows us to compare larger numbers of measures for context: here as well as OM, TWED (with three parameterisations) and X/t, I include LOM (with two parameterisations) and OMv. I also include Lesnard’s dynamic Hamming measure (Lesnard 2010), which is a version of the Hamming distance that calculates the state-space matrix dynamically from the time-dependent transition rates. Where relevant, I apply both the linear and flat state-space matrices.

The resulting correlation matrix is presented in Fig. 5.8 as a heatmap. Rows and columns have been sorted manually to maximise coherence. To get a sense of the scale, the lightest cells correspond to coefficients of 0.98 and higher, intermediate about 0.85, and the darkest in the range from about 0.1 to 0.3.

The first thing to emerge clearly is that Hamming, OM, LOM and OMv (with the same state-space matrix) produce distances with very high correlations, between 0.984 and 0.999. This is true for the linear and flat matrices, but not across matrices: the state-space distance structure matters quite a bit more than the measure, within this set of measures. Effectively, for most pairs of sequences, particularly simple ones, the difference between Hamming and the more complicated distance measures is negligible or zero. This is not to say that, for the pairs where the measures do differ, the time-dislocating measures do not offer significant added value.

The second thing to emerge is that for the flat matrix, TWED is also part of this group, highly correlated with Hamming and OM. Related to this is the fact that for flat *and* linear state-space matrices, TWED is close to the OM/Hamming results for the flat matrix, though the correlation is slightly lower across matrices (e.g., for TWED with the first parameterisation and the linear matrix, the correlation with OM with the flat matrix is 0.937, versus 0.983 when both TWED and OM use the flat matrix). Thus we see evidence that TWED is less affected by the state space matrix for the parameterisations used. Dynamic Hamming also falls in the flat-matrix group, presumably because the dynamic state-space values calculated from

the transition rates are not distinctly different from the flat matrix (i.e., there is no pair of states with distinctly higher transition rates; this can clearly differ with other data sets).

The final piece of information to be gleaned is the complete distinctness of the X/t measure. Its correlations range from 0.078 with OM (linear matrix) to 0.297 with TWED (linear matrix, first parameterisation). It is only with TWED that the correlation exceeds 0.13, suggesting that TWED respects order in a manner like, but rather weaker than, X/t, and more than the other measures.

Parameterising TWED and Similarity to Other Measures

The correlation analysis suggests that while there is a big gulf between X/t and the other measures, TWED may serve as an intermediate measure, though closer to OM et al. than X/t. However, TWED is not well understood, and in particular the effects of its parameterisation is not clear. The v “stiffness” parameter bears on all three operations, making compression and expansion more expensive, and penalising comparisons proportionately with their displacement, and the λ gap penalty adds further to the cost of expansion and compression. However, it is not clear what the consequences of the parameterisation is—while high values will reduce TWED to Hamming distance, it is not clear what lower values do. We have seen so far that changing v between 0.0 and 0.25, holding λ constant at 0.5, has relatively little effect on the correlation of the distances with other measures (values of the parameters not much higher than these constrain the measure to Hamming; this is analogous to raising *indel* with OM, but takes effect at lower values).

To explore this issue, I present an analysis of the correlation structure, comparing OM and X/t, with TWED with multiple parameterisations: each parameter takes values between 0 and 1 (0.0, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5 and 1.0). Figure 5.9 plots the correlation of TWED with OM (using the flat matrix) and X/t. For high levels of either parameter (where high means approximately 0.2 to 1), TWED is very close to OM and is almost indistinguishable as either value approaches 1 (in fact, each is near indistinguishable from Hamming), but for lower values the correlation with OM declines slowly at first but then more quickly, and the correlation with X/t rises initially quickly but finally slowly, peaking at about 0.6 as the parameter values reach zero. We thus see that varying the parameters of TWED can yield distances that differ strongly from OM, in a way that responds to dimensions of sequence similarity that X/t detects (the implication is order) but without abandoning the fundamental sensitivity to timing, albeit with an elastic interpretation.

In this data set, the two parameters seem to be almost interchangeable, such that their sum seems to be driving most of the pattern. This suggests that expansion and compression are doing a lot of the work, particularly at lower values (i.e., further from OM, closer to X/t), since both parameters bear on those operations, while the stiffness parameter, v , alone bears on the residual match operation. With other data

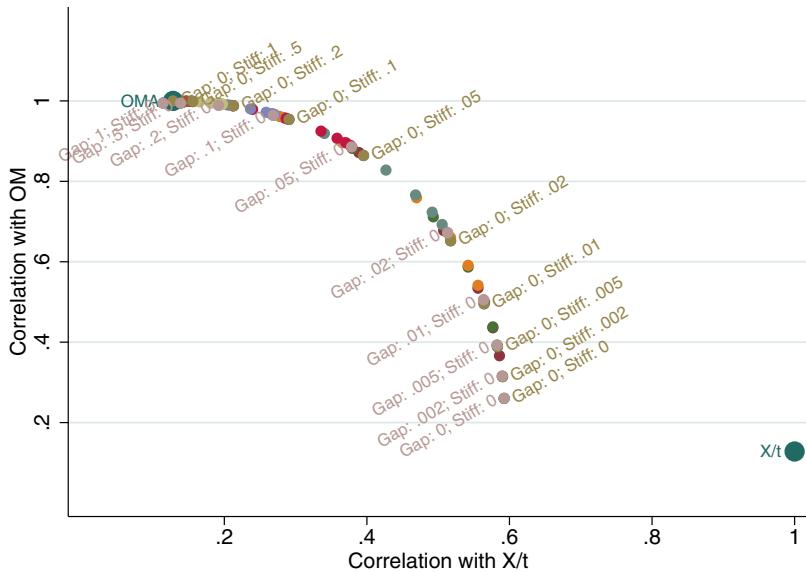


Fig. 5.9 Correlations between TWED, OM and X/t for varying gap and stiffness parameters. Selected points labelled with their λ (gap penalty) and v (stiffness) parameter values

sets, the overall pattern is the same, with a little more evidence of independent effects of the two parameters.

What the Analysis Tells us

The comparison between measures using the mothers' labour market data throws a good deal of light on the difference between measures. One striking finding is that there is little difference between many of the measures, and that the state-space distance matrix is often far more important than the algorithm. The difference between the time-dislocating measures and Hamming can be very small, viewed through the lens of the correlation of the distance matrices, though cluster analysis can amplify small differences (not always pathologically, either, one hopes; such small differences are likely to have to do with sociologically interesting, higher-transition sequences).

The TWED measure has strong commonalities with OM at an algorithmic level, though its narrative is different. We see that at high parameter values it is quite similar to the OM/Hamming complex, though its sensitivity to the state-space matrix seems lower. The combinatorial duration-weighted X/t measure is very distinct: while it picks up some of the same broad patterns at the cluster analysis level (but largely by virtue of isolating the single spell sequences) it gives a very different overall account of sequence similarity, and correlates poorly with all the other measures. TWED, however, provides a bridge, approaching X/t at very low levels of its

gap and stiffness parameters, suggesting that at high levels of the parameters it is picking up on the time-structure of the sequences in a manner very similar to OM et al. and at low levels, picking up on order information in a manner quite similar to X/t.

Conclusion

Inter-sequence distance measures are deterministic algorithms which map state space onto sequence space according to a set of fixed rules. We can understand the mapping in terms of the algorithmic rules, though it is hard to predict what the consequences will be for a given empirical domain or data set. We can also understand the mapping in more heuristic terms, e.g. by considering the algorithm as a model of the data generating process, or as indirectly capturing elements of inter-sequence relationships in a way we can describe more generally.

There may be situations where a distance measure's operation corresponds closely enough to the data generating process that it can be seen as modelling it, but these are rare. Some commentators see the OM algorithm as directly modelling the processes of divergence of DNA across generations. This is not correct: though there are some similarities, OM's operations do not accurately model DNA transcription and mutation, but are driven by algorithmic tractability. What is important for OM's utility in the context is (i) that DNA is indeed a token string and (ii) that DNA does have a pattern of relatedness (closeness due to inheritance) that will show up as potentially displaced matching patterns of tokens. OM is not like, say, the Fisher–Wright model, which does explicitly attempt to model the effect of inheritance on population genetics; rather it is a computationally-efficient heuristic that captures significant features of the phenomenon.

As a heuristic, it will also work for other domains, more or less well depending on the extent to which the token-string representation, with operations on contextless tokens, captures the nature of similarity. Even where it does not work perfectly, it will still capture similarity.

What this paper has been concerned with is both algorithm and the story or narrative with which we can represent the heuristic, and the applicability of distance measures to life course data. In a general sense the narrative is one of mapping between state space and sequence space, starting logically with the simplest mapping, the Hamming distance (full or partial similarity at the same time), and then moving on to measures that allow time-dislocation in similarity. It has become evident that part of OM's success in the sociological sequence analysis (despite worries about its non-applicability) has to do with its achievement of a difficult job, to wit, efficient production of metric distances between sequences. We have seen alternatives that attempt to stay in the same paradigm as OM fail for technical reasons. LOM and OMv attempt to make OM more relevant to lifecourse data by taking account of context, but lose the metric property. The combinatorial approach of X/t has a radically different narrative (the same states in the same order), and while this nar-

rative is sociologically very attractive, in practice the focus on time as order means a weaker connection with time as calendar or developmental scale; calendar and development are often important in life course perspectives. And while X/t also tends to make some very strong distinctions, it also fails to separate complex sequences where it would be sociologically attractive to do so.

In terms of narrative and results, TWED offers a real alternative. While the mechanics of the algorithm are extremely similar to OM, with direct mapping between the three elementary operations, it has a different genealogy: while it works with sequences of discrete tokens, it is consistent with and derived from a continuous-time perspective. Because of this genealogy and because the implementation explicitly takes context into account, it offers us an escape from the contextless-token problem while providing a metric distance. As we have seen, empirically it can range between results that are very close to Hamming and OM, when its gap and stiffness parameters are high, to results that move strongly towards the order-sensitive results of the duration-weighted combinatorial measure, X/t . Thus it gives us a way of moving between calendar- and order-oriented time.

In some ways the result of this investigation is negative: in so far as we started with a worry about the validity of OM for lifecourse data, a part of the finding is that many (but not all) different measures produce remarkably similar results (and that the structure of the state space can matter much more than the algorithm). That is, OM may be technically inadequate but its problems have little consequence; its narrative produces results consistent with quite dissimilar narratives. However, it has also been shown that measure does matter, that differences do exist, particularly between time as scale and time as order, and that there exist viable alternatives to OM's discrete-token string-editing paradigm.

References

- Abbott, A. (1995). A comment on “Measuring the agreement between sequences”. *Sociological Methods and Research*, 24(2), 232–243.
- Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians’ careers. *American Journal of Sociology*, 96(1), 144–85.
- Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology. *Sociological Methods and Research*, 29(1), 3–33.
- Barban, N., & Billari, F. (2012). Classifying life course trajectories: A comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society Series C*, 61(5), 765–784.
- Dijkstra, W., & Taris, T. (1995). Measuring the agreement between sequences. *Sociological Methods and Research*, 24(2), 214–231.
- Elzinga, C. H. (2003). Sequence similarity: A non-aligning technique. *Sociological Methods and Research*, 32(1), 3–29.
- Elzinga, C. H. (2005). Combinatorial representations of token sequences. *Journal of Classification*, 22(1), 87–118.
- Elzinga, C. H. (2006). *Sequence analysis: Metric representations of categorical time series*. Amsterdam: Free University of Amsterdam.
- Elzinga, C. H., & Studer, M. (2013). *Spell sequences, state proximities and distance metrics*. Manuscript.

- Elzinga, C. H., & Wang, H. (2013). Versatile string kernels. *Theoretical Computer Science*, 495, 50–65.
- Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Halpin, B. (2010). Optimal matching analysis and life course data: The importance of duration. *Sociological Methods and Research*, 38(3), 365–388.
- Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2), 147–160.
- Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods and Research*, 38(2), 235–264.
- Kruskal, J. B., & Liberman, M. (1983). The symmetric time-warping problem. In D. Sankoff & J. B. Kruskal (Eds.), *Time warps, string edits and macromolecules* (pp. 125–161). Reading: Addison-Wesley.
- Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods and Research*, 38(3), 389–419.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklady*, 10(8), 707–710.
- Levine, J. H. (2000). But what have you done for us lately? Commentary on Abbott and Tsay. *Sociological Methods and Research*, 29(1), 34–40.
- Lovaglio, P. G., & Mezzananza, M. (2013). Classification of longitudinal career paths. *Quality and Quantity*, 47(2), 989–1008. doi:10.1007/s11135-011-9578-y.
- Marteau, P.-F. (2007). Time warp edit distance with stiffness adjustment for time series matching. *ArXiv Computer Science e-prints*. eprint: cs/0703033. URL. Accessed 26 Jan 2008.
- Marteau, P.-F. (2008). Time warp edit distance. *ArXiv e-prints*. URL. Accessed 8 June 2008.
- Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453.
- Sankoff, D., & Kruskal, J. B. (Eds.). (1983). *Time warps, string edits and macromolecules*. Reading: Addison-Wesley.
- Wu, L. L. (2000). Some comments on “Sequence analysis and optimal matching methods in sociology: Review and prospect”. *Sociological Methods and Research*, 29(1), 41–64.

Part II

Life Course Sequences

Chapter 6

New Perspectives on Family Formation: What Can We Learn from Sequence Analysis?

Anette Eva Fasang

Introduction

Initial skepticism about the applicability of sequence analysis in the social sciences triggered rapid technical progress of the method over the past decade (Aisenbrey and Fasang 2010). The development of powerful user-friendly software in stata (Brzinsky-Fay et al. 2006) and R (Gabadinho et al. 2011), as well as the progressively increasing availability of large-scale longitudinal data certainly were crucial for this development. Given its proliferation in applied social science, it is due time to re-examine what sequence analysis has done for us lately, and whether it helps us to inform old ideas and core theoretical debates in the social sciences.

This chapter engages the field of family formation, one of the more vibrant areas of sequence analysis applications to date, to scrutinize how sequential thinking and sequential methodology can tighten the link between theory and empirical applications in life course research. First, Abbott's theoretical foundation of sequence analysis, captured in the notion of narrative positivism (1988, 1992), is revisited and related to basic premises of the life course paradigm (Elder et al. 2003; Mayer 2009). Second, I review the sequence analysis literature in three areas of family formation research: (a) *multidimensional lives*, (b) *linked lives* within families and social networks, and (c) the *de-standardization* and *pluralization* of family formation. Third, a case study is presented on the de-standardization of family formation before and after the German Reunification. This application seeks to demonstrate how overlapping premises of narrative positivism and the life course paradigm can be fruitfully incorporated in a theory-driven research design using sequence analysis. Specifically, the assumption that interactions are more important than main effects is taken into account by theorizing de-standardization of family formation as the outcome of constellations of macro-structural factors in the two German

A. E. Fasang (✉)
Social Sciences, Humboldt University Berlin, Berlin, Germany
e-mail: fasang@wzb.eu

Demography and Inequality, WZB Social Science Research Center Berlin (WZB),
Berlin, Germany

sub-societies. To implement this empirically, I propose a new approach to specify within-group and between-group sequence distances.

Foundations of Sequential Thinking

The basic circle of inductive and deductive reasoning in scientific progress assumes a cross fertilization of theory and empirical research. Ideally, theoretical questions motivate the development of appropriate methodology, which is used to generate empirical evidence to inform the respective theoretical question. According to Andrew Abbott (1992), mainstream quantitative sociology of the past century has largely failed this ideal. Instead, linear models dominated the field and construed the social world in terms of “general linear reality” (GLR) (Abbott 1988). This has led to the neglect of process in favor of a focus on associations between independent and dependent variables that are attributed causal meaning post-hoc. Abbott continues to argue that this dominance of linear models has gradually imposed on and constrained sociological theorizing (Abbott 2001). He instead calls for bringing back “process” into sociological theory and empirical research. On top of highlighting the value of description in itself, Abbott proposes to rethink causality in temporal terms as a process of emergence that can be reconstructed with fundamentally descriptive procedures (see Savage 2009, p. 161). Sequence analysis is one of several methodological approaches that serve this purpose by unraveling structure and patterning in complex longitudinal sequences.

Abbott was certainly neither the first nor the only one to criticize linear models in quantitative sociology (e.g. Abell 2004). Critical voices come from qualitative and ethnographic research traditions, as well as from mechanism-focused analytical sociology (Hedström and Bearman 2009), proponents of comparative techniques for small N studies including Qualitative Comparative Analysis (QCA) (Rihoux and Ragin 2009), and Abell’s narrative method (Abell 1993, 2004). Abbott fully develops his critique of GLR by developing narrative positivism as an alternative against the premises of GLR (1988, 1992, 2001).¹ As a result, the defining components of narrative positivism counter the defining components of GLR: GLR assumes that the social world is made up of fixed entities, attributes have only one causal meaning, and main effects are more important than interactions. Narrative positivism counters that entities are in constant flux through mergers and dissolution (e.g. think of changing household compositions), a single attribute can have many causal meanings (e.g. gender), and constellations of attributes, i.e. multiple-way interactions, by far outweigh the sociological significance of main effects.² Arguably, Abbott’s most forceful critique of GLR remains its neglect of time and temporality, in particular the assumption that things happen in discrete units of time and the

¹ Narrative is understood in a generic sense as a process or story (Abbott 1992, p. 428) and does not involve complexity of meaning that is inherently unformalizable as in qualitative research. It instead refers to a processual, action-based approach to social reality and formalization should be used to the extent that it advances sociological insight.

² For a full account of narrative positivism see Abbott (1988, 1992, 2001).

order of events does not alter their causal meaning (2001). In addition to a focus on temporality and process, this chapter engages the preeminence of constellations of conditions over main effects in sociological theory and empirical research.

Narrative Positivism and the Life Course Paradigm

In 1992, Abbott challenged the life course literature as a field with an obvious gap between the conceptual importance of process and its empirical neglect (p. 429). Since then much has happened, including the continual refinement of event history models, various multilevel and panel regression models, latent class and latent trajectory models, and the establishment of sequence analysis as a complementary method in life course research. In fact, the life course literature has become the primary area of technical advancement and methodological innovation of sequence analysis (e.g. Bonetti et al. 2013; Brzinsky-Fay and Kohler 2010; Gabadinho et al. 2011; Gauthier et al. 2010; Studer et al. 2011).

Narrative positivism and the life course paradigm clearly share an emphasis on the pivotal importance of the timing and sequencing of events (Abbott 2001; Elder et al. 2003; Mayer 2009; Settersten and Mayer 1997). To date, life course research using sequence analysis has mainly engaged Abbott's work on time and temporality, including his conceptualization of holistic trajectories and turning points (Abbott 2001). Beyond timing and sequencing, basic premises of narrative positivism resonate with other fundamental concerns of the life course paradigm that have been less explicitly noted. One of them is the preeminence of multiple-way interactions against “[...] the odd idea that main effects exist. They are, of course, mere analytical conveniences, not realities” (Abbott 2001, p. 12).

In the life course paradigm, multiple-way interactions are fundamental on all levels of analysis: *multidimensionality* assumes that multiple dimensions of the life course, e.g. family and employment and residential moves, unfold in interaction with one another on the micro level. The notion of *linked lives* implies multiple-way interactions between individuals within networks on the meso-level, such as households or schools (Elder et al. 2003). The premise that historically specific macro-structural contexts jointly shape the *de-standardization* and *pluralization* of collective life course patterns is essentially an assumption of multiple-way context interactions on the macro level (Mayer 2004).³

Despite a notable overlap in the fundamental premises of narrative positivism and the life course paradigm, Abbott's critique of GLR is certainly not essential in showing the usefulness of sequence analysis for family formation research. Sequence analysis has obvious descriptive value as an inductive exploratory pattern search technique (Billari 2001a). However, narrative positivism marks the broader epistemological origin of sequence analysis in the social science. Incorporating its premises, particularly those resonating with the life course paradigm, can thus be useful when moving beyond exploratory description toward using sequence analysis to assemble descriptive evidence more targeted at informing theoretical nar-

³ Cross-level interactions, often of primary theoretical interest, add another layer of complexity.

tives. To illustrate this argument, the remainder of this chapter reviews the sequence analysis literature on family formation and presents a case study on family formation before and after the German reunification that explicitly takes into account constellations of macro-structural conditions (multiple-way interactions) as determinants of family formation.

Family Formation from a Sequential Perspective

Next to employment trajectories (Biemann et al. 2011; Widmer and Ritschard 2009), and some applications on retirement (Fasang 2012; Han and Moen 1999), transitions to adulthood, including family formation and school-to-work transitions, have been one of the most vibrant fields applying sequence analysis (Buchmann and Kriesi 2011). From a life course perspective, three core concerns in the study of family formation are (a) *multidimensional lives*, mostly parallel family and employment trajectories, (b) *linked lives* within families and social networks, and (c) the *de-standardization* and *pluralization* of family formation.

To date, most studies use sequence analysis to measure the de-standardization and pluralization of family formation. A few studies focus on multidimensional lives (Aassve et al. 2007; Gauthier et al. 2010; Pollock 2007; Robette 2010), whereas applications that address linked lives are scarce (Fasang and Raab 2013; Liefbroer and Elzinga 2012). Subsequently, I briefly introduce each of these concepts and review the respective literature. Most of these studies offer new insights on family formation with thick descriptions that pin down population level puzzles in an exploratory fashion. Some seek to explain the observed patterns with largely descriptive evidence rather than with statistical reasoning.

Multidimensional Lives

Lives are multidimensional in that they consist of several parallel processes. Gender and work scholars analyze parallel family and employment trajectories. Transitions to adulthood are generally conceptualized as a composite of household living arrangements, education, employment, and family formation (Billari and Liefbroer 2010; Buchmann and Kriesi 2011). Family formation itself covers at least two dimensions—relationship status and parity.

Several studies have looked at family formation alongside employment trajectories using sequence analysis without explicitly considering joint patterning in the two (e.g. Widmer and Ritschard 2009). Attempts to consider multiple dimensions in a combined alphabet of states start as early as Dijkstra and Taris' (1995) and Elzinga's (2003) studies on transitions to adulthood including residential, educational, and employment trajectories (see also Robette 2010; Stovel et al. 1996).⁴

⁴ Han and Moen (1999) were the first to calculate separate distance matrices for each dimension and the adding them up in an application on retirement.

Further elaborations are proposed in Aassve et al. (2007), Pollock (2007), and Gauthier et al. (2010)—all of who consider family trajectories as one dimension in their example applications. Unfortunately, they do not consistently refer to one another and as a result the literature has remained fragmented. Whereas Aassve et al.’s (2007) innovation lies primarily in a tree-based approach using medoids for the visualization of multidimensional sequences, Pollock (2007) and Gauthier et al. (2010) propose more sophisticated accounts for the cost specification when calculating distances between multidimensional sequences. Aassve et al.’s (2007) analysis of British women’s joint employment and family trajectories shows that women who combine work and family have more complex trajectories and cluster into more heterogeneous groups than women who specialize in either work or family. These findings line up nicely with Moen and Sweet’s (2004) thesis that women’s efforts to combine work and family generate increasingly complex life courses (Aassve et al. 2007).

When introducing multiple sequence analysis (MSA), Pollock (2007) analyzes the four-way sequences of employment, housing, marital status, and the responsibility of taking care of dependent children. The innovation of MSA lies in accounting for multidimensionality not simply by combining different dimensions into one unified state space, but instead by specifying separate substitution costs for each dimension and aligning them simultaneously. Pollock’s (2007) results suggest that the experience of social housing is the decisive factor in differentiating these multidimensional life course trajectories. Gauthier et al. (2010) systematize and elaborate on Pollock’s (2007) approach under the label of multichannel sequence analysis (MCSA) in an analysis of occupational and family trajectories. Their approach also specifies separate substitution cost matrices, and additionally allows for weighing each dimension, such that it contributes differentially to the overall distance. In a systematic comparison of different methods for multidimensional pattern search, they conclude that MCSA performs best when the dimensions are interdependent (statistically correlated monochannels). Moreover, MCSA seems more resistant to increasing noise in data than other approaches. Müller et al. (2012) apply MCSA to jointly study occupational and family trajectories in a small clinical sample of people with psychiatric disorders. They show that the patterns observed for the clinical population differ sharply from those observed in the general population. For instance, a trajectory of a stable family situation and regular work activity—the standard in the general population—is associated with more and not less mental health symptoms, such as depression, among the clinical population.

In sum, multiple and multichannel sequence analyses classify holistic longitudinal experiences in terms of interactions between the dimensions considered (Pollock 2007, p. 176). Multichannel sequence analysis thereby embraces both the focus on process/time and on interaction/interdependence of multiple life dimensions, which is formulated in narrative positivism and as well as in the life course paradigm. Applications to date are largely exploratory but have also generated descriptive evidence that clearly speaks to theoretical arguments (e.g. Aassve et al. 2007).

Linked Lives

“Lives are lived interdependently and socio-historical influences are expressed through this network of shared relationships” (Elder et al. 2003, p. 13). Some links have received particular attention in family formation literature: links between parents and their children addressing intergenerational transmission (e.g. Amato and DeBoer 2001), links between siblings to illuminate shared parental background influence and sibling contagion (e.g. Lyngstad and Prskawetz 2010), and links between partners, for instance in the division of domestic labor (Prince Cooke 2004). Recent configurational approaches highlight more complex network structures, including extended networks of functional and biological kin and non-kin (Widmer 2010), as well as extended multigenerational bonds (Bengtson 2001).

Sequence analysis can be employed when studying linked lives in family formation in three ways. First, by taking the couple or household as the unit of analysis and studying joint patterns (Lesnard 2008). Second, by examining dyadic distances between core dyads of shared relationships, i.e. parent-child, siblings, or partners. Third, for a structural network perspective that places distances among several dyadic sequences within the family system in relation to one another, albeit this has received no actual empirical application so far. Lesnard’s (2008) study on off-scheduling among dual-earner couples provides an exemplary application of the first option by taking couples as the unit of analysis. The couples’ sequences are defined as 24-h days of joint time use, separating joint family and work time. He concludes that off-scheduling reduces conjugal and parent-child time and thereby threatens family solidarity.

The second option of targeted dyadic comparison is a nascent but rapidly growing field. Dyadic sequence distances of clearly specified dyadic units, such as a parent-child dyad, have a more intuitive interpretation than pairwise sequence distances that contain information on collective life course patterns of a population. Further, dyadic sequence distances bridge more readily to mainstream regression based methodology. Just like difference values between dyad members, dyadic sequence distances can easily be analyzed as dependent or independent variables in a dyadic regression model (Card et al. 2008; Kenny et al. 2006). A pioneering application of dyadic sequences in family research explored adolescents’ hypothetical ideal relationship sequence (e.g. going out alone—holding hands—kissing—intercourse) to their actual relationship experiences reported in a later survey wave (Brückner 2007).

Two recent studies compare family formation trajectories of parents and their children in a dyadic sequential approach (Fasang and Raab 2013; Liefbroer and Elzinga 2012). Liefbroer and Elzinga (2012), conclude that in spite of profound social change, including the de-standardization of family formation over the past few decades, there is still similarity in parents’ and their children’s family formation. Fasang and Raab (2013) apply multichannel sequence analysis (MCSA) by taking the parent’s and child’s trajectories as separate dimensions of a dyadic intergenerational sequence. In addition to a group that has strong transmission, where the

parent's and children's family formation is very similar, they also found a group of intergenerational contrast, where children show very different family behavior than their parents. An emotionally distant and conflict-ridden parent-child relationship during adolescence increases the likelihood of intergenerational contrast in family formation. This contrast group is often neglected in the intergenerational transmission literature and could only be identified using a dyadic multichannel approach. The standard approach of calculating average coefficients of parental influence on the timing of the child's family behavior will obscure a polarized pattern into either similarity or contrast. The sequential approach thus offers a new impetus for theory building by nuancing the phenomenon of intergenerational similarity and contrast.

Raab et al. (2013) combine sequence analysis with a sibling design and compare family formation trajectories of siblings and randomly paired persons to disentangle family background effects on family formation. Robette et al. (2013) propose a new dyadic sequence analysis approach to study intergenerational transmission of mothers' and daughters' careers, which is easily transferable to family formation. Pasteels and Mortelmans (2013) use dyadic sequence data on both divorcees' subsequent family trajectories to study life course dynamics after separation.

When going back to the premises of narrative positivism and the life course paradigm, the applications above include a more comprehensive account of the processual nature of family formation, and examine how they unfold in the context of linked lives. So far, the focus is on rather straightforward dyadic sequence constellations, but studying sequence distances among broader family units is certainly a promising and viable direction for future research. The fact that most of the literature above is cited as conference papers indicates that we are just beginning to touch upon potential insights from studying dyadic sequences in family formation research.

De-Standardization and Pluralization of Family Formation

The de-standardization and pluralization of family formation primarily addresses two central themes in the life course paradigm: the historical embeddedness and timing in the life course (Elder et al. 2003; Mayer 2009). The first point highlights that life courses are shaped by the specific macro historical conditions in which they unfold. The second captures the importance of the 'social timing' of lives both for the normative evaluation of life courses in a given society and also for ensuing life chances (Elder 1994).

Standardization can be understood as the process "by which specific states or events and the sequences in which they occur become more universal for given populations" (Brückner and Mayer 2005, p. 32; see also Kohli 1985). Conversely, *de-standardization* means that life states, events, and their sequences "become experiences which either characterize an increasingly smaller part of a population or occur at more dispersed ages and with more dispersed durations" (Brückner and Mayer 2005, pp. 32–33). De-standardization is often associated with a pluraliza-

tion of family forms, where pluralization is understood as the development of new types of family formation patterns. One example of this is parenthood in cohabiting relationships instead of marriage. *Pluralization* is defined as an increase in the synchronous number of family states in a given population (Brückner and Mayer 2005; Brüderl 2004).

Whereas de-standardization refers to a quantitative property of family formation (*how dissimilar?*), pluralization captures a qualitative aspect, i.e. the substantive content of family formation patterns (*in which way is it different?*). Both are relational and dynamic concepts: they refer to differences between individual family formation trajectories that evolve over time. Also, they are properties of populations. One person alone cannot be de-standardized or pluralized. As a result, studies on de-standardization and pluralization often face a small N problem, since the units of analysis are usually countries or subgroups with only a few categories each, such as gender or different birth cohorts. For example, in a gender comparison, there will be only two de-standardization values, one for men and one for women.

Most applications of sequence analysis in the field of family formation address these two concepts. The added value of these studies to date primarily lies in refined measurement and operationalization of de-standardization and pluralization with an inductive focus on exploratory, thick description of complex population level processes, rather than explanation. Arguably, this has three reasons: De-standardization and pluralization are highly complex relational and longitudinal concepts; there is a small N issue as they are properties of larger populations, and they always unfold in historically specific socio-cultural contexts.

Studies using a sequential approach report considerable historical and cross-national variation in the degree of standardization and pluralization of family formation (Aassve et al. 2007; Buchmann and Kriesi 2011; Bras et al. 2010; Billari 2001b; Brüderl 2004; Cook and Furstenberg 2002; Elzinga and Liefbroer 2007; Huinink 2011; Lesnard et al. 2010). In line with the second demographic transition argument (SDT), evidence supports a trend toward de-standardization of family formation across the second half of the twentieth century in most Western countries, albeit with great cross-national variation in the onset and timing of de-standardization (Brüderl 2004; Elzinga and Liefbroer 2007; Lesnard et al. 2010). Several studies show that life courses were highly standardized under the regulative communist regimes in Eastern Europe, where de-standardization and pluralization only kicked in after the collapse of these regimes. For instance, Elzinga and Liefbroer (2007) compare family life trajectories of women born between 1945 and 1964 in 19 Western and Eastern European countries. They find a de-standardization of women's family life trajectories across cohorts for all countries, except for Latvia, Poland, Lithuania, and the Czech Republic under the communist regimes. Studies also show that the amount of de-standardization of employment careers is much smaller than often assumed once de-standardization is rigorously scrutinized with longitudinal data and sequential methodology (Biemann et al. 2011; Brückner and Mayer 2005; Widmer and Ritschard 2009). Family formation, in contrast, seems to be the prime

locus of increased life course variation since the middle of the twentieth century in most Western societies.⁵

Beyond these relatively general, but nonetheless significant findings, the particular results of these studies are not easy to typologize coherently. Attempts to identify covering laws that explain the de-standardization of family formation across a larger number of countries have not been very successful. Results tend to deviate from predictions and across studies. For instance, Elzinga and Liefbroer (2007, p. 245) reported that the de-standardization of family formation does not correspond as expected with established welfare regimes, and that variation is almost as high within as across regimes. In contrast, Lesnard et al. (2010) conclude that pathways to adulthood cluster to a considerable extent in established welfare regime typologies and have converged within Europe over time.

Smaller targeted country comparisons or country case studies in the spirit of differential life course sociology (Mayer 2004) might be more promising to uncover historically specific combinations of the contextual driving forces of de-standardization and pluralization of family formation. These might not reveal general covering laws, but can generate coherent narratives for changes in family formation located in specific socio-historical time and space. Following this perspective, Bras et al. (2010) used historical registry-based data for the Netherlands to show a standardization of pathways to adulthood in the second half of the nineteenth century. They argue that industrialization, increasing institutionalization of the labor market, and improving economic conditions jointly set the stage for young people to establish more standardized normative pathways to adulthood. This study thereby embraces multiple-way context interactions as driving forces of the standardization of family formation in formulating a theoretical narrative to account for the observed patterns.

Overall, the research on de-standardization and pluralization of family formation has mostly used sequence analysis to better account for time and temporality in family formation, highlighted both in narrative positivism and the life course paradigm. Beyond this refined measurement and description of the process, few studies, Bras et al. (2010) being a notable exception, have followed through with serious attempts at explanation by theorizing multiple-way context interactions in in-depth case studies. The following study on family formation before and after the German reunification proposes such an in-depth country case study incorporating both the processual nature of family formation and the constellations of macro-structural contexts that shape the de-standardization of family formation.

⁵ Employment patterns have greatly changed as well but much of this change is expressed in different qualitative patterns, such as a higher prevalence of part-time work in some countries and not in an increase in the quantity of variability across the life course.

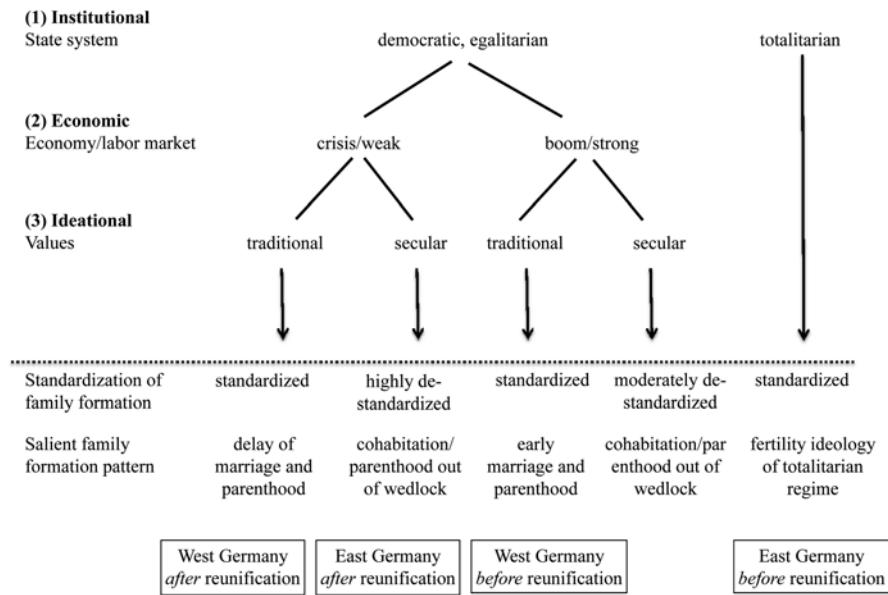


Fig. 6.1 Institutional, economic and ideational determinants of the de-standardization of family formation

De-Standardization of Family Formation Before and After the German Reunification

Three main lines of theoretical argument address how contextual factors structure the process of family formation: institutional, ideational, and economic. Instead of considering them as competing main effects, I use the case of East and West Germany to conceptualize them as constellations of conditions based on the premise that “reality does not happen in main effects but in interactions” (Abbott 1992, p. 439). The German reunification is particularly well-suited to explore this question due to the unique combination of institutional, economic, and ideational contexts in the respective sub-societies before and after the reunification. Subsequently, the three arguments are introduced and related to the situation in divided and reunified Germany (see Fig. 6.1).

1. The *institutional* argument, originally formulated in the life course and welfare state literature (Esping-Andersen 1990; Mayer 2004, 2009), assumes that extensive state regulation generates more continuous and standardized life courses compared to more fluid and de-standardized life courses under restrained government intervention. Strong, regulative states standardize life courses by conditioning access to resources on behavior that conforms to the state ideology. Conversely, de-standardization of life courses is one of the main tenets of theories of individualization, postindustrialism, and postfordism (Beck 1992; Bell

1973; Sennett 1998). The institutional argument is invoked to explain why the SDT has not spread across countries ruled by totalitarian regimes. “The normative and institutional bases of traditional union formation and household structure will systematically weaken in all societies that adopt egalitarian and democratic systems governed by respect for individual choice” (Lesthaeghe 2010, pp. 244–245). An egalitarian and democratic system then is a necessary, but not a sufficient condition for the de-standardization of family formation.

Turning to the case of the German reunification, the regulative communist government in the former East imposed a family ideology aimed at population growth. Pro-natalist family policies conditioned access to state-regulated resources on marriage and parenthood.⁶ There was a strong social norm for women to have children in their early twenties, which was encouraged by generous child benefits and child-care that enabled mothers to work (Huinkink et al. 1995). In contrast, in the democratic social market economy of the former West, a male breadwinner/female homemaker specialization was the core ideology underlying family and housing policies in (Brückner 2004). With the reunification, the regulative pro-natalist regime in the East was suddenly removed and replaced with the West German model (Diewald et al. 2006).

2. The *economic* explanation is most clearly developed with regard to fertility decline in Eastern Europe (Coleman 2004; Kohler et al. 2002). Sudden economic downturn and insecurity depressed fertility and put family formation on hold in the post-communist transition economies (Billingsley 2010; Sobotka 2003). Transferring this argument to the de-standardization of family formation is complicated because there is no obvious relationship between fertility and de-standardization. It depends on larger family formation sequences and their distribution across the population, whether specific fertility levels go along with high or low standardization. On the one hand, if sudden economic decline puts family formation on hold completely, this could standardize a uniform delay pattern. On the other hand, economic downturn and insecurity do not affect everyone equally. People possibly seek out different strategies to adjust family formation to their specific economic situation, which will, for instance, vary with education and family wealth. This could further de-standardize family formation.

In Germany, economic pressures intensified after the collapse of the communist regime, particularly in the turmoil of transition in the East. Unemployment rose sharply during the 1990s, reaching 18.5 % in the former East and 8.4 % in the former West in 2000 (Goldstein and Kreyenfeld 2011, p. 457). Female employment was traditionally high in the communist East, where practically universal public child-care was available and, unemployment was basically non-existent due to a state-controlled labor market. The sudden exposure to unemployment after the system transformation increased economic insecurity but might have also heightened the importance of motherhood as a source of identity, especially among

⁶ There was also broad support for unwed mothers that favored access to housing and parental leave, which set incentives for premarital births and a delay of marriage for the one-year period of these provisions (Huinkink et al. 1995).

lower-educated East German women with poor job prospects (McLanahan and Percheski 2008). In the West, women's employment was consistently lower than in the East. When women were employed, it was often only part-time. This is at least partly related to a limited infrastructure for public childcare, since the male breadwinner ideology underlying the German welfare state foresaw women in the role of homemakers and caretakers. As a result, West German women were more economically dependent on a male breadwinner to sustain themselves and their children, both before and after the reunification.

3. The *ideational* argument (Huinkink 2011; Lesthaeghe 2010) assumes that less-traditional family values will go along with moderate de-standardization, as multiple equally accepted family forms coexist and replace one dominant traditional model of early motherhood secured in marriage. Values were persistently more secular in the East than in the West (Goldstein and Kreyenfeld 2011), suggesting de-standardization in the East once the regulative pro-natalist regime was removed. Massive de-standardization, arguably, is particularly likely when individuals seek various strategies for 'muddling through' a rapidly changing, insecure, and confusing environment, such as in East Germany in the years following the reunification.

Even though East Germany was instantaneously absorbed into the West German model in 1989, neither the economic situation nor attitudes and values have converged among the two German sub-societies as many initially expected (Goldstein and Kreyenfeld 2011): East Germany continues to have poorer economic conditions with higher unemployment, persistently lower wages, and less private property ownership than in the West. In 2008, 74% of East Germans reported to have no religious affiliation compared to 16% of West Germans. Similarly, the proportion of full-time employed mothers was still higher in the East in 2008 (50% compared to 19%) and it was far more common for children aged 0–3 to be in day care than in the former West (41% compared to 12%) (see overview of indicators in Goldstein and Kreyenfeld 2011, p. 457).

Hypotheses

How do these contextual factors jointly shape the de-standardization of family formation? Figure 6.1 shows the expected de-standardization outcomes in the specific institutional, economic, and ideational constellations represented by East and West Germany before and after the reunification. Based on these considerations, I specify three stylized hypotheses on the de-standardization of family formation that correspond to three comparisons before and after the reunification. Clearly, the attempt to explicitly spell out how constellations of conditions shape de-standardization of family formation (Fig. 6.1) necessitates a strong simplification of contextual conditions. Such a general simplified model provides a point of departure for further elaborating various aspects of different contextual conditions in subsequent research.

First, I assume that family formation in totalitarian regimes will be standardized in a pattern that conforms to the family ideology of the totalitarian regime: a standardization of traditional family formation with early marriage and motherhood in the case of communist East Germany (Fig. 6.1). Given a democratic system, the standardization of family formation will depend on the specific constellation of economic conditions and ideational context. In East Germany, secular values coincided with the economic crisis of transition once the regulative pro-natalist regime was removed. In this situation, I assume that East German women will form individual strategies of ‘muddling through’ (Moen and Roehling 2005) economic turmoil and flexibly adopt family formation strategies in accordance with their secular values for instance through motherhood out of wedlock. They might delay costly elaborated weddings, but not motherhood, especially given the importance of motherhood as a source of identity when facing limited labor market prospects. For the *within-group comparison of East Germany before and after the reunification*, I therefore hypothesize massive de-standardization of family formation after the reunification (*hypothesis 1*). This de-standardization will go along with a shift from traditional early marriage and motherhood patterns regulated by the communist regime to alternative family forms of cohabitation and non-marital motherhood. De-standardization will appear particularly drastic against highly standardized family formation under the totalitarian communist regime.

West Germany before the reunification marks a constellation of relative economic growth and stability combined with traditional religiously coined family values. In this situation, family formation will standardize in a traditional pattern of early marriage and motherhood, albeit less standardized than under a totalitarian regime. Stability in employment careers will spill over into more stable and standardized family formation. In addition, welfare services tend to expand in times of economic prosperity and they generally stabilize and standardize life courses (Brückner and Mayer 2005). In the West, a relative economic downturn in the 1990s, even though it was far less drastic than in the East, coincided with persistent traditional values of family formation (Goldstein and Kreyenfeld 2011). In this constellation, women’s family formation will either standardize in a polarized pattern of traditional family formation or a delay pattern. Women will face more pressure to adapt to the labor market and there will be fewer marriageable men who grant economic security in traditional male-breadwinner arrangements. Those who are unable to attain the economic security deemed necessary for marriage and motherhood will uniformly postpone family formation. Family formation is put on hold, as the traditional family ideal of a male breadwinner family cannot be reconciled with economic pressures of the labor market, especially given a lack of public child care. For the *within-group comparison of West Germany before and after the reunification*, I therefore hypothesize a *moderate re-standardization* of family formation after the reunification (*hypothesis 2*). This re-standardization will go along with a shift from

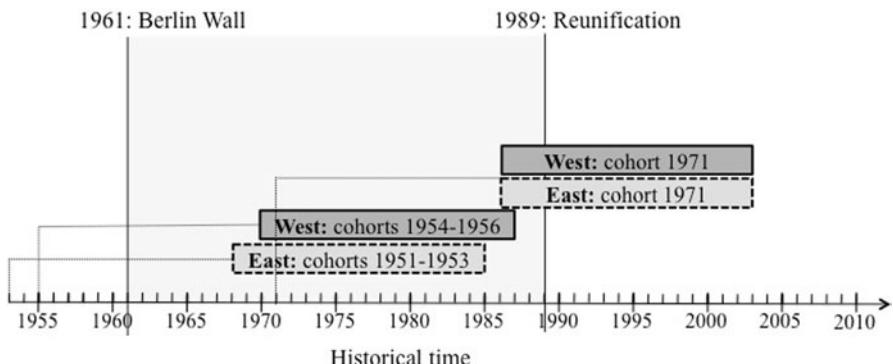


Fig. 6.2 Comparative cohort design of women's family formation between age 15 and 33 before and after the reunification in East and West Germany

traditional family formation to a polarized, either traditional or delayed, family formation pattern.⁷

Finally, as outlined above, economic conditions, ideational orientations, and gender inequalities have not notably converged in the German sub-societies since the reunification (Goldstein and Kreyenfeld 2011, p. 457). For the *between-group comparison of East and West Germany before and after the reunification*, I therefore hypothesize that family formation will be equally different between East and West German women before and after the reunification (*hypothesis 3*).

Data and Methods

Figure 6.2 illustrates the historical comparative cohort design used to analyze the difference in family formation within and between the German sub-societies. The x-axis shows historical time. The grey bars locate the observed family formation trajectories of the study cohorts in historical time: family formation trajectories between age 15 and 33 of women born in the early 1950s compared to the trajectories of East and West German women born in 1971. The cohorts born in the 1950s experienced family formation in divided German sub-societies. Women born in 1971 were just beginning their active family formation in a reunified Germany in 1989.⁸

The data come from the German Life History Study (GLHS) (Mayer 2008). The study uses retrospective life history data for women born 1951–1953 in East Germany collected in 1991/1992, and for women born 1954–1956 in West Germany collected in 1988/1989. The data for the cohorts born in 1971 was collected between

⁷ Note that *hypotheses 1* and *2* imply stable standardization of family formation across all of Germany as a compositional outcome of a more sizeable de-standardization in the smaller East German population and moderate re-standardization in the larger West German population.

⁸ The East German sample of the cohort born 1971 includes women who were born in the GDR and living in East Germany in 1990.

1996 and 1999. They were followed up again with a panel in 2005. I only include women who participated in the panel follow-up. When information was inconsistent in the basic survey and the follow-up, information from the basic surveys is used because it is less prone to recall error.⁹

The analysis sample consists of 485 women born in the 1950s in West Germany, 287 women born in the 1950s in East Germany, and 474 women born in 1971, of which 132 were born in East Germany and 342 were born in West Germany. The family formation sequences are cut at an equal length of 207 months, which corresponds to 17 years between age 15 and 33 and can be observed for all cohorts. The alphabet of states is specified as: S='single'; CNC='cohabiting, no child'; CC='cohabiting, with child', MNC='married, no child', MC='married, with child'; and DW='divorced/widowed'.¹⁰ 'Single' refers to not married and not co-habiting, since the data does not contain information on living-apart-together (LAT) relationships for all cohorts.

This study uses Lesnard's dynamic Hamming distance (2008, 2010) to emphasize the timing and pacing of family formation in sequence similarity. This method employs only substitution operations, no insertion/deletion operations, and specifies time point-specific substitution costs. The substitution of two family formation states is 'cheaper', and thus generates less distance, when transitions between two family formation states are frequent. Substantively, this means that two family formation trajectories will be identified as similar when they go through the same family formation states at the same pace. They will not be regarded as similar if they go through the same family formation states at different speeds. Given that the sequences are censored at age 33, variation in the timing of family formation is of primary interest in this study. The resulting pairwise distances are indicators of de-standardization. Just as the concept of de-standardization (Brückner and Mayer 2005), pairwise sequence distances capture a relational property—the similarity between trajectories—and not a characteristic of individual sequences.

Within and Between Group Comparisons Using Sequence Distances

Figure 6.3 shows a schematic example of two sequence distance matrices to illustrate how they can be used for within-group and between-group comparisons. The distance matrix of the cohorts born in the 1950s is displayed on the left and of the cohort born 1971 on the right. To examine changes in the de-standardization of family formation, I propose to compare different areas of these distance matrices that indicate the de-standardization outcomes under the constellations of conditions represented by the respective German sub-societies. I thereby take into account the

⁹ There were few deviations and in most cases there was only a few months difference in the timing of a change in partnership status.

¹⁰ Divorced and widowed are combined, because widowhood occurs very rarely in this relatively young sample.

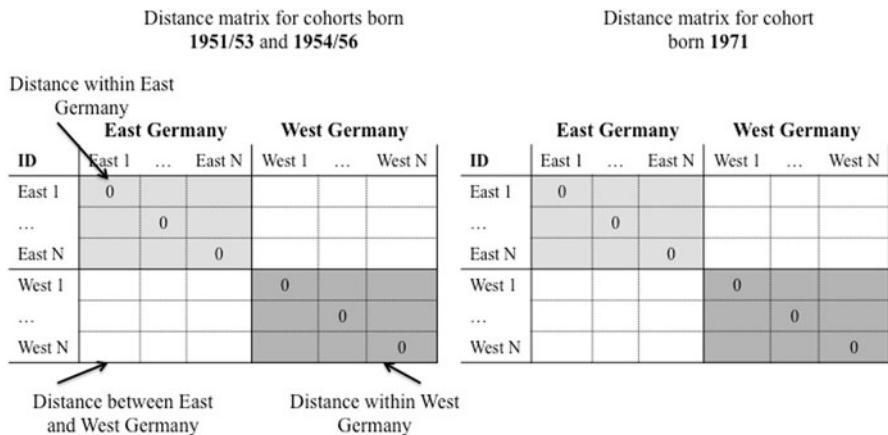


Fig. 6.3 Example sequence distance matrix to illustrate within- and between-group sequence distances

processual nature of family formation, as well as the preeminence of multiple-way interactions emphasized both in narrative positivism and the life course paradigm.

The upper left areas of the matrices, shaded in light grey, compare all East German women to each other. The mean pairwise distance in these areas are an indicator of the degree of de-standardization in East Germany (*hypothesis 1*). Comparing the dark grey areas of the two matrices in the lower right quadrant shows the de-standardization of family formation in West Germany (*hypothesis 2*). Comparing the white areas in the off-diagonals of the two matrices shows the between-group comparison. In the white areas, every East German woman is compared to every West German woman, indicating the difference between them before and after the reunification (*hypothesis 3*). For example, if mean distances of the white areas are smaller for women born in 1971 than for women born in the 1950s, this would indicate a convergence of family formation for East and West Germany since the reunification. Comparing these different sections of the distance matrix allows me to comply to both the “relational” and “dynamic” nature of de-standardization.

To explore whether subgroup differences in mean sequence distances are statistically meaningful, I calculate bootstrap confidence intervals (Efron and Tibshirani 1993). Since sequence distances are calculated between each possible pair of sequences, pairs of sequences become the unit of analysis. The number of possible sequence pairs is given by $N(N-1)/2$. Bootstrapping of sequence distances is different for simple random variables, because the independently observed resampling units (the family formation sequences) are not the values the statistic is calculated from (the pairwise distances). To calculate bootstrap confidence intervals, I draw 1,000 random samples from the original sequences with replacement and calculate the respective mean pairwise sequence distance for each sample (see

Table 6.1 Dynamic Hamming distances, normalized between 0 and 1

	Before the reunification Cohort 1951–1953/1954–1956	After the reunification Cohort 1971	Difference (percent)
East Germany	0.404 (0.375–0.439)	0.573 (0.556–0.597)	0.169 (41.8)
West Germany	0.546 (0.532–0.562)	0.500 (0.488–0.516)	−0.046 (9.2)
Germany	0.514 (0.499–0.527)	0.527 (0.515–0.541)	0.013 (2.5)
East-West difference	0.520 (0.513–0.540)	0.554 (0.512–0.595)	0.034 (6.5)

95 % bias-corrected and accelerated bootstrap confidence intervals in parentheses

Fasang 2012).¹¹ All analyses were conducted with the TraMineR package for sequence analysis in R (Gabadinho et al. 2011).

Results

Table 6.1 shows mean sequence distances as indicators of the degree of standardization of family formation before and after the reunification. They are normalized to vary between 0 and 1 to facilitate comparisons across cohorts. A distance of zero refers to identical trajectories, whereas one indicates the maximum distance. The far right column shows the percentage difference in the means for the cohorts born in the 1950s and in 1971.

In support of *hypothesis 1*, Table 6.1 shows a massive de-standardization of family formation in East Germany after the breakdown of the regulative communist regime. Family formation is more de-standardized by 0.169 which corresponds to a 41.8 % ($0.573/0.404 - 1$) increase compared to communist East Germany. Secular values were established during communism. Yet, they only became apparent in a de-standardization of family formation once the regulative pro-natalist regime was removed and they coincided with the economic crisis of transition. An important part of the East German story is high standardization during communism against which de-standardization after the reunification appears particularly drastic.

In line with *hypothesis 2*, Table 6.1 shows a moderate but statistically significant 9.2 % increase in the standardization of family formation for West German women. This substantiates previous research that a new re-standardized pattern of family formation solidifies in societies that were among the first to experience the SDT (Brüderl 2004; Elzinga and Liefbroer 2007; Lesnard et al. 2010). For West Germany, this re-standardized pattern is a polarization between either traditional marriage and motherhood or delay.

¹¹ For Lesnard's dynamic Hamming distance, substitution costs depend on time point specific transition rates between family formation states, which can vary across bootstrap samples. As a result, the absolute transition costs can vary across bootstrap samples, but the principle of deriving them always remains the same.

As expected, this cross-over of de-standardization for East and West German women averages out in a stable standardization of family formation across all of Germany: the composite mean masks diverging trends across sub-societies. Additional cluster analyses that explore the substantive family formation patterns underlying these de-standardization trends (available from author) further substantiate a shift away from a traditional early marriage, early fertility pattern to cohabitation and motherhood out of wedlock for East German women. For West Germany, the results add further support to a polarized pattern of either delay or traditional family formation.

In line with *hypothesis 3* and Goldstein and Kreyenfeld's (2011) argument of persistent differences between the German sub-societies, East and West German women's family formation is just as different in the two decades following the reunification as it was in divided Germany. Mean between-group distances between East and West German women born in 1971 is even slightly higher at 0.554 than for women born in the 1950s at 0.520 but the bootstrap confidence intervals overlap (0.512–0.595 and 0.513–0.540). This forcefully illustrates the added value of complementing average indicators of isolated family events, such as total fertility rates (TFRs), with a more comprehensive perspective on the larger process they are embedded in—a focus on the processual character of social reality that is central to both narrative positivism and the life course paradigm. On first sight it may seem politically obsolete or even unjustified, to regard East and West Germany as separate sub-societies more than 20 years after the reunification. The empirical reality of such persistent differences resulting from complex constellations of macro-structural conditions suggests otherwise—particularly given cross-over effects that are masked in overall averages.

Discussion

This chapter set out to scrutinize the potential of sequence analysis to inform core theoretical debates on family formation from a life course perspective. First, the premises of narrative positivism and the life course paradigm were revisited. I highlighted several overlaps in the two approaches, focusing on the preeminence of temporality and process and the assumption that social reality happens in interactions rather than main effects.

Based on these premises, I argued that sequence analysis is most promising to tighten the link between theory and empirical research with regard to three theoretical areas in the field of family formation research: (a) *multidimensional lives*, (b) *linked lives* within networks of shared relationships, and (c) the *de-standardization* and *pluralization* of family formation. A review of the literature showed that sequence analysis applications in the study of multidimensional lives and linked lives are a nascent but growing field. To date, most studies using sequence analysis address the de-standardization and pluralization of family formation. They largely follow an inductive, exploratory approach for thick description. This resonates with

Billari's (2001a) notion of a 'pragmatic view' on life courses that is informed by comparative thick description of collective life course patterns.

I conclude that in addition to its well-demonstrated exploratory value, sequence analysis has a lot of untapped potential for generating descriptive evidence targeted at informing theoretical narratives. To this end, it is conducive to incorporate the premises of narrative positivism and the life course paradigm more broadly in theory and research designs using sequence analysis. While incorporating all of them at once might not be feasible in view of Abbott's long list of critiques of general linear reality, even addressing select ones that are assumed most important for the respective research question at hand can advance sociological insight.

To demonstrate this argument, I presented a case study on the de-standardization of family formation before and after the German reunification. This case study incorporated the preeminence of processes and multiple-way interactions by theorizing explicitly about the impact of constellations of macro-structural contexts on the de-standardization of family formation. I proposed a new way to measure within-group and between-group distances in sequences to inform hypotheses about the de-standardization of family formation in the German sub-societies. The results suggest that theorizing about multiple-way interactions between different contextual factors is a promising approach to understand life course patterns. A crucial ensuing question is to what extent it is legitimate and fruitful to simplify complex contextual factors necessary to arrive at manageable combinations of conditions without overly distorting the complexity of social reality. To generate insightful new perspectives on family formation with sequence analysis, researchers should channel more energy into theory-building and research designs based on the overlapping premises of narrative positivism and the life course paradigm.

Acknowledgements I thank Felix Bühlmann, Marcel Raab, and Frans Willekens for detailed and thoughtful comments on earlier drafts of the manuscript. This chapter benefited from previous collaboration on sequence analysis and countless fruitful discussions with Silke Aisenbrey and Tim F Liao. The arguments and empirical study in this chapter were greatly influenced by the guidance of two exceptional mentors: Hannah Brückner and Karl Ulrich Mayer. I gratefully acknowledge support from the Max-Planck Institute for Demographic Research that provided a wonderful academic environment for a research visit, during which most of this chapter was written. The usual disclaimer applies.

References

- Aassve, A., Billari, F. C., & Piccarreta, R. (2007). Strings of adulthood: A sequence analysis of young British Women's work-family trajectories. *European Journal of Population/ Revue européenne de Démographie*, 23(3–4), 369–388. doi:10.1007/s10680-007-9134-6.
- Abbott, A. (1992). From causes to events: Notes on narrative positivism. *Sociological Methods & Research*, 20(4), 428–455. doi:10.1177/0049124192020004002.
- Abbott, A. (1988). Transcending general linear reality. *Sociological Theory*, 6(2), 169–186.
- Abbott, A. (2001). *Time matters. On theory and method*. Chicago: University of Chicago Press.
- Abell, P. (1993). Some aspects of narrative method. *The Journal of Mathematical Sociology*, 18(2–3), 93–124.

- Abell, P. (2004). Narrative explanation: An alternative to variable-centered explanation? *Annual Review of Sociology*, 30(1), 287–310. doi:10.1146/annurev.soc.29.010202.100113.
- Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The “second wave” of sequence analysis. Bringing the “course” back into the life course. *Sociological Methods & Research*, 38(3), 420–462. doi:10.1177/0049124109357532.
- Amato, P. R., & DeBoer, D. D. (2001). The transmission of marital instability across generations: Relationship skills or commitment to marriage? *Journal of Marriage and Family*, 63(4), 1038–1051.
- Beck, U. (1992). *Risk society: Towards a new modernity*. London: Sage Publications.
- Bell, D. (1973). *The coming of post-industrial society. A venture in social forecasting*. New York: Basic Books.
- Bengtson, V. L. (2001). Beyond the nuclear family: The increasing importance of multigenerational bonds. *Journal of Marriage and Family*, 63(1), 1–16. doi:10.1111/j.1741-3737.2001.00001.x.
- Biemann, T., Fasang, A. E., & Grunow, D. (2011). Do economic globalization and industry growth destabilize careers? An analysis of career complexity and career patterns over time. *Organization Studies*, 32(12), 1639–1663. doi:10.1177/0170840611421246.
- Billari, F. C. (2001a). Sequence analysis in demographic research. *Canadian Studies in Population*, 28(2), 439–458.
- Billari, F. C. (2001b). The analysis of early Life courses: Complex descriptions of the transition to adulthood. *Journal Of Population Research*, 18(2), 119–142.
- Billari, F. C., & Liefbroer, A. C. (2010). Towards a new pattern of transition to adulthood? *Advances in Life Course Research*, 15(2–3), 59–75. doi:10.1016/j.alcr.2010.10.003.
- Billingsley, S. (2010). The post-communist fertility puzzle. *Population research and policy review*, 29(2), 193–231. doi:10.1007/s11113-009-9136-7.
- Bonetti, M., Piccarreta, R., & Salford, G. (2013). Parametric and nonparametric analysis of life courses: An application to family formation patterns. *Demography*, 50(3), 881–902. doi:10.1007/s13524-012-0191-z.
- Bras, H., Liefbroer, A. C., & Elzinga, C. H. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography*, 47(4), 1013–1034. doi:10.1007/BF03213737.
- Brückner, H. (2004). *Gender inequality in the life course: Social change and stability in West Germany 1975–1995*. New York: Walter de Gruyter.
- Brückner, H. (2007). *Temporal and contextual dynamics in adolescent risk behavior*. Paper presented at the ASA methodology section spring meeting.
- Brückner, H., & Mayer, K. U. (2005). The de-standardization of the life course: What it might mean and if it means anything whether it actually took place. In R. Macmillan (Ed.), *The structure of the life course. Standardized? Individualized? Differentiated?* (pp. 27–54). Amsterdam: Elsevier.
- Brüderl, J. (2004). Die Pluralisierung partnerschaftlicher Lebensformen in Westdeutschland und Europa. *Aus Politik und Zeitgeschichte*, 19, 3–10.
- Brzinsky-Fay, C., & Kohler, U. (Eds.). (2010). New developments in sequence analysis. *Sociological Methods and Research*, 38(3), 359–364.
- Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis using stata. *The Stata Journal*, 6(4), 435–460.
- Buchmann, M. C., & Kriesi, I. (2011). Transition to adulthood in Europe. *Annual Review of Sociology*, 37(1), 481–503. doi:10.1146/annurev-soc-081309-150212.
- Card, N. A., Selig, J. P., & Little, T. D. (Eds.). (2008). *Modeling dyadic and interdependent data in the developmental and behavioral sciences*. New York: Routledge.
- Coleman, D. (2004). Why we don’t have to believe without doubting in the “Second Demographic Transition”—some agnostic comments. *Vienna Yearbook of Population Research*, 2, 11–24.
- Cook, T. D., & Furstenberg, F. F. (2002). Explaining aspects of the transition to adulthood in Italy, Sweden, Germany, and the United States: A cross-disciplinary case synthesis approach. *Annals of the American Academy of Political and Social Sciences*, 580(1), 257–287.

- Diewald, M., Goedicke, A., & Mayer, K. U. (2006). *After the fall of the wall. Life courses in the transformation of East Germany*. Stanford: Stanford University Press.
- Dijkstra, W., & Taris, T. (1995). Measuring the agreement between sequences. *Sociological Methods & Research*, 24(2), 214–231. doi:10.1177/0049124195024002004.
- Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.
- Elder, G. H. (1994). Time, human agency, and social change: Perspectives on the life course. *Social Psychology Quarterly*, 57(1), 4–15.
- Elder, G. H., Johnson, M. K., & Crosnoe, R. (2003). The Emergence and Development of Life Course Theory. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the life course* (pp. 3–19). New York: Kluwer Academic/Plenum Publishers.
- Elzinga, C. H. (2003). Sequence similarity: A nonaligning technique. *Sociological Methods & Research*, 32(1), 3–29. doi:10.1177/0049124103253373.
- Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population/Revue européenne de Démographie*, 23(3–4), 225–250. doi:10.1007/s10680-007-9133-7.
- Esping-Andersen, G. (1990). *The three worlds of welfare capitalism*. Princeton: Princeton University Press.
- Fasang, A. E. (2012). Retirement patterns and income inequality. *Social Forces*, 90(3), 685–711. doi:10.1093/sf/sor015.
- Fasang, A. E., & Raab, M. (2013). Beyond intergenerational transmission: Intergenerational patterns of family formation. Accepted at Demography New Orleans. <http://paa2013.princeton.edu/papers/132178>. Accessed 24 Sept 2013.
- Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Gauthier, J.-A., Widmer, E. D., Bucher, P., & Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology*, 40(1), 1–38. <http://doi.wiley.com/10.1111/j.1467-9531.2010.01227.x>.
- Goldstein, J. R., & Kreyenfeld, M. (2011). Has East Germany overtaken West Germany? Recent trends in order-specific fertility. *Population and development review*, 37(3), 453–472.
- Han, S.-K., & Moen, P. (1999). Clocking out: Temporal patterning of retirement. *American Journal of Sociology*, 105(1), 191–236.
- Hedström, P., & Bearman, P. (Eds.). (2009). *The oxford handbook of analytical sociology*. Oxford: Oxford University Press.
- Huinink, J. (2011). New patterns or no patterns? Changing family development and family life in Europe. *Studi interdisciplinari sulla famiglia*, 25, 49–70.
- Huinink, J., Mayer, K. U., Solga, H., Sørensen, A., & Trappe, H. (1995). *Kollektiv und Eigensinn. Lebensverläufe in der DDR und danach*. Berlin: Akademie Verlag.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: The Guilford Press.
- Kohler, H.-P., Billari, F. C., & Ortega, J. A. (2002). The emergence of lowest-low fertility in Europe during the 1990s. *Population and Development Review*, 28(4), 641–680.
- Kohli, M. (1985). Die Institutionalisierung des Lebenslaufs. Historische Befunde und theoretische Argumente. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 37(1), 1–29.
- Lesnard, L. (2008). Off-scheduling within dual-earner couples: An unequal and negative externality for family time 1. *American Journal of Sociology*, 114(2), 447–490.
- Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3), 389–419. doi:10.1177/0049124110362526.
- Lesnard, L., Cousteaux, A.-S., Chanvril, F., & Le Hay, V. (2010). Do transitions to adulthood converge in Europe? An optimal matching analysis of work-family trajectories of young adults from 20 European countries. SciencesPo. Working paper 2010–2004.
- Lesthaeghe, R. (2010). The unfolding story of the second demographic transition. *Population and development review*, 36(2), 211–251.
- Liefbroer, A. C., & Elzinga, C. H. (2012). Intergenerational transmission of behavioural patterns: How similar are parents' and children's demographic trajectories? *Advances in Life Course Research*, 17(1), 1–10. doi:10.1016/j.alcr.2012.01.002.

- Lyngstad, T. H., & Prskawetz, A. (2010). Do siblings' fertility decisions influence each other? *Demography*, 47(4), 923–934.
- Mayer, K. U. (2004). Whose lives? How history, societies, and institutions define and shape life courses. *Research in Human Development*, 1(3), 161–187.
- Mayer, K. U. (2008). Retrospective longitudinal research: The German life history study. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement and analysis* (pp. 85–106). San Diego: Elsevier.
- Mayer, K. U. (2009). New directions in life course research. *Annual Review of Sociology*, 35, 413–433.
- McLanahan, S., & Percheski, C. (2008). Family structure and the reproduction of inequalities. *Annual Review of Sociology*, 34(1), 257–276. doi:10.1146/annurev.soc.34.040507.134549.
- Moen, P., & Roehling, P. V. (2005). *Career mystique: Cracks in the American dream*. Oxford: Rowman & Littlefield.
- Moen, P., & Sweet, S. (2004). From 'work-family' to flexible careers'. A life course reframing. *Community, Work & Family*, 7(2), 209–226.
- Müller, N. S., Sapin, M., Gauthier, J.-A., Orita, A., & Widmer, E. D. (2012). Pluralized life courses? An exploration of the life trajectories of individuals with psychiatric disorders. *The International journal of social psychiatry*, 58(3), 266–277. doi:10.1177/0020764010393630.
- Pasteels, I., & Mortelmans, D. (2013). *A dyadic analysis of repartnering after divorce. Do children matter?* Paper presented at the workshop on "life-course transitions after separation", Berlin, July 4 and 5, 2013.
- Pollock, G. (2007). Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1), 167–183. doi:10.1111/j.1467-985X.2006.00450.x.
- Prince Cooke, L. (2004). The gendered division of labor and family outcomes in Germany. *Journal of Marriage and Family*, 66(5), 1246–1259.
- Raab, M., Fasang, A. E., Karhula, A., & Erola, J. (2013). Similarity of siblings' family formation. Paper prepared for the 2013 XVII IUSSP International Population Conference Busan, Republic of Korea. http://www.iussp.org/sites/default/files/event_call_for_papers/2013_IUSSP_extended_abstract_Erola_et.al_0.pdf. Accessed 24 Sept 2013.
- Rihoux, B., & Ragin, C. C. (Eds.). (2009). *Configurational comparative methods. Qualitative comparative analysis (QCA) and related techniques*. Thousand Oaks: Sage Publications.
- Robette, N. (2010). The diversity of pathways to adulthood in France: Evidence from a holistic approach. *Advances in Life Course Research*, 15(2–3), 89–96. doi:10.1016/j.alcr.2010.04.002.
- Robette, N., Bry, X., & Leliévre, È. (2013). Like mother, like daughter? A dyadic sequence analysis approach to uncover patterns of mothers and daughters careers. http://nicolas.robette.free.fr/Docs/Mother_daughter_RobetteBryLelievre.pdf. Accessed 24 Sept 2013.
- Savage, M. (2009). Contemporary sociology and the challenge of descriptive assemblage. *European Journal of Social Theory*, 12(1), 155–174. doi:10.1177/1368431008099650.
- Sennett, R. (1998). *Corrosion of character. The personal consequences of work in the new capitalism*. New York: W.W. Norton.
- Settersten, R. A., & Mayer, K. U. (1997). The measurement of age, age structuring, and the life course. *Annual Review of Sociology*, 23, 233–261.
- Sobotka, T. (2003). Re-emerging diversity: Rapid fertility changes in Central Europe collapse of the communist regimes. *Population*, 58(4), 451–485.
- Stovel, K., Savage, M., & Bearman, P. (1996). Ascription into achievement: Models of career systems at Lloyds bank, 1890–1970. *American Journal of Sociology*, 102(2), 358–399.
- Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, 40, 471. doi:10.1177/0049124111415372.
- Widmer, E. D. (2010). *Family configurations. A structural approach to family diversity*. London: Ashgate Publishing.
- Widmer, E. D., & Ritschard, G. (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life Course Research*, 14, 28–39. doi:10.1016/j.alcr.2009.04.001.

Chapter 7

Developmental Psychologists' Perspective on Pathways Through School and Beyond

Julia Dietrich, Håkan Andersson and Katariina Salmela-Aro

Educational and career transitions represent major developmental milestones on young people's pathways and related developmental sequences to adulthood. Given their influence on outcomes over the course of adult life, the career choices that young people make are critical for their future development in the career domain and beyond (e.g., Abele and Spurk 2009). In the psychological literature, making career decisions and going through related transitions is considered a key developmental task for young people (Havighurst 1948). In the past years, the psychological transition to adulthood has become more and more heterogeneous and uncertain (Cohen et al. 2003). Heterogeneity of transitional pathways is especially common in the career domain (e.g., Brzinsky-Fay 2007; Fouad and Bynner 2008). For example, there are manifold pathways young people take after school when they enter college, work, apprenticeship, or other types of tertiary and vocational education with or without going through a moratorium activity such as a gap year (Wells and Lynch 2012). In the wake of growing heterogeneity and uncertainty in the world of work, there has been a rising interest in identifying patterns of people's career development (e.g., Biemann et al. 2012; Bühlmann 2008; Stovel et al. 1996). The present study applies the idea of career patterns to the educational transitions of young Finns during school and beyond. Our study presents an application of sequence analysis in the field of developmental psychology linking two types of longitudinal data: the sequences of young people's educational and employment activities and the longitudinal data of psychological resources. We followed participants from ages 16 to 20 as they went through two key educational transitions: first, the transi-

J. Dietrich (✉)

Institute of Educational Science, Department of Educational Psychology,
University of Jena, Am Planetarium 4, 07743 Jena, Germany
e-mail: julia.dietrich@uni-jena.de

H. Andersson

Department of Psychology, Stockholm University, Stockholm, Sweden

K. Salmela-Aro

Department of Psychology, University of Jyväskylä, Helsinki, Finland

Helsinki Collegium for Advanced Studies, University of Helsinki, Helsinki, Finland

tion from comprehensive school to upper secondary education, and, second, the transition from upper secondary education to employment or to further studies. The aims of the study were threefold: First, what are typical pathways during educational transitions in Finland? Second, to what extent do psychological resources predict taking a certain educational trajectory? Third, do psychological resources show differential change among young people on different educational pathways?

The Role of Psychological Resources During Educational Transitions

During transition periods, such as from one school form to another or from education to employment, individuals can engage in a variety of intentional behaviors which are assumed to benefit their success in dealing with the demands of the transition (e.g., Heckhausen et al. 2010; Salmela-Aro 2009; Savickas 2011). Such behaviors include, for example, exploring occupational options, forming occupational goals, and investing effort in the pursuit of these goals. A number of theories in the fields of developmental and vocational psychology aim to understand the role that young people's intentional engagement plays for education and career transitions (for a review, see Dietrich et al. 2012). The research lines of career development (e.g., Savickas 2011; Vondracek and Porfeli 2008) and developmental regulation (e.g. Heckhausen et al. 2010; Nurmi 1992; Salmela-Aro 2009) have accumulated significant knowledge on the range of relevant behaviors, their development, and their predictive power for related transition success. In summary, these theories suggest that it is adaptive for a person to set occupational goals and form an occupational identity, to actively pursue their goals and actualize their identity, and to reconstruct goals and identity commitments if necessary. These kinds of adaptive goal- and identity-related behaviors have been termed as phase-adequate engagement (Dietrich et al. 2012). In this chapter, we approach phase-adequate engagement from the perspective of developmental regulation theories. That is, we focus on the role that adolescents' career-related goals play in their educational trajectories as identified using sequence analysis of educational and employment activities.

Generally, personal goals are states that people strive to achieve (Austin and Vancouver 1996). According to developmental regulation theories, age-graded normative transitions, such as the transition after high school, guide young people in selecting broad developmental goals. These, in turn, give rise to a set of subordinate goals that structure young people's everyday lives and help them to channel their resources into current developmental tasks (Heckhausen et al. 2010; Little 1983; Salmela-Aro 2009). Research on personal goals is not only concerned with the contents of young people's goals (e.g., "going to college" or "becoming a nurse"), but also focuses on the ways in which they think about their goals and their engagement towards goal pursuit. This process has been termed goal appraisal (Austin and Vancouver 1996). In research practice, goal appraisal is typically examined by asking people to write down the personal goals they currently pursue, followed by

questions about how they think and feel about them. This idiosyncratic approach to goals and related appraisals was also adopted in the current study in the context of career-related goals. Young people reported one current career-related goal, which had a focus on their education or occupational planning. Our interest for this chapter was then on how important that career-related goal was to young people, their beliefs on the attainability of the career goal, the effort they invested in striving for the goal, their perceptions of the progress they made toward accomplishing it, and the amount of stress they experienced during goal pursuit (see Nurmi et al. 2009, for typical goal-appraisal dimensions).

The phase-adequate engagement, and, particularly, developmental regulation models (see Dietrich et al. 2012) contend that goal appraisal has consequences for life outcomes, such as people's transitional pathways. Giving high importance to one's goals and setting goals that are attainable can be assumed to influence transitional pathways directly and indirectly. Having important and attainable career goals may directly influence one's choice of educational track (Eccles 2009) but may also influence one's educational pathway through the amount of effort one invests into goal pursuit and, in turn, the levels of goal progress (Lent et al. 1994). For example, people are likely to put more effort into something they value, eventually leading to success in the task at hand (Sheldon 2002). Evidence suggests that people who place importance on career-related goals and engage actively in their pursuit more easily find a job during the transition from education to employment (e.g., Nurmi et al. 2002). Similarly, high perceived progress with career goals has been shown to be associated with a successful transition to university (Vasalampi et al. 2012; Vasalampi et al. 2009). While importance, attainability, effort, and progress appraisals facilitate attainment, strain and stress with one's goals where demands exceed a person's resources, is an adverse experience. Goal-related stress can thus be assumed to pull people away from those high demands (e.g., the demands of academic track education and university) or education in general (Salmela-Aro and Suikkari 2008).

Personal goals are also sensitive to transitions (Nurmi 1992; Salmela-Aro 2009). When a goal is attained, people typically capitalize on that success and set new goals (Heckhausen et al. 2010). However, changing demands or going through specific phases of a transition can make prior goals inefficient and futile (Salmela-Aro and Suikkari 2008). Young people thus frequently reconstruct their career-related goals in times of transition (Dietrich et al. 2012; Nurmi and Salmela-Aro 2002). Also, the appraisals of such goals can be expected to change. However, very few studies have examined changes in goal appraisal during educational and career transitions (see also Vasalampi et al. 2010). In one study, Dietrich et al. (2012) found that the importance of work-related goals was very high before young people made a transition to employment and decreased thereafter when other life tasks such as partnership and family gained importance. However, change in a young person's goal appraisal likely depends on the pathway that the person takes through the education system. This hypothesis is explored in the current chapter. Before more detailed hypotheses are laid out, we introduce the context in which this study was conducted, the educational system of Finland.

The Finnish Education System

At the age of 16, after nine years of comprehensive schooling, most young people in Finland (around 90 %) make a transition to upper secondary education. Very few adolescents make a transition to employment or voluntarily attend a tenth school year in comprehensive school before they are ready for the transition to upper secondary school. Upper secondary education has two tracks: academic and vocational schools, in which young people are educated for 3 or 4 years. The academic track typically leads to university, while vocational school usually serves as a direct route to employment. Many students also apply to polytechnics, which offer higher education with a practical orientation. In Finland, education at all levels is state-provided and tuition-free.

While during the transition from comprehensive school there are two major choice options (vocational and academic secondary education), there are a multitude of activities for young people after graduation from upper secondary school. These include university, employment, vocational education, and moratorium activities, such as gap years. Typical of the Finnish educational system is the interrupted educational pathway after upper secondary school before continuing to higher education. Partly due to the tough entrance examinations of universities and polytechnics, more than 60 % of the students completing the matriculation examination in upper secondary education pass through at least one gap year before or during their current studies (Statistics Finland 2010). Moreover, in Finland, military or civil service is mandatory for males and usually completed after secondary education.

Research Questions and Hypotheses

Using data from a longitudinal study of Finnish young people (FinEdu study), this chapter seeks to link young people's educational pathways to their career-related personal goals. Students were followed from age 16 (end of comprehensive school) to age 20, when most of them had completed either upper secondary (academic track) or vocational school. Sequence analysis, an optimal matching procedure which empirically identifies typical developmental sequences in the data, is an ideal tool for studying young people's career pathways (Abbott and Hrycak 1990; Aisenbrey and Fasang 2010; see also the chapters in this volume). It has been used to examine the role of socio-economic predictors such as gender, age, education, and family-related variables, in career transitions (e.g., Biemann et al. 2012; Bühlmann 2008; McVicar and Anyadike-Danes 2002; Stovel et al. 1996). Psychologists have only started to recognize the potential of analyzing sequences (e.g., Huang et al. 2007; Salmela-Aro et al. 2011), and, therefore, research focusing on transition patterns through the education system and to employment has been limited to socioeconomic predictors. Little is known about how psychological factors such as career-related goal appraisals contribute to individuals' educational transitions

patterns. Furthermore, there is a scarcity of knowledge regarding developmental changes in goal appraisal and to what extent these differ between different educational transition pathways (i.e., sequences). With our study, we seek to answer the following research questions:

1. What are typical pathways (i.e., patterns or sequences) during educational transitions as Finnish youth go through two key career transitions: first, the transition from comprehensive school to upper secondary education, and, second, the transition from upper secondary education to further studies, employment, or other post-school activities? We will analyze sequential data on young people's educational/employment activities in each 6-month period from the beginning of secondary education (around age 16–17) to the age of 20. We expect to find vocationally-oriented and academically-oriented pathways. Students graduating from the academic track face more possible post-school options than vocational students; thus, pathways from the academic track might be more diverse. Moreover, research on career patterns and transitions typically finds 'smooth' or 'express' pathways, pathways characterized by breaks or gaps, and pathways characterized by disengagement from education and career (e.g., Brzinsky-Fay 2007; Huang et al. 2007; McVicar and Anyadike-Danes 2002). In other words, it can be hypothesized that, while some people make smooth and immediate transitions through the education system, others experience delays and instabilities—for example, at the entrance to university.
2. Does career-related goal appraisal predict educational transition pathways? We hypothesized that placing high importance on career goals, setting attainable career goals, putting effort in these goals, and perceiving them to progress well are psychological resources that increase the likelihood that a person experiences smooth transitions and minimal delays (e.g., Dietrich et al. 2012b). A lack of these resources or high levels of career goal-related stress can be assumed to increase the likelihood of disengagement from education. When examining the role of career goal appraisal for educational transition pathways, we control for a number of covariates known to be typically related to transition pathways. These are gender, socioeconomic status, and academic achievement.
3. Does career-related goal appraisal show differential changes for young people on different educational pathways? We expected that changes in goal appraisal differed between vocational and academic pathways and between smooth pathways vs. delayed and unstable pathways. Career goals were assessed at ages 16 years (comprehensive school), 18 years (upper secondary school), and 20 years (post-school) and could change at two transition points: during the transition from comprehensive school to upper secondary school and during the transition from upper secondary school to post-school activities. Goal importance can be assumed to increase in salience and importance when the educational context becomes more occupation-focused (Dietrich et al. 2012b). This happens across the first transition for vocational students, while, for youths in the academic track, this shift should occur across the second transition. Likewise, career goal attainability can be expected to increase at the first transition point for young

people on vocational pathways and to increase at the second transition point for academic pathways. However, because motivation is often boosted when a person experiences success (Bandura 1997), goal attainability should rise only on smooth but not on struggling pathways. Following the same logic, goal effort and progress can be hypothesized to change in the same manner as goal attainability. Moreover, drawing on research of school burnout (Salmela-Aro and Tynkkynen 2012), it can be expected that career goal-related stress increases on academic pathways and decreases on vocational pathways due to changes in the academic demands associated with both kinds of pathways. Finally, career goal importance, attainability, effort, and progress can be assumed to be low and even decreasing in a disengagement from the education pathway, while goal stress could be initially high but decreasing as the disengagement continues.

Method

Participants and Procedure

The present study is part of the ongoing Finnish Educational Transitions (FinEdu) study. At the beginning of the study, in spring 2004, the participants were 9th-graders (median age=16) facing the transition to upper secondary education. We recruited all the 9th-grade students ($N=954$) in a medium-sized town (population 88,000) in Central Finland for the study. Five measures have been carried out: one before the transition to upper secondary school and four after the transition to upper secondary education at ages 17, 18, 20, and 23. At the first two measurement points, the questionnaires were administered to the students in their classrooms during regular school hours, and, at the last three measurement points, the questionnaires were sent to their respective postal addresses. The present analyses are based on information collected at age 16 (Time 1), 18 (Time 2), and 20 (Time 3).

At Time 1,687 adolescents (327 females, 360 males) out of the 954 students attending the nine comprehensive schools participated in the study (participation rate 72%); at Time 2, the number was 749 (368 females, 381 males; participation rate 79%); and, at Time 3, the number was 611 (332 females, 279 males; participation rate 71%). The analysis sample comprises of 602 participants who filled in the life calendar at Time 3 (see below).

The majority of the participants (99%) were native Finnish-speaking, 1% of them having some other native language. This ratio corresponds well to the figures for ethnic minorities living in Central Finland. The attrition analyses showed that those who participated at each measurement point had higher grade point averages $t(833)=5.67, p < .001$ and that they were more likely to be female $\chi^2 (1)=27.77, p < .001$ compared to those who participated fewer times.

Measures

Life History Calendar

At age 20 (Time 3), participants completed the Life History Calendar (Caspi et al. 1996). They reported on their educational and work trajectory retrospectively for each 6-month period from Fall 2004 to Fall 2008, resulting in sequences of nine states. Participants indicated the six-month period in which a certain activity (e.g., studying in vocational track) began, how long it lasted, and when it ended (i.e., when they graduated or dropped out from that activity). The educational situation was coded using the following states: (1) voluntary 10th grade in comprehensive school, (2) academic upper secondary education, (3) vocational upper secondary education, (4) higher vocational school, (5) university, (6) polytechnic school, (7) open university, (8) full-time work, (9) army, (10) exchange year abroad, (11) no educational or employment activity.

To reduce the number of possible states, we combined similar categories. States (3) and (4) were recoded to vocational upper secondary education. States (5) and (6) were recoded to higher education. States (7), (9), (10), and (11) were recoded to career moratorium activities – that is, they were related neither to education nor employment. Open University is recoded as moratorium as it does not lead to a degree and does not include an entrance examination, but is available to all interested persons. The final states that built the alphabet for the sequence analysis were 10th grade, academic upper secondary education, vocational upper secondary education, higher education, work, and moratorium.

Career Goal Appraisal

At each time point, participants filled in the revised personal projects analysis inventory (Little 1983). First, they were asked to name one current personal goal related to their career, comprising educational and work goals. Second, participants were requested to rate this career-related personal goal on 9 appraisal items (1 = *very little*, 7 = *very much*), which reflected: (1) goal importance (2 items, “How important is the goal to you?” “How committed are you to this goal?”); (2) goal attainability (2 items, “How able do you think you are in attaining your goal?” “How probable would you say it is that this goal will come true?”); (3) goal effort (2 items, “How much time and effort have you spent on this goal?” “To what extent have you worked for your goal?”); (4) goal progress (1 item, “To what extent have you made progress in reaching this goal?”), and, (5) goal stress (2 items, “How stressful is this goal?” “How tiring or burdening is this goal?”). The Cronbach’s alphas for these scales at ages 16/18/20 were .75/.77/.76 for goal importance, .75/.73/.76 for goal attainability, .85/.87/.88 for goal effort, and .82/.86/.86 for goal stress.

Socioeconomic Status (SES)

Socioeconomic status (SES) was measured by parents' occupations. For the analyses of the study, a variable was created according to the highest SES of the parents. The variable was coded so that 1 indicated that either or both of the parents were in a blue-collar occupation, 2 indicated a lower white-collar occupation, and 3 a higher white-collar occupation, respectively. All other occupations were coded as missing values.

Academic Achievement (GPA)

Academic achievement was measured by asking the participants to report the grade point average (GPA) stated on their comprehensive school diploma.

Results

Transition Patterns

We applied sequence analysis to identify clusters of typical transition patterns. The analysis was conducted with the TraMineR package in R (Gabadinho et al. 2008). Figure 7.1 shows the state distributions for each time point.

To compute the distance matrix which was the basis for cluster analysis, we specified weights, substitution and indel (insertion and deletion) costs, reflecting the similarity between each pair of sequences (Abbott and Tsay 2000). Substitution costs were based on the transition rates between the states, and indel costs were set to 1. The optimal matching algorithm yielded a distance matrix between each pair of sequences. Based on this distance matrix, we ran a hierarchical cluster analysis using Ward's algorithm. Our choice of the final cluster solution was based on the proportion of explained variance, visual inspection of the dendrogram, and clarity of interpretation. To achieve a manageable, interpretable, and parsimonious cluster solution, we chose a solution with seven clusters. This solution explains 67% of the variance. Figure 7.2 depicts the distribution of sequences within each transition pattern and Table 7.1 shows descriptive statistics. We identified seven transition patterns:

1. **Vocational pathway:** Vocational pathways form the largest but also most heterogeneous cluster ($n=175$, discrepancy [degree of within cluster variation]=2.37). The most typical sequence corresponds to a vocational education for three years, followed by six months of moratorium and then the transition to employment. Also typical is a sequence of vocational education followed by 1.5 year of moratorium. This reflects young people who, after vocational education, have not yet

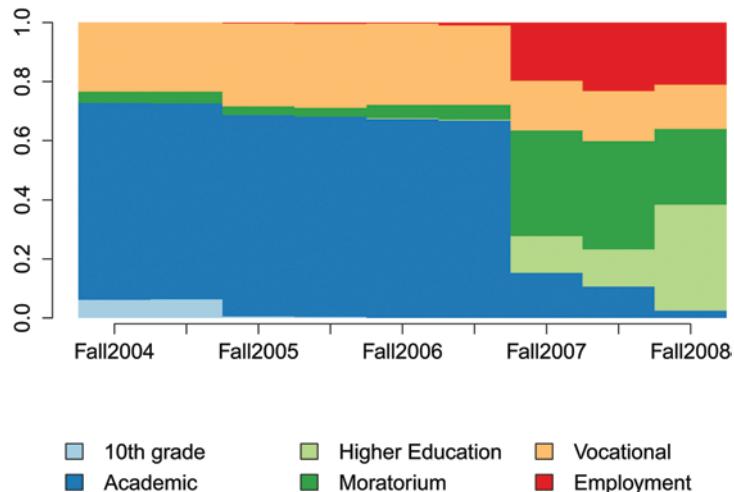


Fig. 7.1 Distributions of states from Fall 2004 (age 16) to Fall 2008 (age 20)

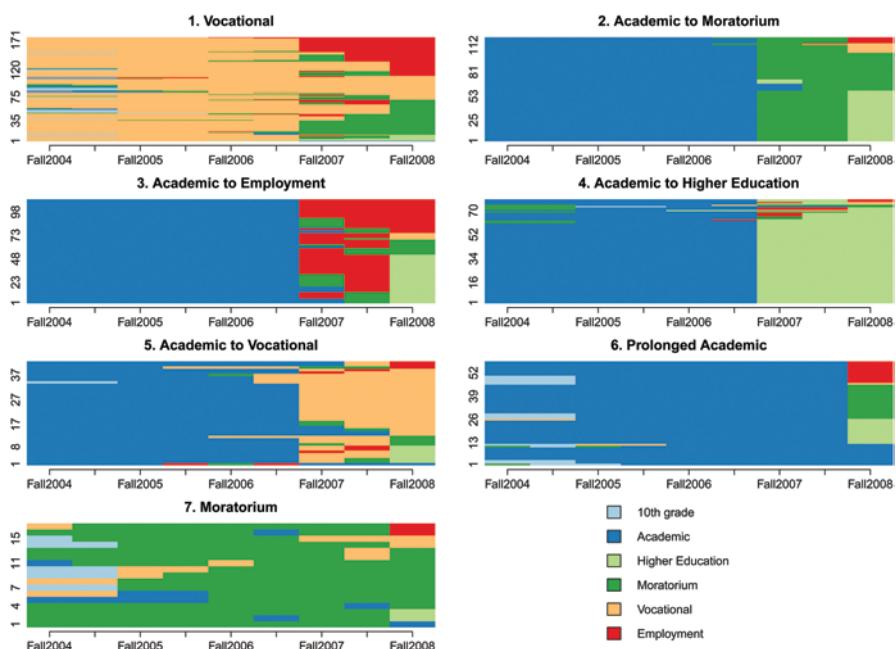


Fig. 7.2 Distribution of sequences within each transition pattern. (Fall 2004– Fall 2008)

Table 7.1 Descriptive statistics for the seven transition patterns on career goal appraisal in comprehensive school (age 16) and background variables

Transition pattern	N	Achievement	Gender	Parents' SES	Career goal importance	Career goal effort	Career goal attainability	Career goal progress	Career goal stress
			% male/ female	% low/medium/ high	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Total sample	602	81.58 (7.97)	46/54	12/45/37	5.76 (0.85)	4.62 (1.23)	5.51 (1.03)	4.84 (1.23)	4.30 (1.52)
1. Aca to HE	78	88.12 (5.40)	17/83	5/36/58 ^a	5.87 (0.73)	4.77 (1.07)	5.76 (0.72)	4.58 (1.17)	4.46 (1.40)
2. Aca to work	111	85.13 (4.86)	47/53	9/41/43	6.00 (0.62)	4.93 (1.13)	5.72 (0.81)	5.02 (1.18)	4.42 (1.60)
3. Vocational	175	73.60 (5.72)	56/44	19/51/20	5.52 (1.02)	4.48 (1.18)	5.19 (1.18)	5.15 (1.14)	4.08 (1.50)
4. Aca to moratorium	121	85.49 (5.43)	64/36	11/42/43	5.79 (0.91)	4.58 (1.39)	5.69 (1.07)	5.21 (1.17)	4.63 (1.53)
5. Prolonged academic	58	80.89 (6.91)	38/62	2/45/48	5.74 (0.68)	4.46 (1.24)	5.33 (1.06)	4.47 (1.45)	4.37 (1.55)
6. Aca to vocational	42	83.38 (5.77)	17/83	19/52/29	5.88 (0.62)	4.64 (1.24)	5.64 (0.86)	4.58 (1.36)	3.91 (1.38)
7. Moratorium	17	73.69 (10.21)	47/53	18/59/12	5.71 (0.89)	4.08 (1.74)	5.21 (1.01)	4.25 (1.55)	3.83 (1.60)

^a For parents' SES percentage may not add up to 100% due to missing cases. Achievement (GPA scores) multiplied with 10

made the transition to employment. In this cluster, males are overrepresented while young people with high SES background are underrepresented.

2. **Academic to Moratorium pathway:** The second largest cluster ($n=121$) is fairly homogeneous (discrepancy=.68) and consists of young people who go through one or more gap years after academic upper-secondary education. Most of them enter higher education after one gap year, and most of them are males.
3. **Academic to Employment pathway:** Also common is the move to paid work as a transitory activity subsequent to graduating from an academic track ($n=111$, discrepancy=1.19). This cluster is similar to the previous cluster, as the most common sequence corresponded to a passage to higher education after one year of fulltime working.
4. **Academic to Higher Education pathway:** Less frequent are immediate transitions from the academic track to higher education ($n=78$, discrepancy=.87). Almost all young people in this cluster, after three years of academic track education, move directly to university or polytechnic education. Females and young people with high SES background are the majority in this cluster.
5. **Academic to Vocational pathway:** Less than 10% of young people move to vocational school after completing an academic upper secondary education ($n=42$, discrepancy=1.63). Again, females are the majority in this cluster.
6. **Prolonged academic pathway:** Also infrequent is a pathway where young people, mainly females, stay four years in the academic track before graduation ($n=58$, discrepancy 1.40). The majority of people in this cluster move to moratorium activities, but transition to work and higher education is also common.
7. **Moratorium pathway:** It is very rare to be on a pathway characterized primarily by moratorium activities ($n=17$, discrepancy=2.37). Young people shift between these moratorium activities and vocational education.

Predicting Transition Patterns

As a first analysis step in examining the association between career goal appraisal at the end of comprehensive school (age 16) and transition patterns, we inspected descriptive statistics of the predictor variables (goal appraisals, GPA, gender, and SES) overall and separated by cluster (see Tables 7.1 and 7.2). Goal appraisals are positively correlated both among each other and with GPA. As Table 7.1 shows, the Moratorium cluster is very low in all career goal appraisals except importance. However, this cluster contains only 17 cases and therefore had to be excluded from further analyses. People on moratorium-dominated pathways are a small but interesting group that should be further considered in studies with larger samples.

With the remaining six clusters, we used multinomial logistic regression to predict transition patterns. Each cluster was used as a reference cluster: thus, we compared each cluster to all other clusters. Gender, parents' socioeconomic status (SES), and grade point average (GPA) were used as covariates together with the five goal predictors (goal importance, goal effort, goal attainability, goal progress, and goal stress). Missing values on SES ($n=36$), GPA ($n=37$) and goal appraisal

Table 7.2 Sample correlation matrix for career goal appraisal and academic achievement in comprehensive school (age 16)

Variable	1	2	3	4	5	6
1. Goal importance	—	.50***	.43***	.44***	.13**	.23***
2. Goal effort		—	.41***	.60***	.27***	.20***
3. Goal attainability			—	.52***	-.06	.32***
4. Goal progress				—	.13**	.28**
5. Goal stress					—	.10*
6. Achievement						—

* $p < .05$, ** $p < .01$, *** $p < .001$

($n=117$) were imputed by using multiple imputation with 20 imputed data sets in SPSS 21. As Table 7.3 shows, GPA and gender are the most consistent background factors in predicting which transition pattern a person will pursue. For parents' SES, we find the main difference between low- and high-SES homes.

The effects of goal appraisal are, in general, small when the background factors are controlled for. Lower career goal effort increases the odds of belonging to the Academic to Moratorium than to the Vocational transition pattern, holding the other predictors constant. This seems to contradict the descriptive results presented in Table 7.1, which indicate that goal effort is actually lower in the Vocational pattern than in the Academic to Moratorium pattern. However, the Vocational pattern also has lower GPA scores. Thus, controlling for GPA, higher levels of goal effort make it rather more than less likely for a given individual to enter a vocational track.

Furthermore, lower career goal progress increases the odds of moving to the Academic to Vocational pattern compared to the Academic to Higher Education and Vocational patterns.

Finally, higher career-goal stress makes individuals less likely to end up in the Academic to Vocational pattern compared to the Academic to Moratorium and the Academic to Higher Education pattern. This finding is also reflected in the raw mean differences between both clusters (see Table 7.1).

Changes in Goal Appraisal in Different Transition Patterns

In the final analysis step, we used ANOVA to examine whether changes in career goal appraisal were related to transitional patterns. Again, the Moratorium cluster was not included in the analyses due to the small sample size. Results presented below are based on the original sample instead of the imputed data sets, as both methods yielded the same results.

The dependent variable was one of the four goal appraisals at a time (measured at ages 16, 18, and 20 years), with time as the within-subjects factor and transitional pattern as the between-subjects factor. First, we used goal importance as the dependent variable. As sphericity (equal variances across transition patterns) could not be assumed, the Huynh-Feldt correction was applied to the results. We find a main effect of time,

Table 7.3 Odds ratios (OR [95% CI]) for multinomial logistic regressions of transition patterns on goal appraisal and background variables

Reference cluster	Comparison cluster	Achievement	Gender (ref.=female)	Parents' SES (low) (ref.=high)	Parents' SES (medium) (ref.=high)	Career goal importance	Career goal effort	Career goal attainability	Career goal progress	Career goal stress
1. Vocational	2.	<i>1.48***</i> [<i>1.37, 1.59</i>]	<i>3.26***</i> [<i>1.64, 6.45</i>]	.60 [.22, 1.66] .41+	1.16 [.55, 2.47] .93	.61* [.72, 1.87] .60+	[.40, 93] [.79]	1.12 [.71, 1.78] 1.05	.93 [.59, 1.48] .94	1.29+ [1.00, 1.67] 1.09
3.		<i>1.42***</i> [<i>1.32, 1.53</i>]	<i>1.48</i> [<i>7.6, 2.86</i>]	.41 [.145, 1.17] .20*	.93 [.44, 1.95] .84	1.05 [.51, 1.24] 1.31	[.64, 1.72] [.60+, 1.24]	1.17 1.13	.94 [.60, 1.47] 1.19	
4.		<i>1.55***</i> [<i>1.42, 1.68</i>]	<i>.37*</i> [<i>16, .85</i>]	.37* [.05, .83] .130	.84 [.36, 1.95] 1.30	.60+ [.73, 2.35] 1.49	[.67, 2.04] [.36, 1.01]	1.17 1.29	[.60*, 1.87] [.69, 1.87]	[.86, 1.39] [.88, 1.61] .85
5.		<i>1.35***</i> [<i>1.24, 1.47</i>]	<i>.31*</i> [<i>12, .89</i>]	.31* [.12, .89] .05** [.39, 4.39]	.60 [.68, 4.73] .67	.94 [.58, 1.55] 1.42	[.75, 2.22] [.79, 2.83]	[.75, 2.22] 1.29	[.60*, 1.87] [.69, 1.87]	[.86, 1.27] [.88, 1.61]
6.		<i>1.24***</i> [<i>1.16, 1.33</i>]	<i>.81</i> [<i>40, 1.64</i>]	.81 [.01, .43] .69	.67 [.31, 1.42] .80	.75 [.49, 1.15] 1.38	[.56, 1.53] [.61, 1.53]	.97 1.30	[.56, 1.29] 1.01	[.63, 1.16] 1.11
2. Academic to moratorium	3.	<i>.96</i> [<i>91, 1.02</i>]	<i>.45***</i> [<i>26, .80</i>]	.45*** [.27, 1.77] .34+	.69 [.44, 1.47] .73	.85, 2.25 1.13	[.62, 1.84] .98	[.62, 1.40] 1.04	[.60, 1.45] 1.22	[.68, 1.04] .92
4.		<i>1.05</i> [<i>.98, 1.12</i>]	<i>.11***</i> [<i>05, .24</i>]	.11*** [.09, 1.22] 2.17	.09 [.36, 1.46] 1.54	.65, 1.95 [.66, 1.95] 1.29	[.65, 1.49] [.62, 1.74] 1.54+	[.62, 1.74] 1.15	[.56, 1.29] 1.01	[.86, 1.45] 1.05
5.		<i>.92*</i> [<i>.85, .99</i>]	<i>.09***</i> [<i>04, .24</i>]	.09*** [.66, 7.11] .09*	.09 [.62, 3.83] .57	.67, 2.49 1.23	[.96, 2.46] 1.22	[.68, 1.95] 1.22	[.41, 1.02] 1.22	[.48, 90] 1.22
6.		<i>.84***</i> [<i>.78, .90</i>]	<i>.25***</i> [<i>12, .51</i>]	.25*** [.01, .74] .49	.25*** [.27, 1.21] .91	.78, 1.91 [.71, 2.12] .82	[.78, 1.91] [.54, 1.39]	.91 1.12	[.82, 1.79] [.72, 1.18]	
3. Academic to employment	4.	<i>1.09*</i> [<i>1.02, 1.16</i>]	<i>.25***</i> [<i>12, .52</i>]	.25*** [.13, 1.80] .13	.49 [.46, 1.80] 1.93	.54+ [.47, 1.43] 1.93	[.50, 1.15] [.68, 1.83]	.65+ [.81, 1.78]	[.56, 1.29] 1.01	[.66, 1.13] 1.09
5.		<i>.95</i> [<i>.88, 1.02</i>]	<i>.21**</i> [<i>27, 1.11</i>]	.21** [.02, 1.09] 6.44*	.316+ [.34, 1.51] 2.12	.54+ [.49, 1.61] 1.14	[.56+ [.60, 1.49] 1.14]	.64+ [.58, 1.49] 1.10	[.41, 1.00] 1.02	[.58, 1.05] 1.02
4. Academic to Higher Ed	5.	<i>.87***</i> [<i>.80, .95</i>]	<i>.54+</i> [<i>.82, .93</i>]	.54+ [.29, 2.46] .84	.96, 10.46 [.78, 4.74] .72	.49, 1.78 [.75, 1.87] .89	[.75, 1.87] [.71, 2.14] .94	.93 [.71, 2.14] 1.23	[.60, 1.36] 1.00	[.78, 1.33] 1.02
6.		<i>.80***</i> [<i>.74, .87</i>]	<i>.221+</i> [<i>92, 5.31</i>]	.221+ [.03, 2.68] .04**	.96, 10.46 [.82, 5.46] 2.6	.57, 2.30 [.93, 2.64] .08	[.60, 2.03] [.60, 1.47] 1.24	.53** [.33, 85] 1.24	[.33, 85] 1.24	[.52, 99] 1.24
5. Academic to vocational	6.	<i>.92*</i> [<i>.85, .99</i>]	<i>2.62</i> [<i>94, 7.28</i>]	2.62 [.00, .38] 14.99	.95 [.46, 1.95] 14.99	.79 [.47, 1.33]	.75 [.43, 1.33]	.1.40 [.88, 2.23]	1.31 [.94, 1.82]	

6 = Prolonged academic cluster
+p<.10, *p<.05, **p<.01, ***p<.001

$$F(2, 756.20) = 8.32, p < .001, \eta_p^2 = .02,$$

showing that career goal importance increases linearly over time

$$F(1, 385) = 15.70, p < .001, \eta_p^2 = .04.$$

We find also a main effect of transitional pattern,

$$F(5, 385) = 3.00, p < .05, \eta_p^2 = .04.$$

The Games-Howell post hoc test (equality of error variances could not be assumed) shows that the Academic to Higher Education cluster $M = 6.12, SE = .07$ reported a higher career goal importance than the Vocational cluster $M = 5.82, SE = .06, p = .006$. Finally, we find no interaction between time and transitional pattern,

$$F(9.82, 756.20) = 0.93, p = .46, \eta_p^2 = .01.$$

Next, goal effort was used as the dependent variable, and, again, the Huynh-Feldt correction was applied. In contrast to goal importance, we find no main effect of time,

$$F(2, 768.00) = 0.74, p = .48, \eta_p^2 = .00.$$

Although the main effect of transitional pattern reaches statistical significance

$$F(5, 384) = 2.30, p < .05, \eta_p^2 = .03,$$

the post hoc test shows no statistical differences between the groups. Finally, we find no interaction between time and transitional pattern,

$$F(10, 768.00) = 1.83, p = .05, \eta_p^2 = .02.$$

In the third analysis, we employed goal attainability as the dependent variable. A main effect of time

$$F(2, 770.00) = 4.43, p < .05, \eta_p^2 = .01$$

shows that career goal attainability increases linearly over time

$$F(1, 385) = 8.35, p < .01, \eta_p^2 = .02.$$

Again, the main effect of transitional pattern reaches statistical significance,

$$F(5, 385) = 2.90, p < .05, \eta_p^2 = .04,$$

but the post hoc test shows no statistical differences between the groups. Moreover, we find an interaction between time and transitional pattern

$$F(10, 770.00) = 1.86, p < .05, \eta_p^2 = .02,$$

suggesting that goal attainability develops differently across transitional patterns. The time trends for the different clusters are plotted in Fig. 7.3 (upper panel). As the figure shows, goal attainability increases in the Vocational cluster at the moment of transition from comprehensive school to vocational school and levels off thereafter, while, for example, in the Academic to Employment cluster goal, attainability does not change across that transition but increases across the transition from upper secondary school to employment.

The results for goal progress show no main effect of time

$$F(2, 756.00) = 1.60, p = .20, \eta_p^2 = .00$$

and no interaction between time and transition pattern

$$F(10, 756.00) = .78, p < .65, \eta_p^2 = .01.$$

However, a main effect for transition pattern

$$F(5, 378) = 3.22, p < .01, \eta_p^2 = .04$$

indicates that the Academic to Higher Education cluster reported higher goal progress $M = 5.21, SE = .11$ than the Vocational $M = 4.80, SE = .09, p < .05$, the Academic to Vocational $M = 4.65, SE = .16, p < .05$, and Prolonged Academic clusters $M = 4.60, SE = .15, p < .05$.

Finally, we tested goal stress as the dependent variable. Similar to goal importance and attainability, a main effect of time

$$F(2, 769.43) = 15.85, p < .001, \eta_p^2 = .04$$

shows that career goal stress increases,

$$F(1, 385) = 28.16, p < .001, \eta_p^2 = .08.$$

In addition, a main effect of transitional pattern

$$F(5, 385) = 4.17, p < .01, \eta_p^2 = .05$$

indicates that the Vocational cluster $M = 4.17, SE = .11$ reported lower career goal stress than Academic to Higher Education $M = 4.85, SE = .14, p < .01$, Academic to Employment $M = 4.68, SE = .12, p < .05$, Academic to Moratorium $M = 4.64, SE = .12, p < .05$, and Prolonged Academic clusters $M = 4.77, SE = .18, p < .05$.

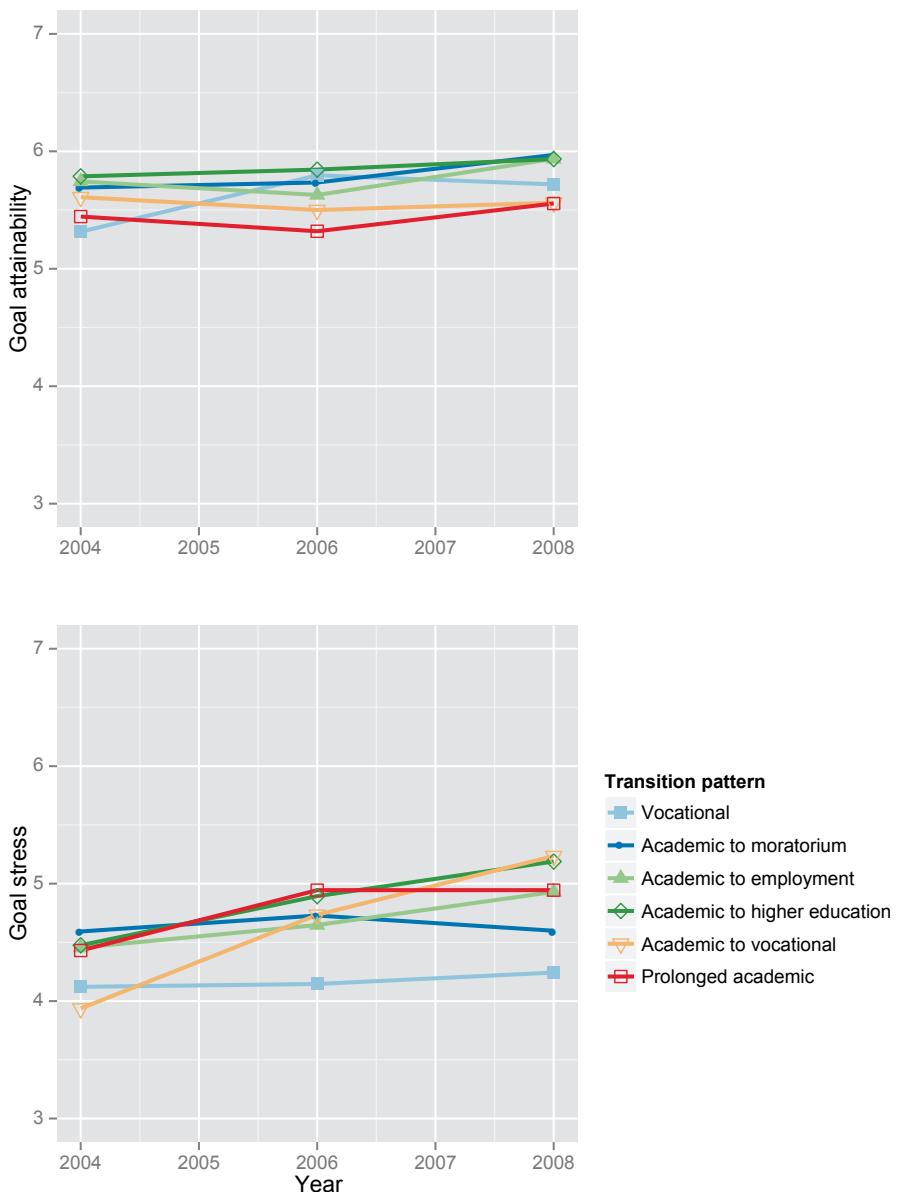


Fig. 7.3 Mean changes of career goal attainability and career goal stress by transition pattern

Finally, we find an interaction between time and transitional pattern,

$$F(9.99, 769.43) = 2.01, p < .05, \eta_p^2 = .03.$$

Inspection of the plotted time trends for the different clusters (see Fig. 7.3, lower panel) reveals that two clusters, Academic to Vocational and Vocational, start lower than the other clusters; however, the Academic to Vocational cluster increases in goal stress over time while the Vocational cluster stays low in goal stress over the three time points. The only other group with no mean change in goal stress is the Academic to Moratorium cluster.

Discussion

This research examined sequential pathways through school and beyond from a developmental psychologists' perspective. It sought to answer three major research questions. First, what are typical sequential patterns during educational transition for young Finns? Second, does young people's career goal appraisal predict transitional patterns beyond the effects of demographic characteristics and school achievement? Third, do developmental changes in career goal appraisal depend on the transitional pathway a person takes?

Transitional Pathways through School and Beyond

Our findings concerning research question 1 revealed seven transitional pathways, among them five patterns related to academic upper secondary education, one pattern related to vocational upper secondary education, and one pattern dominated by moratorium activities. This is in line with our expectation that, first, pathways from the academic track would be more diverse than those from the vocational track. This finding is, second, in line with the expectation to find express pathways (e.g., the Academic to Higher Education pattern), pathways characterized by breaks and gaps (e.g., the Academic to Moratorium pattern), and a pathway of disengagement from education (i.e., Moratorium pattern) (cf. Brzinsky-Fay 2007; McVicar and Anyadike-Danes 2002).

Career Goal Appraisal as Antecedent of Transitional Pathways

Our results further showed small but significant effects of young people's goal appraisal on their transitional patterns (research question 2). High career goal stress in comprehensive school (i.e., at age 16) increased the odds that young people, after graduating from the academic track, would make a transition to moratorium activi-

ties compared to making a transition to vocational education. This supports our hypothesis that the experience of career goal-related stress and strain can make young people disengage from the career domain (i.e., education and employment), at least temporarily. This interpretation, however, does not hold for those young people who made an immediate transition from school to higher education: in comprehensive school, they also reported high goal stress. This seems surprising, but corresponds with existing research showing that the most successful young Finns are also the most stressed (Tuominen-Soini and Salmela-Aro [in press](#)).

Moreover, when holding achievement and demographic background characteristics constant, high effort invested in career goals already in comprehensive school increased the odds of moving to the vocational track compared to taking a pathway from academic track to moratorium. That lower career goal effort would be associated with a transition to moratorium is in line with our expectation and with the predictions of psychological theories (Dietrich et al. 2012; Heckhausen et al. 2010; Nurmi 1992; Salmela-Aro 2009). The Academic to Moratorium transition pattern is characterized by an interruption in young people's educational careers that could indicate a less successful post-school transition. However, this pattern also consisted of males serving in the military or civil service, although our analyses suggest an effect of goal effort on transition pathway independent of gender. Regardless, the role of gap years and the extent to which these are detrimental or beneficial for young people's career development (Wells and Lynch 2012) should be explored in greater depth before final conclusions on this pattern can be made.

Finally, higher levels of goal progress increased the odds of taking up vocational education or moving from academic track school directly to higher education. These two patterns could consist of young people who are determined to strive for their career goal and who thus rate the attainment of this goal higher than those first taking up academic and then vocational education.

In summary, although there was some indication that early career-goal appraisal plays a role in young people's subsequent educational pathways, academic achievement was the strongest predictor of transition pathway. Also, as is well established in the literature, we found effects of gender and SES: females and adolescents from high-SES homes are more likely to pursue high levels of education.

Transitional Pathways and the Development of Career Goal Appraisal

Concerning our final research question, 3, the results showed that the development of career goal attainability and goal stress was associated with the transitional pathways young people took. The attainability of career goals increased from ages 16 to 18 for people taking up vocational education after comprehensive school but stayed stable for all other groups. Moreover, young people who made a transition to employment after graduation from the academic track showed an increase in

attainability during that transition, i.e., from ages 18 to 20. This suggests that transitions to more work-related environments, compared to academic environments, possibly enhances the belief that one will succeed with one's career goals. There are at least two explanations for this. It could be that young people making transitions to vocational education or employment perceive such work-related environments as less demanding than their previous academic environments (see also Salmela-Aro and Tynkkynen 2012). In turn, they show rising attainability beliefs for their career goals (Tynkkynen et al. 2012). Another explanation is social comparison effects (i.e., Marsh 1987), which could have operated while adolescents were still in comprehensive school. In comprehensive school, for low-achieving adolescents, there are manifold opportunities for comparing oneself to higher-achieving students. This typically dampens their expectation to succeed. However, when low-achieving students make a transition to academically less-demanding environments, the options for upward comparison decrease and their success expectations increase. It is possible that such comparison effects also operate for the attainability appraisals of career-related goals. At age 16, in comprehensive school, those with lower achievement also have low goal-attainability beliefs. Then, after the transition to vocational school, they experience a rise in those appraisals.

With respect to career goal stress, we found that young people on academic pathways reported more stress than those on vocational pathways, which corresponds with research on school burnout (Vasalampi et al. 2009). Moreover, we found that goal stress increased for most young people entering academic education, again in line with previous research on school burnout (Salmela-Aro and Tynkkynen 2012). Interestingly, young people in the Academic to Vocational pattern, although they started with very low stress levels, showed the strongest rise, ending up at the same levels of goal stress as their peers moving to university after academic upper secondary school. It is possible that adolescents shifting from an academic to a vocational pathway experience misfit between their career goals and educational track; thus, those goals are increasingly associated with strain and stress. Finally, goal stress started on high levels in comprehensive school for those in the Academic to Moratorium pattern but did not increase over time. This finding, again, could be an indication that adolescents moving to moratorium activities after school temporarily disengage from educational demands and take gap years (Wells and Lynch 2012).

In conclusion, the findings of this study were twofold. They suggest, first, that career goal appraisal (although only weakly when achievement and background characteristics were taken into account) predicted transition sequences through the Finnish education system. Second, our findings suggest that career goal appraisal was also shaped by the series of educational transitions young people went through. Our study thus presented an application of sequence analysis in the field of developmental psychology which linked two types of longitudinal data: the sequences of young people's educational and employment activities and longitudinal data on career goal appraisal.

References

- Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data. *American Journal of Sociology*, 96, 144–185.
- Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29, 3–33. doi:10.1177/0049124100029001001.
- Abele, A. E., & Spurk, D. (2009). The longitudinal impact of self-efficacy and career goals on objective and subjective career success. *Journal of Vocational Behavior*, 74(1), 53–62. doi:10.1016/j.jvb.2008.10.005.
- Asisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The “Second Wave” of sequence analysis bringing the “Course” Back Into the Life Course. *Sociological Methods & Research February*, 38, 420–462.
- Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120(3), 338–375. doi:10.1037/0033-2909.120.3.338.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Biemann, T., Zacher, H., & Feldman, D. C. (2012). Career patterns: A twenty-year panel study. *Journal of Vocational Behavior*, 81(2), 159–170. doi:10.1016/j.jvb.2012.06.003.
- Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review*, 23(4), 409–422.
- Bühlmann, F. (2008). The corrosion of career? Occupational trajectories of engineers and business economists in Switzerland. *European Sociological Review*, 24(5), 601–616.
- Caspi, A., Moffitt, T., Thornton, A., Freedman, D., Amell, J. W., Harrington, H., Smeijers, J., & Silva, P. A. (1996). The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research*, 6, 101–114.
- Cohen, P., Kasen, S., Chen, H., Hartmark, C., & Gordon, K. (2003). Variations in patterns of developmental transmissions in the emerging adulthood period. *Developmental Psychology*, 39(4), 657–669.
- Dietrich, J., Jokisaari, M., & Nurmi, J.-E. (2012a). Work-related goal appraisals and stress during the transition from education to work. *Journal of Vocational Behavior*, 80(1), 82–92. doi:10.1016/j.jvb.2011.07.004.
- Dietrich, J., Parker, P., & Salmela-Aro, K. (2012b). Phase-adequate engagement at the post-school transition. *Developmental Psychology*, 48, 1575–1593. doi:10.1037/a0030188.
- Eccles, J. S. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, 44(2), 78–89. doi:10.1080/00461520902832368.
- Fouad, N. A., & Bynner, J. (2008). Work transitions. *American Psychologist*, 63, 241–251. doi:10.1037/0003-066X.63.4.241.
- Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. (2008). Mining sequence data in R with the TraMineR package: A user’s guide. <http://mephisto.unige.ch/traminer>. Accessed 01 Apr 2013.
- Havighurst, R. J. (1948). *Developmental tasks and education*. New York: Longman.
- Heckhausen, J., Wrosch, C., & Schulz, R. (2010). A motivational theory of life-span development. *Psychological Review*, 117(1), 32–60. doi:10.1037/a0017668.
- Huang, Q., El-Khoury, B. M., Johansson, G., Lindroth, S., & Sverke, M. (2007). Women’s career patterns: A study of Swedish women born in the 1950s. *Journal of Occupational and Organizational Psychology*, 80, 387–412. doi:10.1348/096317906X119738.
- Lent, R. W., Brown, S. D., & Hackett, G. (1994). Towards a unifying social cognitive theory of career and academic interest, choice and performance. *Journal of Vocational Behavior*, 45, 79–122. doi:10.1006/jvbe.1994.1027.
- Little, B. R. (1983). Personal projects—a rationale and method for investigation. *Environment and Behavior*, 15(3), 273–309. doi:10.1177/0013916583153002.

- Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–388. doi:10.1016/0883-0355(87)90001-2.
- McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society Series A, 165*(2), 317–334. doi:10.1111/1467-985X.00641.
- Nurmi, J.-E. (1992). Age differences in adult life goals, concerns, and their temporal extension: A life course approach to future-oriented motivation. *International Journal of Behavioral Development, 15*, 977–991.
- Nurmi, J.-E., & Salmela-Aro, K. (2002). Goal construction, reconstruction and depressive symptoms in a life-span context: The transition from school to work. *Journal of Personality, 70*(3), 385–420. doi:10.1111/1467-6494.05009.
- Nurmi, J.-E., Salmela-Aro, K., & Koivisto, P. (2002). Goal importance and related achievement beliefs and emotions during the transition from vocational school to work: Antecedents and consequences. *Journal of Vocational Behavior, 60*(2), 241–261. doi:10.1006/jvbe.2001.1866.
- Nurmi, J.-E., Salmela-Aro, K., & Aunola, K. (2009). Personal goal appraisals vary across both individuals and goal contents. *Personality and Individual Differences, 47*(5), 498–503. doi:10.1016/j.paid.2009.04.028.
- Salmela-Aro, K. (2009). Personal goals and well-being during critical life transitions: The four C's—channelling, choice, co-agency and compensation. *Advances in Life Course Research, 14*(1–2), 63–73. doi:10.1016/j.alcr.2009.03.003.
- Salmela-Aro, K., Kiuru, N., Nurmi, J.-E., & Eerola, M. (2011). Mapping pathways to adulthood among Finnish university students: Sequences, patterns, variations in family- and work-related roles. *Advances in Life Course Research, 16*(1), 25–41.
- Salmela-Aro, K., & Suikkari, A. (2008). Letting go of your dreams—Adjustment of child related goal appraisals and depressive symptoms during infertility treatment. *Journal of Research in Personality, 42*(4), 988–1003. doi:0.1016/j.jrp.2008.02.007.
- Salmela-Aro, K., & Tynkkynen, L. (2012). Gendered pathways in school burnout among adolescents. *Journal of Adolescence, 35*(4), 929–939. doi:10.1016/j.adolescence.2012.01.001.
- Savickas, M. L. (2011). *Career counseling*. Washington: American Psychological Association.
- Sheldon, K. M. (2002). The self-concordance model of healthy goal-striving: When personal goals correctly represent the person. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 65–86). Rochester: University of Rochester Press.
- Statistics Finland. (2010). Education 2010. http://www.stat.fi/til/sijk/2008/sijk_2008_2010-03-23_tie_001_en.html. Accessed 10 May 2013.
- Stovel, K., Savage, M., & Bearman, P. (1996). Ascription into achievement: Models of career systems at Lloyds Bank, 1890–1970. *American Journal of Sociology, 102*(2), 358–399.
- Tuominen-Soini, H., & Salmela-Aro, K. (in press). Schoolwork engagement and burnout among Finnish high school students and young adults: Profiles, progressions, and educational outcomes. *Developmental Psychology*. doi:10.1037/a0033898.
- Tynkkynen, L., Tolvanen, A., & Salmela-Aro, K. (2012). Trajectories of educational expectations from adolescence to young adulthood in Finland. *Developmental Psychology, 48*(6), 1674–1685. doi:10.1037/a0027245.
- Vasalampi, K., Salmela-Aro, K., & Nurmi, J.-E. (2009). Adolescents' self-concordance, school engagement and burnout predict their educational trajectories. *European Psychologist, 14*, 332–341. doi:10.1027/1016-9040.14.4.332.
- Vasalampi, K., Salmela-Aro, K., & Nurmi, J.-E. (2010). Education-related goal appraisals and self-esteem during the transition to secondary education: A longitudinal study. *International Journal of Behavioral Development, 34*(6), 481–490. doi:10.1177/0165025409359888.
- Vasalampi, K., Nurmi, J.-E., Jokisaari, M., & Salmela-Aro, K. (2012). The role of goal-related autonomous motivation, effort and progress in the transition to university. *European Journal of Psychology of Education, 27*(4), 591–604. doi:10.1007/s10212-011-0098-x.

- Vondracek, F. W., & Porfeli, E. J. (2008). Social contexts for career guidance throughout the world: Developmental-contextual perspectives on career cross the lifespan. In J. A. Athanasou & R. van Esbroeck (Eds.), *International handbook of career guidance* (pp. 209–225). New York: Springer.
- Wells, R. S., & Lynch, C. M. (2012). Delayed college entry and the socioeconomic gap: Examining the roles of student plans, family income, parental education, and parental occupation. *The Journal of Higher Education*, 83(5), 671–697. doi:0.1353/jhe.2012.0028.

Chapter 8

Sequence Analysis and Transition to Adulthood: An Exploration of the Access to Reproduction in Nineteenth-Century East Belgium

Michel Oris and Gilbert Ritschard

Studying the “Roads to Reproduction” in Historical Demography

Until the Industrial Revolution, within the Malthusian frame, population well-being was always threatened by a ‘naturally’ excessive demographic growth in a world of scarce resources. To avoid the positive check of mortality, the prudent restraint of marriage appeared as the only efficient option. From the second edition of his famous *Essay* published in 1802, Malthus had this intuition, that was considerably strengthened one-hundred-fifty years later by Hajnal (1965), when the author drew a line going from Saint-Petersburg to Trieste and identified on the west side the so-called *European Marriage Pattern* of late marriage and high proportion of final celibacy. From the 1970s, Peter Laslett and many followers demonstrated that ‘nuclear households’ were dominant in the West for centuries. In an effort that became decisive for a proper integration of historical demography and family history, Hajnal came back in 1983 with a text on household formation that described the central position of the life cycle service (i.e., a domestic service at teenage and early adult age) in delaying the age at first marriage, while providing various capital to young adults and distancing the generations until young people could reach, at a relatively late age (27–30), a neolocal establishment (i.e., the new-married establishing households of their own, independent from the parental one). However, Todd (1990) showed that in several parts of preindustrial Western Europe, there also

This paper is part of the research works conducted within IP14 of the National Centre of Competence in Research “LIVES—Overcoming vulnerability: life course perspectives”, which is financed by the Swiss National Science Foundation.

M. Oris (✉) · G. Ritschard
Institute for Demographic and Life Course Studies, LIVES,
University of Geneva, Geneva, Switzerland
e-mail: michel.oris@unige.ch

G. Ritschard
e-mail: gilbert.ritschard@unige.ch

existed a nuclear family system without life cycle service, as well as stem family societies, a multiple family household system, and some marginal types. This last synthesis has been severely criticized and in 1998, Wall and Fauve-Chamoux explicitly wrote that they could not provide a better alternative and that it was better to give up. The typological approach of family systems had reached a dead end.

This is one of the reasons demography moved from populations to individuals as units of analysis and offered its analytical tools to investigate the associations of behaviors and structures forming processes at various moments of individual life courses, the latter being part of family dynamics according to the ‘linked lives’ principle (Elder et al. 2003). Recent studies have focused on configurations and living arrangements and used multivariate statistical regressions to identify the factors that explain the occurrence and timing of several transitions along the life course. Life transitions, however, are still too often considered isolated from each other (Kok 2007).

In this paper, we look at marriage and household formation processes from an original perspective. From the previous research, we retain that marriage is still seen as the most important brake to population growth in Western historical populations. However, its interaction with mobility and its role in the household formation must also simultaneously be taken into account. Italian and Spanish scholars have been pioneers of this integrative approach and more explicit while looking at the impact of various migration systems on family lives and demographic regimes. Carlo Corsini (2000, p. 18) concluded:

Any family system could not maintain itself but through controlling marriage and migration solutions. Migration and marriage appear to be the most sensitive and important factors of family behavior... in historical populations, characterized by ‘natural’ fertility and when mortality is depending mostly on external elements.

Moreover, all societies have formal and socially accepted events sanctioning transitions from one stage of life to another. Marriage has always represented one of the most important transitions in the individual life course, both in past and in contemporary societies (Hareven and Masaoka 1988). Indeed, we believe that a correct approach to studies on marriage and its role should begin from the fact that in almost every society, marriage has represented the socially accepted access to reproduction, and was consequently an essential precondition for the biological survival and continuity of families and populations. For this reason too, marriage should be seen as a transition point on the ‘road to reproduction’ and cannot be analyzed for itself, but in relation to other turning points in the life course process of the so-called ‘transition to adulthood’, i.e., leaving home, family formation, household headship, and inheritance transmission (Shanahan 2000; Dribe et al. 2014, 2010).

Our research sees marriage as one possible event, others being leaving the parental home and first (legitimate) birth, their calendar and successions forming trajectories, or the roads young people could take to access to legitimate reproduction, i.e. to a first legitimate birth. We make use of individual longitudinal data. The settings include rural samples in East Belgium.

Ardennes and the Pays de Herve were the two rural areas in this region, separated by the river Vesdre, on which was located a pioneering centre of the Industrial Revolution in continental Europe: the growing agglomeration of Verviers. The two regions experienced in the first half of the nineteenth century a collapse of their

proto-industrial activities that could not resist to the competition with the modern factories, and a consequent ‘ruralisation’. After a period of tensions and decline in economic well-being, the second half of the nineteenth century was a time of betterment, because the excess population emigrated to industrial towns and the growing demand for food from the urban areas benefited the peasants. The agricultural crisis of 1873–1890 just accelerated on-going changes (Neven 2002).

Data and Data Management

Local Sources and Data

Population registers are the original data sources. In a few villages of Ardennes, they were implemented by the local communities to control the right of access to the common lands. From 1846, the Central Committee for Statistics and more specifically its most prominent figure, Adolphe Quetelet, obtained a legal organization and its general use in each Belgian municipality. The basic principle is to start from a population census, to copy the household census sheets in large books, one page per household, one line per individual, and then to start a continuous update. Newborns are added below on their household page and immigrant families are indicated on a new page each; deaths and emigrations are mentioned in special columns; marriages and new households also have to be reported. For each event, the exact date—and when relevant, the location—are indicated. The aim was to have a continuous, complete, and reliable view of the population for statistical and administrative purposes. Theoretically it is a wonderful tool.

In the real world, the succession of additions, new information, new individuals, or new households often resulted in a quite confused document, sometimes very difficult to read and interpret. Moreover, if most of the people moving in took the initiative to declare their arrival in a locality, emigrants regularly just disappeared. In addition, for the local civil servants, this was a register of the living population that they updated more or less regularly, sometimes every four to five months; therefore, newborns who died at an early age tended to be underreported.

The authorities realized that completely replacing censuses with the population registers was a dream. A new census was organized approximately every ten years from 1856, with a new series of population registers reestablished on this reliable base each time. Many rural municipalities, however, wanted to reduce their workload and maintained their population registers over 20 years.

In our databases, we first entered the successive population registers for the nineteenth century. Then we verified the completeness of the marriages, births, and deaths through a comparison with the civil registers. Third, we linked individual records in the successive population registers, so we identified people who ‘appeared’ as well as those who ‘disappeared’. For them, we attributed dates of immigration or emigration using imputation techniques (Alter et al. 2009). We will see below that the results of this long and tedious work are globally excellent.

Table 8.1 Codification of the states in the roads to reproduction

Code	Label	Experienced Events
H	At home	None
LH	Left home	LH
CH	Child at home	FB
C	Child out of the home	LH and FB
MH	Married at home	FM
M	Married	LH and FM
MCH	Married w/child at home	FM and FB
MC	Married w/child	LH, FM and FB

For the purpose of this chapter, we have integrated two samples, one from Ardennes in the commune of Sart (Alter et al. 2004) and one from the Pays de Herve with a cluster of three municipalities: Charneux, Clermont, and Neuchâteau (Neven 2003). We have reconstructed life-sequence data from 1812 to 1900 in addition to the socioeconomic status (SES) of the individuals. SES classification reflects both a simple social structure in the studied villages and the absence of systematic tax information in the Belgian local archives. We could only use the occupation, which is systematically mentioned in our sources, although it implies many missing data for teenagers and women. In our typology, low SES persons are mainly daily laborers and other unskilled workers. The medium SES represents the heart of those rural societies, because this includes the peasants (cultivators, a few farmers). Artisans are in this group in the Ardennes sample and in the low category in the Pays de Herve sample, because in this region, proto-industrial workers faced quite hard times during the studied period as a result of the Industrial Revolution in the neighboring city of Verviers. An upper class (high SES) is present in our typology but has a quite limited demographic weight.

Data Management and Critical Evaluation

The analysis is based on a range of 34 years with annual state sequences from age 12 to 45. The states are defined by three basic events: leaving home (LH), first marriage (FM) and first childbirth (FB). They are defined in Table 8.1. We used TraMineR (Gabadinho et al. 2009, 2011) for managing the sequence data as well as for producing all plots and sequence analysis results shown in this article.

Analyses of the trajectories from leaving the parental home until legitimate reproduction could be severely biased by left censoring, while we dropped all the individuals who experienced one of the three events (leaving home, first marriage, first birth) before they immigrated to our studied villages. When observation started after age 12 and individuals had not experienced any of the events at their first observation, we completed the sequences backward with the state H (Home). Similarly, for cases falling out of observation before age 45 but that had experienced all three events at their last known state, we completed the sequences forward with the

Table 8.2 State to event transformation matrix: Each cell lists the events assumed to occur when we change from the origin row state to the destination column state

	H	LH	CH	C	MH	M	MCH	MC	*
H	“H”	“LH”	“FB”	“LH, FB”	“FM”	“LH, FM”	“FM, FB”	“LH, FM, FB”	“*”
LH	“H”	“LH”	“”	“FB”	“”	“FM”	“”	“FM, FB”	“*”
CH	“”	“”	“FB”	“LH”	“”	“”	“FM”	“LH, FM”	“*”
C	“”	“”	“”	“LH, FB”	“”	“”	“a”	“FM”	“*”
MH	“”	“a”	“”	“”	“FM”	“LH”	“FB”	“LH, FB”	“*”
M	“”	“”	“a”	“”	“a”	“LH, FM”	“a”	“FB”	“*”
MCH	“”	“”	“a”	“a”	“”	“”	“FM, FB”	“LH”	“*”
MC	“”	“”	“”	“a”	“”	“”	“a”	“LH, FM, FC”	“*”
*	“H”	“LH”	“FB”	“LH, FB”	“FM”	“LH, FM”	“FM, FB”	“LH, FM, FC”	“*”

The “*” stands for the missing state and event ‘a’ corresponds to transitions which were observed in the data but should not occur by definition. Indeed, each state has been calculated at each individual’s anniversary only with the information available at the given moment, without considering the individual status before or after that date

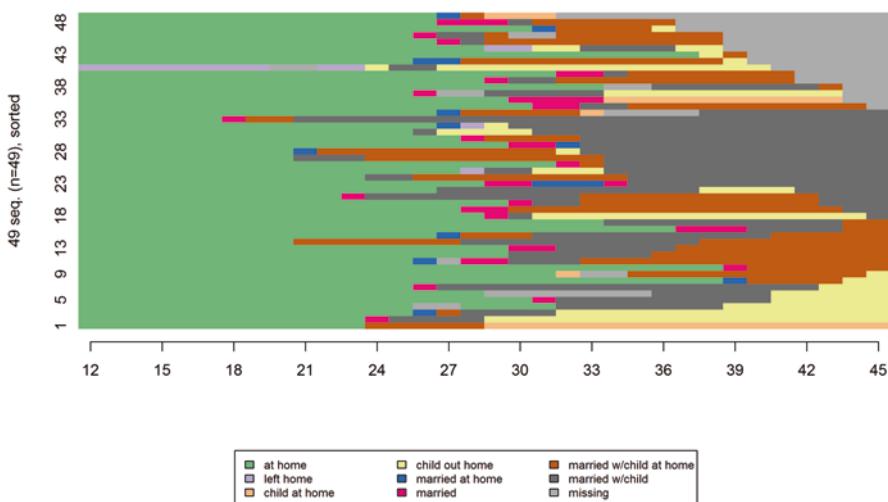


Fig. 8.1 Trajectories with invalid transitions

state MC (married w/child). Then we selected only cases with less than 15 missing states in their sequence. This left 2,511 individual trajectories.

From the state sequences we reconstructed the sequences of original events using the method described in Ritschard et al. (2009) with the transformation matrix in Table 8.2.

We are consequently in a situation where we can make an ex-post evaluation of the data coherence. We found 49 incoherent trajectories, shown in Fig. 8.1, which we dropped out of the following analyses. Considering the complexity of the original data sources, excluding only 1.9 % of the individuals because of inconsistencies

Table 8.3 Distribution of all possible trajectories (%) from living in the parental home to experiencing a first legitimate birth (Nineteenth-century, rural East Belgium)

Trajectories		Men	Women
N	Types		
1	LH-FM-FB	5.7	3.8
2	LH-FB-FM	0.7	0.4
3	FM-LH-FB	9.8	11.1
4	FM-FB-LH	13.8	12.8
5	FB-FM-LH	1.7	3.5
6	FB-LH-FM	0.2	0.7
7	LH/FM-FB	27.9	28.9
8	FB-LH/FM	2.1	3.2
9	FM-FB	33.8	29.6
10	FB-FM	3.9	5.7
11	OTHER	0.2	0.2
	TOTAL	100.0	100.0

demonstrates the high reliability of our documentation, especially if we take also into account that we study here the turbulent years from teenage to adulthood.

The Roads to Reproduction: Cross-Sectional Expectations, First Longitudinal Results

Table 8.3 provides a first picture of the trajectories to legitimate reproduction through a simple alphabet (the one described above in Table 8.1) with an elementary sequential structure (– means a succession, / a simultaneity of events). The results immediately illustrate the many roads young people could take to reach a first legitimate birth in nineteenth-century East Belgium. While the common view is that those rural societies severely controlled the behaviors of the young adults (Servais and Alter 2005; Oris et al. 2014), we observed more than 11 combinations. Moreover, this large variety is accentuated by the absence of a clearly dominant pattern.

Indeed, a first marriage while staying in the parental home, followed for the newlywed by a first birth (trajectory 9, FM-FB) was the most common pattern, but counted for only one-third of the men and a bit less for the women. Moreover, while globally, East Belgium was an area dominated by a culture of a family system without life cycle service, this kind of trajectory is what we expect to find in a very different family system, the stem system, where one children remains in the parental home to receive the inheritance and to become the head of the family. Marriage marks the designation of this heir (or more rarely, heiress) and the new couple cohabitantes with the parents until their death (see Fauve-Chamoux and Ochiai 2009 for a more extensive discussion of the stem family system).

The second-most frequent trajectory from the parental home until legitimate reproduction is type 7, LH/FM-FB. Some 28% of the young adults left the parental home when they married, and this neolocal establishment was followed by a first birth. From our previous knowledge based on cross-sectional data, this is what we

expected to be the dominant pattern. We have here a classical illustration of a discrepancy between transversal and longitudinal data. Indeed, in a demographic regime characterized by a late age at first marriage (average age 27 for women, almost 30 for males) and a low life expectancy (around 40–45), cohabitation between a newly married couple and the parents (of the groom or the bride) tended to be short. This implies that when we look at the household structures in the census each ten years, we miss most of those short phases. However, when we use longitudinal data we see that they were indeed frequent in the life courses of young adults as an intermediary phase on their roads to autonomous settlement.

In nineteenth-century rural East Belgium, type 9 was a bit more frequent than type 7. It becomes even clearer if we consider that type 4 (with 13/14% of the trajectories) expresses the same pattern as type 9.

We also note that in 8% of the male trajectories and 13% of the female paths (types 5, 6, 8, and 10) the first birth was an illegitimate one, occurring before the marriage. However, it ended with a formal union, followed by a first legitimate birth. Those high proportions in a nineteenth-century rural context are coherent with a very late age at marriage, which automatically implies an extension of the duration at risk of delivering an illegitimate child (Oris et al. 2014).

We face here rural, supposedly traditional societies with high levels of complexities. At least three results emerge from Table 8.2. First, in a context of demographic pressure (more births than deaths each year) and reduction of the economic opportunities for settling a household ('ruralisation' with the disappearance of artisanal activities and a tendency to fewer but larger farms), young adults who did not emigrate to the industrial towns accessed marriage late and were frequently obliged to cohabit with parents, waiting their death to inherit the farm.

Second, in this highly constrained context, spaces for freedom or human agency existed. Young adults could eventually reach legitimate reproduction through many roads. An international team calculated the measures shown in Table 8.2 for preindustrial settings in Sweden, Italy, and Japan and the comparison demonstrated the higher liberty of the young living in nineteenth-century rural East Belgium (Dribe et al. 2010).

Third, freedom was both obvious and relative, because around one in ten broke the religious and legal rules and had an illegitimate child. They regularized their situation later. This pattern demonstrates the ability of the local societies to integrate some marginal trajectories and reconstruct a threatened social order.

More globally, this is also an expression of the tensions between parents and their adult children, between rural communities and their young adults, between the local notables and the lower classes (daily laborers). On one hand, at a community level, preserving the balance between population and scarce resources implied delaying the access to reproduction and at an individual/family level to oblige the fiancés to wait for an economic unit to become available before marrying and settling in a household of their own. On the other hand, young adults had to wait for long periods and sometimes used premarital sexual relations to impose their own choice of a partner and/or obtain an early marriage. We know that many brides were pregnant when they married, because some 40% had a first birth in the eight months following the wedding (Oris et al. 2014).

Sequence Analysis of Access to Legitimate Reproduction

Although those first results are of great interest and already represent a decent contribution, two limitations have to be immediately acknowledged. First, we just observed the sequences between the events but not the status duration. In other words, two ‘roads to reproduction’ that followed the pattern LH-FM-FB could seem exactly the same. However, it is possible that one left home at 14, stayed in life cycle service for 15 years before a marriage at 29, while for the second, those values could have been 18, 2, and 20, respectively.

Second, if we cross-tabulate the possible trajectories to first birth with a pertinent variable, socioeconomic status for example, we face a multitude of little numbers that are difficult, if not impossible, to interpret. And without considering the timing of events and social position variables (e.g., SES), our approach to the articulation of individual trajectories with social and demographic regulations remains superficial. Especially, we do not see how the individual dynamics, or relative individual freedom, resulted in a global adjustment between demographic trends and economic resources at a community level. Capturing this articulation is crucial to our understanding of the delicate trade-offs observed in all the previous studies on preindustrial Western societies. In the next pages, we show that sequence analysis emerges as a highly relevant tool to reach this objective.

The series of cross-sectional distributions in the top panels of Fig. 8.2 shows that both males and females exhibited a pattern typical of a nuclear family system without life cycle service. The ‘left home’ status seems almost absent. From ‘at home’ (being at home, single, without children) to ‘married w/child’ (married with child in a household of his/her own), only a limited area of transitional statuses appears. Women experienced their transitions younger than men; otherwise the figures for males and females are very similar.

However, below the series of cross-sectional distributions are also drawn the index-plots, which add relevant information by depicting the individual trajectories and their diversity. The statuses associated with stem phases of cohabitation ('child at home', 'married at home', 'married with child at home') are more frequent at higher ages and more frequent among men than women. The first observation supports the interpretation that those people stayed with their parents because they waited for their inheritance of the family property. From a legal point of view, women were not disadvantaged during the successions, but brothers usually compensated their sisters, so the transfers tended to be in the male line.

We conducted a second separate cluster analysis for males and females using optimal matching (OM) pairwise dissimilarities between the sequences. Since the states were derived from the experienced events (leaving home, first marriage, first birth), we used substitution costs reflecting the number of mismatches on events characterizing each state, i.e., the Euclidean distance between the rows of the property (see Table 8.4).

We set the indel cost as half the maximum substitution cost.

We obtained the clusters by partitioning around medoids (pam) (Kaufman and Rousseeuw 2005; Studer 2013). Examining the range of solutions for $k=2$ to 10

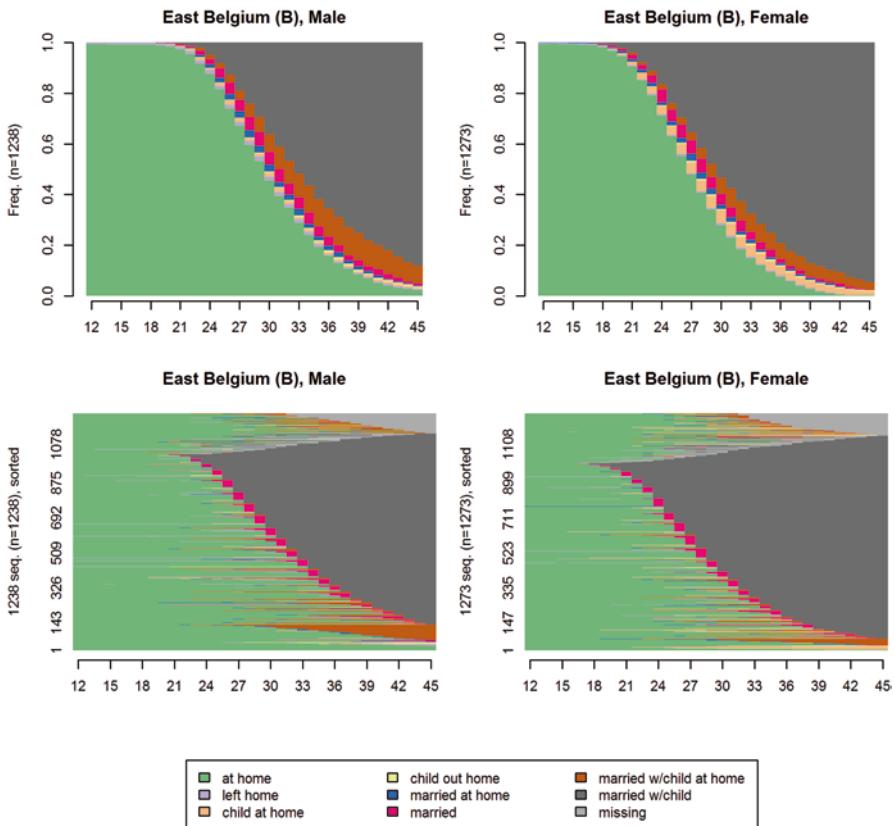


Fig. 8.2 Plots by sex. *Top*: Time evolution of cross-sectional distributions. *Bottom*: Index-plots sorted by state from right to left

Table 8.4 State properties in terms of occurred events

	Left home	Married	Birth
H	0	0	0
LH	1	0	0
CH	0	0	1
C	1	0	1
MH	0	1	0
M	1	1	0
MCH	0	1	1
MC	1	1	1

groups, the four-group partition appeared to be, for men as well as for women, a good compromise solution regarding a series of partition quality measures. The average silhouette width of the four-cluster solution is .38 for men and .35 for women and the R^2 (proportion of explained discrepancy) respectively 45 and 47% (see for

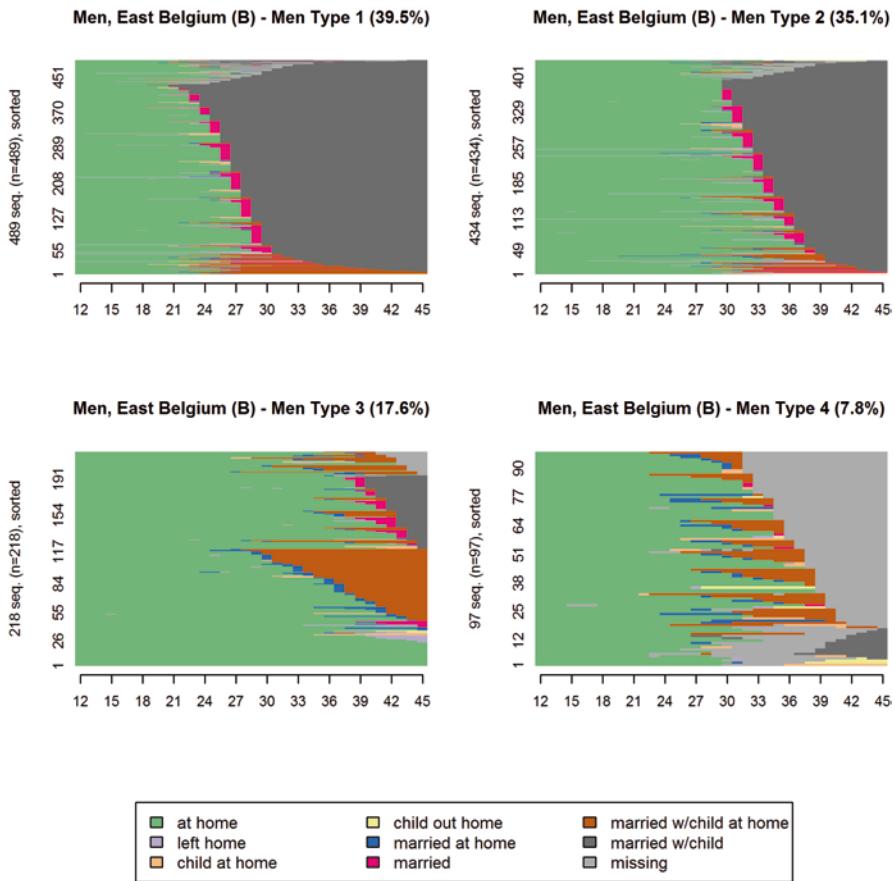


Fig. 8.3 Clusters of sequences. *I-plots, sorted by state from right to left. Men*

example Studer 2013 for details about the measures). The obtained clusters are visualized with index-plots in Fig. 8.3 for men and in Fig. 8.4 for women. The advantage of well-sorted index plots over chronograms is that they render the discrepancies within the clusters in addition to the between-clusters differences.

Among men (Fig. 8.3), the timing of transitions to adulthood is the discriminating factor for the three first groups. Indeed, while the average age at first legitimate birth for males was around 32 (with a median at 31), type 1 brought together almost 40% of the trajectories, with this transition observed at 27 on average. For type 3, with 17.6% of the male population, the turning point was very late (mean=38.64 with the median at 40). Type 2 (35%) was close to the global male average values (Table 8.5). Finally a little group (less than 8%) brings together people who emigrated from the villages at some moment in their life trajectory, creating situations of right censures and consequently missing information.

Tamara Hareven explained at which point marriage and the access to reproduction are normative transitions, especially with a ‘proper age at marriage’ and social

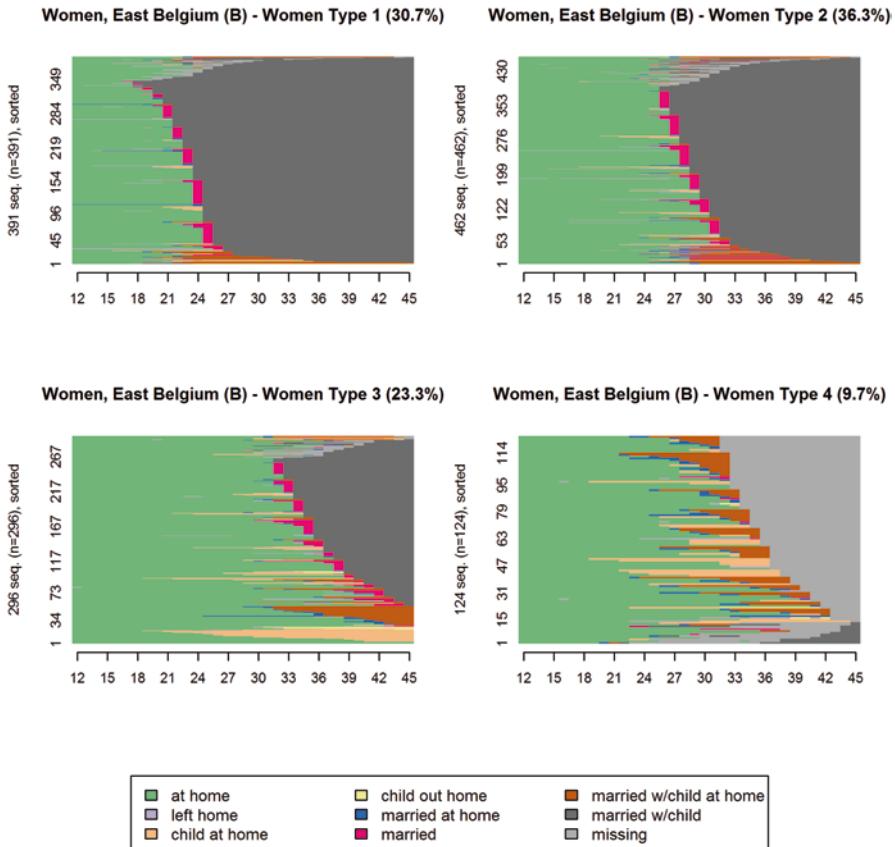


Fig. 8.4 Clusters of sequences. *I-plots, sorted by state from right to left.* Women

Table 8.5 Age at first legitimate birth, by sex and cluster

	Men				Women			
	Median	Mean	Sd	Cv	Median	Mean	Sd	Cv
Type 1	27	27.23	2.68	0.10	24	23.75	2.25	0.09
Type 2	34	34.09	2.87	0.08	29	29.44	2.31	0.08
Type 3	40	38.64	4.35	0.11	36	36.54	3.48	0.10
Type 4	31	32.57	5.27	0.16	31.5	32.12	5.21	0.16
Total	31	31.75	5.31	0.17	29	29.33	5.51	0.19

stigma on those who did not experience the transition at the right time, both the young (suspected of sexual debauchery) and the ‘old’ unmarried (named old spinsters, old bachelors and seen as the ‘residuals’ from the matrimonial market, the ‘untaken’) (Hareven and Masaoka 1988). Those social representations are strong enough to create a survival curve for first marriage and childbirth which has a first

plane segment (everybody unmarried without child), followed by a brutal decline during a short period (more or less ten years) during which most of the marriages and first births happen (i.e. the proper age), then a new flat portion at a lower level once the proper age is passed and those who did not succeed in making the transitions during the typical period of their life course have no other choice than final celibacy and taking care of their nephews or nieces (Oris and Ochiai 2002).

Here we observe a clear distinction between a significant proportion (almost 40 %) of young males who broke the secular pattern of late marriage and were pioneers of a modernization of the household formation rules in those remote rural areas of East Belgium (type 1), versus those who respected the norms and married late (type 2), while a little group avoided final celibacy and biological extinction at the very last moments (type 3). Those results are quite original and unexpected. Although one of the two authors of this paper has worked on those populations and databases for more than 20 years, he never suspected the existence of this group 3 of very late access to parenthood, those who reached the objective largely after the proper age.

Index plots show that the stem family phases were present everywhere but clearly more in this type 3. It suggests once again that those men waited for their inheritances and faced the ambivalent situation of seeing their parents surviving until they were old enough that they lost their power and authorized their waiting boy to finally access marriage and reproduction. The sons then took the headship of the household and family business and cared for the elderly parents until their deaths. This kind of transfer, when both generations were still alive, was not known till now in nineteenth-century rural East Belgium but has frequently been observed elsewhere in Europe at the same time, especially in the Scandinavian countries (Dribe 2000). In our case, it becomes a new reasonable interpretation that will require further investigation in the archives, mainly in the notarial acts. However, we will see below that a complementary explanation also emerges.

The female pattern (Fig. 8.4) was similar, with the first 31 % of women transitioning early, around 24, and abruptly, with a little coefficient of variation around the mean (see type 1 on Fig. 8.4 and Table 8.3). Some 36 % had a late transition around 29 (type 2), and 23 % a very late transition to motherhood around 36 or 37 (type 3). Once again, type 4 brought together the cases with missing information on the right; they are less than 10 %. In that group, we see the trajectories of unwed mothers. The yellow colour means they were isolated from their parents and were very probably not old-rooted members of the local communities. Eventually they emigrated after their pregnancies to escape from the public shame, to go back to their own native communities, or to search for a new life in the industrial towns. We also see some yellow lines in the other clusters, designating the few who obtained marriage and regularization.

But in cluster 3, a block of orange lines is concentrated. It represents the unwed mothers who lived with their parents. Their kin were not always able to constrain the gallant to do his duty, but those women were not rejected by their parents. In cluster 3, we also see the stem family phases—less frequent than among their male counterparts, however, for the reasons already discussed above.

Table 8.6 Univariate discrepancy analysis

	Categories	R ²	F	Sig.
Sex	2	0.018	47.2	0.000
Modal SES	4	0.014	11.9	0.000
Last SES	4	0.009	7.9	0.000

Relationships Between Trajectories, Socioeconomic Status, and Sex

From the previous analysis, it seems that only timing really distinguished men and women with transitions two to three years earlier for the latter. We wanted to verify this interpretation of the figures statistically, but also to see if the observed clusters and therefore the internal variance within the local community could be explained by SES. Indeed, when Thomas Malthus ([1803/1992](#)) explained in his essay that the poor were unable to control their sexual appetites, married too early, and made more children than they were able to raise—children consequently condemned to poverty—he just expressed an idea widely shared by the elites of his time all across Europe, including in East Belgium ([Alter and Oris 1999](#); [Neven 2003](#)). A more recent and less moral interpretation is that landless had no land or patrimony to protect, and they could not use their children to work on land they did not have. Parents thus had no rationale to delay the leaving home and marriage of their children ([Oris et al. 2014](#)). That is why we suspect cluster 1 (both for males and females) brings together trajectories of young adults from the low SES, from daily laborer families.

To test those hypotheses and see at which point sex and SES explained the diversity in timing and household formation rules, we first proceeded to a univariate ANOVA-like discrepancy analysis ([Studer et al. 2011](#)) for each of the variables: sex, most frequent SES in the sequence, and last SES in the sequence. The results, shown in Table 8.6, demonstrate that sex and SES explain only a small but statistically significant part of the discrepancy between trajectories. The most frequent SES looks more informative than the last observed SES.

We then grew a regression tree for our sequences ([Studer et al. 2011](#)). Regression trees are obtained by searching first among the covariates for the one that permits the best binary split, i.e., the split which explains the biggest part of the discrepancy and therefore generates the highest R². This operation is then repeated locally at each obtained node. The growing stops when the split is not significant. We used a 0.5% p-value limit and set the maximum tree depth as 4.

As shown in Fig. 8.5, sex explodes the initial root and divides the male and female trajectories. Rules and distributions were, however, quite similar; only the more precocious timing of the women made a real difference. Among men, the splitting procedure separated the few males without occupation and the daily labourers from the peasants and the modest local elites (the middle class). The former had more abrupt and earlier transitions, which tends to confirm the interpretation that daily labourers were freer to marry because they did not have land or patrimony

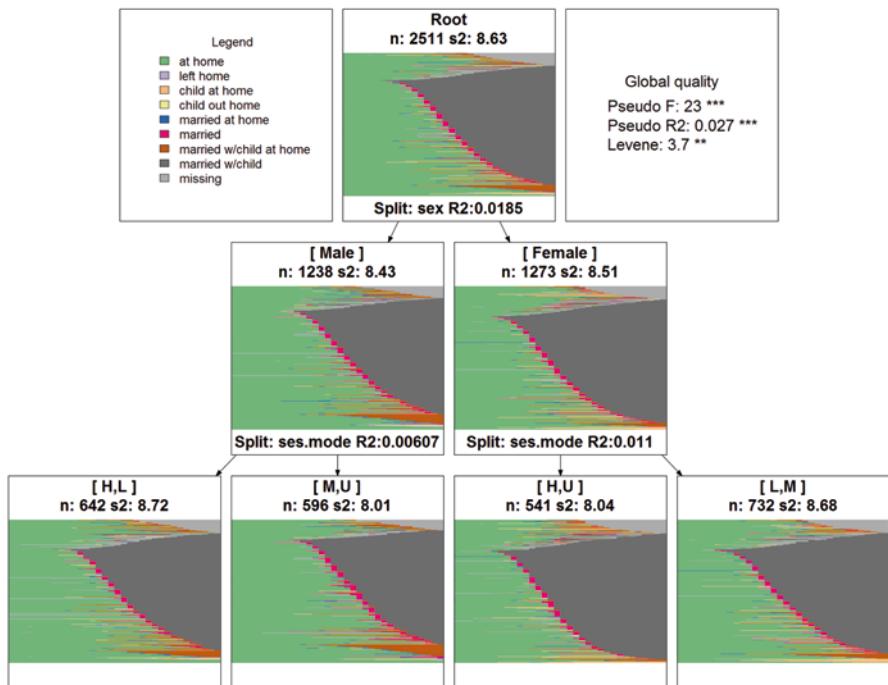


Fig. 8.5 Sequence regression tree. I-plots, sorted from end

to preserve. We find a confirmation of the whole process of transition from leaving the parental home to legitimate reproduction.

However, on the female side, we do not find the same separation. On one side are the few women from the upper classes and those who did not declare an occupation, which was a part of the nineteenth century culture of respectability among the middle classes. On the other side are together the medium (peasants) and low (daily labourers) SES. Moreover, if we look at the average age at first legitimate birth and the variation coefficient (Cv) around this mean, we see few differences between the leaves (terminal nodes of the tree) and we observe a large standard deviation (sd) within each of those leaves (Table 8.7). This is confirmed by the low pseudo R^2 of only 0.027 (Fig. 8.5, Global quality) for the final partition into four leaves, meaning that the tree explains only 2.7% of the total discrepancy.

However, I-plots from the tree provide new information. Indeed, the phases of cohabitation with grand-parents are also present among the men of low SES who had nothing or little to leave as an inheritance. Going back to the original data, we observe a pattern which is not a stem pattern but an adaptation of the nuclear family rules. Indeed, newly married couples with children welcomed a widowed parent in their household, usually a grand-mother. The latter was not abandoned in an ‘empty nest’ and probably took care of the grand-children so that the mother could work more (Oris and Ochiai 2002).

Table 8.7 Age at first legitimate birth, by leaf of the tree

	Median	Mean	Sd	Cv
Male, H or L	30	31.08	5.33	0.17
Male, M or U	32	32.48	5.19	0.16
Female, H or U	28	28.23	4.98	0.18
Female, L or M	30	30.16	5.74	0.19

Conclusion

The main objective in this chapter was to look at internal variances within the population considered. We also wanted to explore the relationships between family systems and socioeconomic structures through an analysis of longitudinal individual-level data. Using optimal matching to analyze the trajectories systematically, and then cluster analysis and a regression tree to classify them, we demonstrated that internal variations are important to understand a social and economic organization. It appears that a system is a combination of individual differences in behaviors.

Peasants and daily laborers from Ardennes and the Pays de Herve lived in a demographic regime dominated until the end of the nineteenth-century by a very late age at marriage. In a context of uncontrolled fertility, it was the only solution to reduce the family size. However, this global pattern was created by an impressive diversity of trajectories, a diversity that illustrates the importance of individual agency in a highly constrained context for those who decided to stay in the villages. The completely unexpected discovery of a group, both on the male and the female sides, able to triumph over the norm of a proper age at marriage—moreover, in a demographic regime where marriage played such a crucial role—proves the performance of the data-mining techniques (Ritschard and Oris 2005). We were so structured by the literature and previous results that we simply never looked at those people until they appeared in group 3 of Figs. 8.3 and 8.5. A new narrative became part of the story as a whole.

This contribution also shows the point at which the usual approach of household structure or living arrangements can hide the complexity of the real world and real life trajectories. If we come back to Fig. 8.2 (Top), we face an ideal type: a proto-nuclear family system without life cycle service. But looking inside, we discovered that on the roads to reproduction, men and women could respect various set of rules, and that an internal diversity could be demonstrated and measured. In East Belgium, the internal diversity was striking, mainly due to variations in the timing of transitions in the former case. We investigated the impact of sex and SES; however, the statistical evaluation of our decomposition shows that the larger part of the individual variability is not captured by those variables.

According to Buckx (2009), the transitions of youth into adulthood cannot be fully understood if the parental-children relations are not properly taken into account. Our own results about the cohabitation with parents and grand-parents, resulting from a stem family logic or from an adaptation of the ‘nuclear’ family rules,

support this view. Moreover, previous analyses of leaving home and first marriages using event-history methods demonstrated the importance of the young adult's position in their sibling group (Alter and Oris 1999; Bras and Neven 2007). The next methodological and substantial challenge will be to properly consider those linked lives dimensions.

References

- Alter G., Oris M. (1999). Access to marriage in the East Ardennes during the nineteenth century. In I. Devos & L. Kennedy (Eds.), *Marriage and rural economy. Western Europe since 1400* (pp. 133–151). Turnhout: Brepols.
- Alter, G., Neven, M., Oris, M. (2004). Mortality and modernization in Sart and surroundings. In T. Bengtsson, C. Campbell, & J. Lee (Eds.), *Life under pressure. Mortality and living standards in Europe and Asia, 1700–1900* (pp. 173–208). Cambridge: MIT Press.
- Alter, G. C., Devos, I., & Kvetko, A. (2009). Completing life histories with imputed exit dates: A method for historical data from passive registration systems. *Population*, 64(2), 293–318.
- Bras, H., Neven, M. (2007). The effects of siblings on the migration of women in two rural areas of Belgium and the Netherlands, 1829–1940. *Population Studies*, 61(1), 53–71.
- Buckx, A. J. E. H. (2009). *Linked lives: Young adult's life course and relations with parents*. (Unpublished doctoral dissertation). Utrecht University, Utrecht, Netherlands.
- Corsini, C. (2000). Introduction. In M. Neven & C. Capron (Eds.), *Family structures, demography and population. A comparison of societies in Asia and Europe* (pp. 9–21). Liège: University of Liège.
- Dribe, M. (2000). *Leaving home in a peasant society. Economic fluctuations, household dynamics and youth migration in southern Sweden, 1829–1866*. Södertälje: Almqvist & Wiksell International.
- Dribe, M., Manfredini, M., Oris, M., & Ritschard, G. (2010). Pathways to reproduction in pre-transitional Europe. A sequential approach. Rethinking reproductive change in historical and contemporary populations. Social Science History Association, Chicago, Nov 17–20.
- Dribe, M., Manfredini, M., & Oris, M. (2014). The roads to reproduction: Comparing life course trajectories in preindustrial Eurasia. In S. Kurosu, C. Lundh et al. (Eds.), *Similarity in difference: Marriage in Europe and Asia, 1700–1900*. Cambridge: MIT Press.
- Elder, G. J., Kirkpatrick, J. M., & Crosnoe, R. (2003). The emergence and development of life course theory. In J. T. Mortimer & M. J. Shanahan, (Eds.), *Handbook of the Life course* (pp. 3–19). New York: Plenum.
- Fauve-Chamoux, A., & Wall, R. (1998). Nuptialité et famille. In J. P. Bardet & J. Dupâquier (Eds.), *Histoire des populations de l'Europe* (pp. 344–368). Paris: Fayard.
- Fauve-Chamoux, A., & Ochiai, E. (Eds.). (2009). *The stem family in Eurasian perspective: Revisiting house societies, 17th–20th centuries*. Bern: Peter Lang.
- Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2009). *Mining sequence data in R with the TraMineR package: A user's guide*. Geneva: Department of Econometrics and Laboratory of Demography, University of Geneva.
- Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Hajnal, J. (1965). European marriage patterns in historical perspective. In D. Glass & D. E. C. Eversley (Eds.), *Population in history* (pp. 101–146). Chicago: Aldine.
- Hajnal, J. (1983). Two kinds of preindustrial household formation systems. In R. Wall, J. Robin, & P. Laslett (Eds.), *Family forms in historic Europe* (pp. 65–104). Cambridge: Cambridge University Press.

- Hareven, T., & Masaoka, K. (1988). Turning points and transitions: Perceptions of the life course. *Journal of Family History*, 13, 271–289.
- Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data*. Hoboken: Wiley.
- Kok, J. (2007). Principles and prospects of the life course paradigm. *Annales de Démographie historique*, 1, 203–230.
- Malthus, T. R. (1992). *An essay on the principles of population*. Cambridge: Cambridge University Press (Original work published 1803)
- Neven, M. (2002). Espaces ruraux et urbains au XIXe siècle: Trois régimes démographiques belges au cœur de la révolution industrielle. *Popolazione e storia*, 2, 35–62.
- Neven, M. (2003). *Individus et familles: Les dynamiques d'une société rurale. Le Pays de Herve dans la seconde moitié du 19e siècle*. Genève: Droz.
- Oris, M., & Ochiai, E. (2002). Family crisis in the context of different family systems: Frame-works and evidence on “When Dad died”. In R. Derosas & M. Oris (Eds.), *When dad died. Individuals and families coping with distress in past societies* (pp. 17–79). Bern: Peter Lang.
- Oris, M., Alter, G., & Servais, P. (2014). Prudence as obstinate resistance to pressure. Marriage in nineteenth century East Belgium. In S. Kurosu, C. Lundh et al. (Eds.), *Similarity in difference: Marriage in Europe and Asia, 1700–1900*. Cambridge: MIT Press.
- Ritschard, G., & Oris, M. (2005). Life course data in demography and social sciences: Statistical and data-mining approaches. In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, & E. Widmer (Eds.), *Towards an interdisciplinary perspective on the life course* (pp. 283–314). Amsterdam: Elsevier.
- Ritschard, G., Gabadinho, A., Studer, M., & Müller, N. S. (2009). Converting between various sequence representations. In Z. Ras & A. Dardzinska (Eds.), *Advances in data management* (pp. 155–175). Berlin: Springer-Verlag.
- Servais, P., & Alter, G. (Eds.). (2005). *Le mariage dans l'Est de la Wallonie XVIIIe-XIXe siècles*. Louvain-la-Neuve: Bruylants-Academia.
- Shanahan, M. (2000). Pathways to adulthood in changing societies: Variability and mechanisms in life course perspective. *Annual Review of Sociology*, 26, 667–692.
- Studer, M. (2013). Weighted Cluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. LIVES Working Paper 24. Switzerland: NCCR LIVES.
- Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research*, 40(3), 471–510.
- Todd, E. (1990). *L'invention de l'Europe*. Paris: Seuil.

Part III

Political Sequences

Chapter 9

Trajectories of the Persecuted During the Second World War: Contribution to a Microhistory of the Holocaust

Pierre Mercklé and Claire Zalc

Introduction

The mobilization of prosopographic methods remains relatively uncommon in historical work on the Holocaust, especially in France (Anders and Dubrovskis 2003). This means that the data on which the present analysis is based unquestionably have an exemplary character: a long-term investigation (Mariot and Zalc 2010), in fact, had made it possible to reconstitute the “trajectories of persecution” of the one thousand Jews living in Lens in 1939, a mining town in the north of France with a population of around thirty thousand before the Second World War.

In order to understand the specific character of the history of the Jewish community in Lens during the War, we have first to consider its tragic conclusion: the Jews of Lens took a still greater toll from Nazi persecution than others elsewhere, since out of the 991 individuals that made up the community before the War, 478 were arrested and 467 of these deported, of whom only 18 returned from the extermination camps. In total, only 528 of the 991 Lens Jews survived the War. From this point of view, Lens is not representative: over half the Jews living there in 1939 were deported, whereas the proportion for Jews present in France as a whole is estimated at around one quarter (Klarsfeld 2012).

The question is to explain why Lens is not representative and to understand the incredible harshness of the persecution there. The aim of our work has been to approach the dilemmas faced by Jews in Lens not primarily as psychological phenomena, but rather as choices dependent not only on the particular contexts in which they were made, but also on the social and demographic characteristics of those who made them: occupation, family configuration, and structure of group affiliations. The aim of this study, therefore, is in a sense to model persecution.

P. Mercklé (✉)
Centre Max Weber, ENS de Lyon, CNRS, Lyon, France
e-mail: pierre.merckle@ens-lyon.fr

C. Zalc
Institut d'histoire moderne et contemporaine, ENS Ulm, CNRS, Paris, France
e-mail: claire.zalc@ens.fr

This is why we chose a quantitative approach to explain circumstances that are usually associated with the singularity of suffering. Some other authors have done this, especially in Netherlands (Gross 1994; Croes 2006; Tammes 2007), in order to explain differences in survival rates. Yet, their attempts at quantitatively modelling persecution rely on what Abbott calls the “standard programme” of sociology (Abbott 2001): their approaches are mainly based on regression models, with a few incursions into event history analysis (Tammes 2007).

The first part of this chapter will review the difficulties encountered in mobilizing such tools: is it appropriate to reduce choices made under tragic circumstances to their social determinants? How accurate and realistic is our account of persecution if it reduces its causes to age, income level or family size? The benefits of quantification must not conceal the problems raised by the linear patterns of causality assumed by the techniques we initially tested (correspondence analysis, logistic regression). To conceptualize our data as “trajectories of persecution” seems to offer an interesting prospect for overcoming some of these difficulties. In the second part of the chapter, we will therefore describe how we translated our database into a corpus of “sequences”. Our aim was to formalize successions of time sequences into trajectories of individuals confronted with persecution, in order to identify classes of trajectories, patterns and trends. By moving from a logic of properties to a logic of sequential states, and from a logic of causes to a logic of paths, we describe, order and interpret the plurality of trajectories without abandoning quantification. While some previous difficulties may be resolved, new ones may emerge due to certain properties of our sources (missing data, problems with formalizing events into states...), which we also discuss. In the last part of the chapter, we indicate the benefits that may be drawn from approaches mobilizing optimal matching analysis, in order to discuss the contribution of these modes of quantification to a better understanding of interactions, at a local level, between victims and persecutors, and in order to address the possible contribution of an approach in terms of trajectories to a microhistory of the Holocaust.

A Study That Changed Shape

The Original Prosopographic Approach

The study on which this analysis is based had its origin in the desire to reconstitute the individual biographical trajectories of all Jews living in Lens at the start of the Second World War. It was supported by the patient gathering of the greatest possible amount of materials and documents that enabled these trajectories to be described. How did we do this? There is a well-known book by Daniel Mendelsohn, entitled *The Lost: A Search for Six of Six Million* (Mendelsohn 2006). He attempts to rescue a single family’s story from oblivion by digging into the details of their lives. Our survey had a similar ambition, but multiplied by 300. It is the story of 300 families: 991 people. Our work thus consisted in tracking one thousand people based on a wide range of

sources, from local and French national archives (such as “Aryanization” files, or naturalization files) through to files documenting deportation from France and Belgium, archives of the United States Holocaust Memorial Museum (USHMM), Auschwitz archives, Swiss refugee records, Yad Vashem testimonies, etc. To reconstitute the individual trajectories of these one thousand Jews through the War, we tracked their names through an average of 10 to 15 different sources per person, recorded in different places and contexts, on different dates and by different persons.

There are certain obvious patterns that structure these trajectories: the Jews living in Lens in 1939 were mainly immigrants who had arrived from Eastern Europe in the 1920s. This mining region of northern France was more generally a land of Polish immigration, and most of these Jews specialized in textiles and sold their goods to Polish miners. They did so within a relatively clearly bounded “interethnic commerce” because they spoke the same language: Polish. More than 80 percent of the Lens Jews lived in the city centre, and all of them within 1,300 m of the train station. Yet, downtown Lens cannot be considered as a “ghetto”, since Jews only accounted for 3 % of the whole city population, so that most lived in highly mixed neighbourhoods.

At the outbreak of the Second World War, a significant number of these Jewish immigrants volunteered for service in the French Army. Joining the general exodus from northern France at the time of the German invasion in May 1940, around 40 % of the Jews of Lens left the area at that time and never returned, many of them settling in southern France. We followed these individuals throughout their subsequent lives and found out that leaving Lens in 1940 did not ensure permanent safety: some of them would unfortunately be trapped in later roundups and deportations elsewhere in France.

As for the Jews who stayed in Lens, or returned by late summer 1940, they were soon confronted with an ever-growing list of discriminatory measures. The chronology of the War itself is somewhat specific in Lens since the town was part of the “zone interdite” (forbidden zone), annexed to Belgium by the Germans under the terms of the Armistice. But it differs relatively little from the well-known chronology of the “zone occupée” (occupied zone): autumn 1940 saw the promulgation of the first statute on Jews, the census of Jews in the northern zone, the first Jews excluded from certain professions, the Aryanization of companies, and the internment of some foreign Jews; June and July 1941 saw the second statute on Jews, the extension of quotas and expulsions from the professions, and a new census; spring and summer 1942 saw the implementation of curfews and the requirement that Jews wear the star of David. It saw frequent roundups, and the handing of Jews over to the Germans. The authorities identified them, isolated them, reduced them to misery, and ultimately arrested and sent half of them to their deaths at Auschwitz-Birkenau. The massive arrests and deportations ran from summer 1942 to 1944 and hit the Lens Jewish community with full force.

Yet, even given this overall pattern, the patient study of individual trajectories through persecution revealed no uniform and mechanical logic. On the contrary, when we followed each one step by step, with the purpose of understanding the concrete lives of individuals facing persecution, it appeared that the town’s Jews faced

several options, which we tried to retrace: when confronted with the German invasion in spring 1940, some sold their businesses and embarked on a new exodus, while others stayed and continued their pre-War occupations; among those who left, some returned to Lens (now located in a “forbidden zone” for refugees); when confronted with the successive censuses of Jews, some declared themselves to the authorities and some did not, etc. These questions, with which each of the 991 Jews living in Lens in 1939 was confronted, were nothing less than a matter of life or death.

In order to understand how we managed to follow individuals through their entire life courses, it is necessary to take a closer look at the archives. The originality of the present work is that it relies on historical sources that were not at all intended to serve our purposes: our data were not drawn from survey questionnaires. Historical data are, by definition, not conceived for any subsequent quantitative analysis. On one hand, this could be seen as a problem. But on the other hand, that is why they are so rich, and so helpful in understanding the relationship between executioners and victims in the Holocaust process. That is the main reason why we favoured an archival approach to the trajectories, against the dominant trend of revering the direct testimonies of Holocaust victims which, likewise by definition, are driven by a will to tell a story of survival (Pollak 1990; Browning 2010). To multiply sources means to multiply points of view, in terms of exogenous, endogenous and action variables. Among sources that provide information about persecution (what we called “exogenous variables”), censuses are precious, since they allow us to determine whether the Jews of Lens were identified and located in Lens at these specific moments, or not: December 1940, April 1941, January 1942; likewise the list of stars of David given to Jews (August 1942). Even more crucial are lists of names on deportation convoys, since they allow us to determine deportation and internment trajectories. A particularly hard part of the research involved following these Jews during the deportation process. Thanks to the opening of new archives—especially the Bad Arolsen records digitalized by the United States Holocaust Historical Museum three years ago—it became possible to describe what happened to Jews deported from Lens from their arrival at Auschwitz until their death there. Of the 467 deported Lens Jews, 108 were registered and held at Auschwitz. That means that more than 75% went immediately to the gas chambers. And of those 108 Jews from Lens who were registered and held at Auschwitz, 84% died. Among those whose date of death is known, a quarter died in the first month and 85% within four months, while only 7% survived longer than a year. At that point in our research, we no longer searched for names in the archives, but for numbers: the serial numbers that were given to Jews when entering a camp, and tattooed on their forearms.

While they provide information on exogenous constraints on their trajectories, sources such as Aryanization files or identification censuses may also provide more or less accurate information on the social properties of Lens Jews, such as gender, occupation, address and so forth (what we called “endogenous” variables). And in addition to exogenous and endogenous variables, elements of individual trajectories and actions may be learnt from two kinds of specific source: first, the refugee files of Jews who fled to Switzerland, and second the naturalization files that were compiled after the War.

From the time that we gave ourselves the objective of restoring to observable trajectories and decisions their social density, quantification offered several possibilities. It was not just a matter of counting how many individuals were despoiled, hidden or deported, but also who they were and in what way they were distinguished (or not) from those who were not. Quantification thus makes it possible to break with an individual logic, a considerable advantage in dealing with controversial questions that are also issues of memory. To start from individual trajectories involves the risk of only preserving the most “exemplary” cases, those most “outside the norm”, or those which left the strongest or most palpable traces. In a perspective that is not, we should make clear, in any way hostile to or exclusive of a more qualitative approach, we have sought to define individual characteristics with a view to understanding the possible determinants of their trajectories. But how should this data be analysed?

Dead Ends of the ‘Standard Programme’

Conducting a prosopographic work on the victims of persecution, in fact, raises questions of both a practical and an epistemological order (Mariot and Zalc 2012). We came up against the difficulties inherent in the analysis of the social that Abbott terms the “standard programme” of American sociology. The criticisms addressed by Abbott to the schema of linear causality go together with a subtle deconstruction of the inferred role of “variables” in sociological analysis. He shows first of all the difficulty of producing stable explanatory variables of the social, given that the reality observed is eminently mobile. This is what he calls the “temporal horizon” (Abbott 2001), recalling how often the same variables are used, depending on researcher and context, to equate behaviours that are sometimes completely different. We can even press this criticism somewhat further: on one and the same terrain, variables can work differently at different times. Following the various stages of the persecution that the 991 Jews of Lens were subjected to invites us to reflect on the explanatory weight attributed to the “variables” initially applied in describing and explaining behaviours. In fact, the weight of certain of these variables may fluctuate greatly according to the particular moment. Take socio-economic status, for example. This variable provides a potential indicator of factors of survival, in the sense that it denotes a space of resources that can be mobilized in the face of persecution: wealth (Croes 2006). This is one of the hypotheses that we wanted to test in the early stages of our research: did the rich emerge better than the poor? But there is nothing self-evident about this question. The effect of socio-economic status, in fact, can be noted at certain moments in the persecution, but not at others. Even if more weakly than age, and especially household size, socio-economic status did play a role in the fact of leaving or remaining in Lens: 62% of independent professionals left, against only 53% of wage-earners. At the time of arrest, on the other hand, socio-economic status no longer played any significant role: more or less equal proportions of professionals, wage-earners and non-employed (around 45%) escaped arrest. The same variable, therefore, played a different role at different stages of persecution.

When we focus on the role of nationality, we find a still more significant yet reversed type of interaction. Nationality played only a small role in the decision whether to remain in or leave Lens, but it remained a very strong predictor of arrest, despite the fact that French citizens were in theory foreigners like any other in the “forbidden zone”—this region being, we recall, attached to the German command in Brussels, and thus treated by the occupants as a zone annexed to Belgium. However, “only” 36% of Lens Jews of French nationality were arrested, as against 59% of Poles and 63% of other nationalities (Romanian, Czech, Russian, etc.). Finally, if we now examine not just the act of leaving, but the date of departure from Lens, the “nationality” variable is once again strongly discriminating: an average of 20% of Lens Jews left in 1942, but these were made up of 41% of French citizens as against 19% of Poles and 8% of other nationalities. Undoubtedly, as we have hypothesized, the former continued for longer to feel relatively protected by their French nationality.

Our study of Lens, moreover, shows that from one day to the next, even the modalities of the variables that characterize individuals can change. To take the example of wealth, if it is possible to establish a socio-economic categorization of Lens Jews in autumn 1939, thanks to the Aryanization files and declarations of occupation that figure in the different censuses carried out throughout the occupation period, the measures of confiscation of goods on the one hand, and professional bans on the other, rendered the majority of these socio-economic classifications null and void by the end of 1940. In this sense, socio-economic characterization changed over time, and this variable cannot be treated as a static attribute that was stable throughout the period under consideration. Quite the contrary: the impossibility of reconverting economic capital in the Vichy context, or even the disappearance of all means of subsistence, enables us to understand a certain number of behaviours: for example, how forced registration in a *Groupement des Travailleurs Étrangers* (GTE) could represent an escape route for certain individuals.

As we see, the question that vexes the analyst (“What determined that some people survived and others did not?”) can only be resolved with difficulty via an approach of the “linear causal” type, to use Abbott’s term. His critique of the standard methods involves a second point that proves pertinent for describing the difficulties encountered in attempts to explain the itineraries of the Lens Jews in the face of persecution: the difficulty of isolating an indicator or ascribing it a particular status in explanations (Abbott 2001). For example, in a correspondence analysis, how to distinguish the “active” variables from the “illustrative” ones, when possible “outcomes” of persecution (such as identification, or departure from Lens) may become explanatory factors of further persecution, or of escape from it?

Trajectories are indeed a product of the interaction of three types of factors: exogenous variables (being identified, aryanized, arrested), which depended on the context and on the authorities applying the anti-Semitic policy; endogenous variables, which could again be qualified as “individual properties” and which, as we saw with the example of socio-economic status, could develop in a matter of days or months (young/old, single/married, French/Polish, rich/poor); action variables that describe the behaviours of individuals (declaring oneself, leaving, remaining,

hiding, crossing into Switzerland, etc.). Yet, the handling of the causal connections between these three groupings is by no means a trivial matter. It is theoretically and empirically impossible to distinguish causes from effects, and the logics of trajectories are a function of the complex articulation or combination through time of the three sets of variables.

Another difficulty inherent in the “standard model” is that it proceeds from the starting point and as a function of an objective that aims to reveal the specific effect of variables in a causal logic. The problem makes itself felt here in a particularly sensitive way: in a certain sense, proposing a logistic regression that bears on arrest involves seeking to disassociate the supposed effects of different variables on the fact of being arrested or not. But this amounts to obscuring, if not misconstruing, other events that are determinant prior to the moment of arrest. For example, leaving Lens or remaining.

We are faced here with the third pitfall in causal explanations, as described and considered by Abbott: applying an explanation, by way of logistic regression, that seeks to cast light on effects of causality, comes up against the specifically arbitrary character of the persecution policy. The failure of socio-economic status to have any effect on arrest, as we have demonstrated above, already shows this arbitrary character. For summer 1942, does it make sense to decompose variables? Were there unlikely situations whose possibility should nevertheless be examined? Or again, to take the famous quotation that Maurice Halbwachs attributes to François Simiand, is this method adapted to our study, when it “consists in studying and comparing the behaviour of a reindeer in the Sahara with a camel at the North Pole” (Desrosières 2001)? There are a number of hierarchies that underlie the models of quantification that attribute fixed and uniform causal connections between variables. The life of the Lens Jews, however, sometimes seems to hang on a single thread: the boy William Sharfman was rounded up in Lens with his mother on 11 September 1942, but saved by a railway-worker on the station platform. This chance determined his survival. But how would we seek to explain it? This shows how the process of persecution cannot be, from our point of view, analyzed simply as a final result (survival or not) but should be understood as a trajectory. These aspects and difficulties of univocal interpretation contributed to inflecting our study by attempting other ways of reading the data, formulating ideas and modelling. Our goal thus became that of working to formalize the different “trajectories of persecution” in terms of a “quantitative approach”.

Trajectories of Persecution: From Archives to Sequences

Biographies to Trajectories

From this point on, the formatting of the prosopographic data gathered on these 991 Jews had to be changed. In the first phase of research, the formatting was conducted with a database, the so-called “individuals base”, which, in a big spreadsheet, com-

piled the different bodies of sources used and plotted individuals (in rows) against information given in these sources (in columns): date and place of birth, household composition, occupation, address, nationality, sometimes date of arrival in France. In the course of compiling this database, however, several of the difficulties mentioned above became evident: how could one put in a single box marked “address” the several addresses that appeared in naturalization files after the War, tracing the trajectories of flight? But if this logic right away appeared ill adapted, the fact is that the majority of quantitative treatments used databases formatted according to a sociological questionnaire supplied to those studied (Lemercier and Zalc 2008). Historical data, by their very nature, escape such simple “completion” of questionnaires. But is the historian, particularly when quantifying, not tempted to offer his or her “subjects” a retrospective questionnaire with closed questions, at the price of forgetting the connection with the sources and filling in the boxes, at any cost, to avoid “empty spaces”, ignoring that the sources are very often patchy? As we followed our sources, the “individuals” database became steadily transformed into a “trajectories” database. Adoption of a longitudinal logic that marked the different steps in the trajectories of persecution finally succeeded in most closely matching the specificities of our sources. From information segmented on the principle of individuals faced with persecution, for whom we sought to determine the factors favouring survival, we managed to break down this information into the form of “trajectories of persecution”. Without for all that abandoning quantification: here again, as shown by the totality of work on occupational careers, it is possible to take into account the totality of sequences observed in order to reveal a fine typology of itineraries. This is why an analysis of trajectories of the “optimal matching analysis” type imposes itself as a solution adapted to our terrain (Abbott 1990).

In order to understand how data taken from different sources are articulated together, examination of individual cases proves particularly enlightening. Take for example Charles Dembinski, whose naturalization file enables us to determine both his place and date of birth, then the whole of his residential trajectory until the end of the War. But the police report included in his naturalization file does not mention that he was arrested on 10 September 1942, that he escaped and joined his wife in Périgueux, and that from this point until his return to Bully-les-Mines he was therefore living a clandestine existence. This information is essential, as it is what makes it possible to determine the succession of “states” that make up the “trajectory of persecution” of Charles Dembinski (Table 9.1).

In the language of sequence analysis, this trajectory is finally formalized in the form of an ordered succession of particular states or “spells” of an individual relationship to persecution, each of these spells being occupied for a certain duration. Thus, the biographic trajectory of Charles Dembinski, as this was finally coded in SPS format by using R (R Core Team 2013) and the TraMineR package software for data analysis (Gabadinho et al. 2011), takes the following form: (F,492)-(R,1)-(F,12)-(R,8)-(N,1)-(C,24). In other words, starting from birth, Charles Dembinski was free during 492 months, then identified during 1 month, then free during 12 months, then identified during 8 months, then interned during 1 month, and eventually underground during 24 months, i.e. until the end of the War.

Table 9.1 The “trajectory of persecution” of Charles Dembinski

Spell	Place	State	Start	End
1	Kowal (Pol.)	Birth	4 Dec. 1899	1922
2	Angevillers	Free	1922	1924
3	Lens	Free	1924	Oct. 1931
4	Vimy	Free	Oct. 1931	Dec. 1931
5	Lens	Free	Dec. 1931	1934
6	Saint-Berain	Free	1934	21 Dec. 1936
7	Bully-les-Mines	Free	21 Dec. 1936	Feb. 1941
8	Paris	Free		16 June 941
9	Bully-les-Mines	Free	16 June 1941	10 Sept. 1942
10	Malines	Interned	10 Sept. 1942	Oct. 1942
11	Périgueux	underground	Oct. 1942	
12	?	underground		
13	Bully-les-Mines	Free	30 Oct. 1944	

Trajectories of Persecution

The body of information that the analysis draws on compiles 991 individual trajectories of Jews from Lens, each of these trajectories being itself defined as a succession of distinct moments or “spells”, characterized by a start date, an end date, an individual “state” and a particular place of residence. These 991 sequences are thus made up of a total of 5,875 distinct spells, that is, an average of almost exactly six spells for each sequence. The shortest sequences include just one spell (birth), and the longest, that of Soria Salik (*née Schor*), has sixteen, from her birth in Poland in 1903 to her death in Lens in 2002. Between these two extremes, the distribution of the number of spells per sequence has a very clear mode, since almost half (47%) of the trajectories contain five or six spells (including birth and death). Each of these spells distinguishes a different “status” of relations in which individuals found themselves in the face of persecution (Table 9.2).

The trajectories that we thus try to describe are indeed “trajectories in the face of persecution”: from freedom and census registration through to internment and deportation, the spells that make them up are always defined by the relationship to persecution. That said, while these “states” were certainly determined by the factual situations in which the Jews of Lens found themselves in the face of persecution at each moment in their biographical trajectories, they were equally determined by the manner in which we managed to gather the information that has enabled us to determine these situations. The “Identified” status is emblematic of this ambiguity: it denotes both the source of information (the census lists of Jews drawn up by the Lens authorities during the conflict), the fact that the individuals who were in this “situation” lived in Lens, and the fact that they were not exactly in the same situation as those who escaped the census (and are thus defined as “Free”). They were subject to the constant work of identification and censuses carried out by both French and German authorities, particularly between December 1940 and September 1942, and

Table 9.2 States of individuals in the face of persecution

Status	N	%	Mean dura- tion (months)	Status	N	%	Mean duration (months)
Birth	991	16.9	—	Underground	37	0.6	24
Free	2638	44.9	189	Interned	495	8.4	2
Army	35	0.6	9	Deported	475	8.1	3
Prisoner	13	0.2	23	Refugee	53	0.9	26
Identified	594	10.1	15	Deceased	528	9.0	—
Assigned	16	0.3	6	<i>Total</i>	5875	100.0	

it is because of the traces left by this administrative and police work that we are familiar with these parts of their trajectories... In some cases, it is in fact completely impossible to separate what bears on the “real” trajectory from what is a function of the archival artefact, even when the intersection of different sources indicates contradictions: in a non-negligible number of cases, in fact, censuses continue to identify individuals whom we know from other sources were no longer present in Lens. This is the case with Markus Adlerfligel, who appeared on the Lens census in December 1940 though he was in fact a prisoner-of-war in Germany throughout the War, from May 1940 to January 1945. Similarly, the naturalization file for Baruch Stolik, despite his appearing regularly on the Lens censuses until January 1942, indicates that he was actually in Lyon between 1940 and 1945: he was certainly “Free”, but he was also “Identified”, which corresponds to a very real situation in the face of persecution, if only in so far as it can precisely explain his departure from Lens. We must however be well aware that in the present case, appearing under “Identified” actually meant being “registered in Lens”, inasmuch as the reconstitution of trajectories is based on archive material that is initially local. This constitutes a possible bias in the data from which trajectories can be reconstituted, inasmuch as an individual who was not “Identified” in Lens is described in our material as “Free” during this period, whether or not he was “Identified” elsewhere in France.

Trajectories: Defining and Dating

In theory, each spell is defined by a start date and an end date. But in practice, we were not able to reconstitute each trajectory with full precision as an exhaustive series of perfectly time-delimited spells. At the start, less than half (46.5%) of the spells were bonded by both a start date and an end date. The initial data, moreover, as gathered from primary sources, already incorporated a share of hypothetical extrapolation, at least each time we extrapolated continuously enduring states from discrete events. For instance, if we knew a date of birth in Lens and then only a date of departure from Lens, even several years after, we supposed a continuous

residence in Lens between these two dates, instead of considering these spells as missing.

Yet, despite this first way of imputing dates to incomplete spells, we were still confronted with a significant proportion of missing dates. Also, to the extent that TraMineR only accepts spells that are bounded by both a start and an end date, we applied certain simple procedures of imputation. First of all, the end date of the final known moment of each trajectory has been set by convention as 1 January 2012. Despite this, for a very large portion of other spells (42.4% in the raw data, and still 27.7% after imputing end dates to the final spell), only the start date is known, and not the end date. This is quite simply explained by the fact that entry into a given state generally corresponds to an event whose date is known, and which therefore defines its start. The end of this moment, on the other hand, is far less frequently obtainable. For the same reasons, it is also the case with deportation, for which we know the start date, particularly from the lists of convoys, but generally do not know the end date, except in the very rare cases where the date of decease in Auschwitz is known. Faced with these many uncertainties, we imputed to the incomplete spells a start date equal to the end date of the preceding sequence, and an end date equal to that of the start of the following one, if these were known.

There is still a final debatable effect of this simple procedure of imputation of dates: in certain cases, when a moment without a start or an end date is interpolated between a previous moment with an end date and a following moment with a start date, this interpolated moment is then imputed both a start and an end date, which amounts to considering that it actually is adjacent to the two others and occupies the whole interval of time that separates them. In the case of Nelly Hornstein, for example, we know only that she was in Lille during the War, but not for how long or under what status. But, as the end date of the previous moment and the start date of the following moment are known, the imputation procedure here has the consequence of supposing that the moment of residence in Lille (which originally had neither a start nor an end date) covers the whole period of the War, from May 1940 to September 1944.

Despite these few restrictions, the imputation procedure thus applied does have significant effects on the whole, since it makes it possible to end up with a total of 83.6% of complete states instead of only 61.0% at the start. The final procedure then consists in imputing a conventional duration of one month to spells for which only one of the two time boundaries is known. It is thereby supposed that for the rest of the previous or following time, the state is not determined. This final procedure has two effects: on the one hand, it makes it possible to complete almost the totality of spells (98.5% of spells are now bounded by both a start and an end date), but on the other hand it generates a non-negligible proportion of “gaps” in the sequences. We thus find several “holes” of this type at the start of sequences with individuals for whom we know only the date of birth in Poland and then the date of departure from Lens. Between the month following birth and the month preceding departure from Lens, for example in May 1940, the status is not determinable.

From Missing Values to “Biographical Gaps”

The most evident property of our data is that they are thus characterized by a quite unusual proportion of missing values. As in any body of longitudinal data, these are of three types, and it is necessary therefore to distinguish between unknown states prior to the first known state (left-censored), unknown states posterior to the last known state (right-censored), and unknown states located between these limits, within trajectories, otherwise known as “gaps”. In the majority of cases, the missing values on the left precede birth, and could thus be treated as void values; in the same way, values missing on the right correspond in the majority of cases to death, generally in deportation but at an unknown date, which means that deportation remains the last known state, for which we originally know only the start but not the end. For these first two types of missing values, we could have applied the relatively customary strategy which consists in treating them as void values and removes them from the analysis. But as far as the “deceased” states and missing values on the right are concerned, that would have amounted to not taking death and date of death into the analysis, whereas in the perspective of a study of the persecution and extermination of Jews this is obviously a fundamental datum. We consequently decided to consider these missing values as positive states, which makes it possible on the one hand to take account of date of birth and thus of the age of individuals, and on the other hand of the moment of death or disappearance.

There remains the third type of missing values, in other words “gaps”. For some Jews from Lens, biographical information, as gathered from archival sources, is considerably more brief and patchy than for others. This is one of the problems that the formatting of data into sequences confronted us with: we do not have equal information on trajectories for all individuals. The remaining biographical gaps are substantial in quantitative terms: disappearance (missing states after the last known states) affects some 15% of individuals from the end of 1942 on, and biographical gaps (missing values) steadily rise during the whole wartime period, reaching practically 20% by mid-1940 (Fig. 9.1). Before the work of identification and census registration of Jews in Lens began in December 1940, we no longer know where over one third of these were. The proportion of gaps experiences a first significant dip from December 1940 on, with the lists regularly drawn up by the authorities enabling us to locate one part of them in Lens; and the second, more spectacular fall, corresponds to the two roundups of 11 and 25 September 1942.

Standardizing the Recording of Dates

After imputation, we thus have available a database made up of sequences that include only complete spells (with both a start and an end date) and possible biographical “gaps”, that is, spells determined both by a start and an end date, but corresponding to states and places of residence that are unknown. This said, the dates that border the known spells are of variable precision. We have chosen to standardize

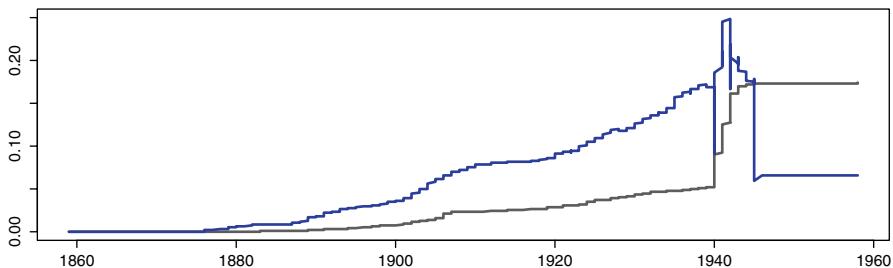


Fig. 9.1 Disappearances (*plotted in grey*) and missing values, or gaps (*plotted in blue*)

the precision of all dates to the month, by applying the following simple rule: the dates that are known only to the year have been rounded to the mid year, that is, to the month of July in the year in question.

This rule of standardization of dates does have consequences, however. Thus, for a certain number of very important spells, the start date and the end date are known to the day, but these two dates fall in the same month. This is the case for rather more than 700 spells out of the 5,875 in total. These spells are particularly the internments in Malines between 11 and 15 September 1942, after the big roundup in Lens, and the ensuing deportations to Auschwitz, between 15 September 1942 and a date of death in the camp that was only a few days later. For all the sequences concerned, we resorted to a small artifice of coding “Free” until August 1942, “Interned” in September 1942, “Deported” in October 1942, and “Deceased” if appropriate from November 1942. This artifice offers the advantage of not obliterating the trajectories of internments lasting less than a month, even if it produces a certain number of distortions in relation to the starting data, in particular an overestimation of the duration of internment and deportation, and a shift of a month in the dates of deportation and decease in a certain number of cases.

Results

What is a “Normal” Life-Course Sequence?

One of the most evident contributions of sequence analysis lies in its ability to offer an extremely synthetic view of status transformations over time. In the histogram below (Fig. 9.2), where time is represented on the horizontal axis, each vertical line shows the distribution of Lens Jews between possible states for each given month between 1859 and today. A glance at the periods preceding and following the 1940s reveals that War appears as a perturbation in a histogram of states distribution that otherwise reflects what life courses usually are: as time goes by, people enter our sample, thus leaving their initial “Unborn” state (this decline being figured in grey at the bottom left of the histogram) and experiencing a lifelong “Free” state until they either eventually die, or enter “Unknown” or “Lost” states in case we lose

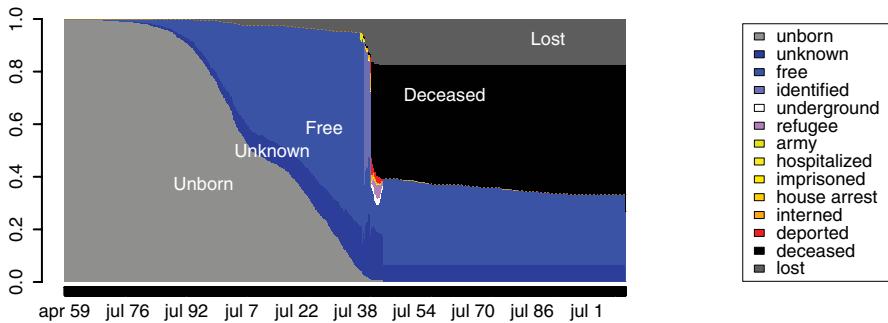


Fig. 9.2 Distribution of states of the Lens Jews between 1859 and 2012

their tracks. That is what we can observe before July 1939: the proportion of “Unborn” individuals steadily declines over time, whereas the proportions of “Free” and “Unknown” states grow, until these eventually represent almost the totality of the population.

After 1945, states are also distributed in a very stable fashion between alive (“Unknown” and “Free”) and dead (“Deceased” and “Lost”) states. The increase in the proportion of dead is very slow, in fact almost imperceptible, over a 5-year period of time after the War. The very stable “Lost” state corresponds to people who most probably died in the camps, but whose date of death remains unknown and thus whose last known state is “Deported”.

What sequence analysts call “entropy” may represent an interesting measure of this pre- and post-War stability against wartime perturbation. It tends to 0 when all cases are in the same unique state and it is maximal when the same proportion of cases is in each different state. Entropy can thus be seen as a measure of the diversity of states observed at the considered time point (Fussell 2005). Before the War, entropy in the present case slowly declined as individuals left the “unborn” state. The War caused entropy to increase in steep steps, among which the most noticeable occurred in December 1940 when the authorities began their systematic enterprise of identifying the Lens Jews, and in September 1942 with the Lens roundups. After the roundups, entropy was at its highest, as significant numbers of people entered new states such as “Underground” and “Refugee” (those escaping to Switzerland). Afterwards it decreased slowly until the end of the War, as the population tended to stabilize in four remaining states (“Unknown”, “Free”, “Deceased”, and “Lost”).

Focus on the War

When we then focus on the War period, it appears that the most dramatic change in the state distribution clearly occurs in September 1942, the month of the great Lens roundup (Fig. 9.3).

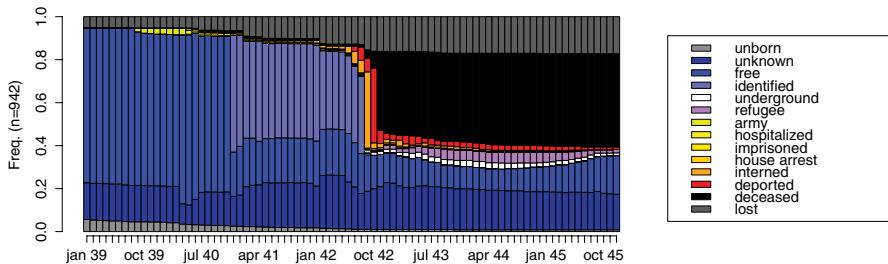


Fig. 9.3 Distribution of States 1939–1945

Prior to the roundup, more than 80% of Lens Jews were free; after September 1942, this was only the case with one in five, and between September and October 1942, half of the Lens Jews were exterminated. As far as they were concerned, the local application of the Final Solution had a massive and almost instantaneous effect. In August 1942, 34% of Lens Jews were “identified” by the authorities, which means that they were asked at the beginning of the month to retrieve their yellow stars. In September, they were arrested and interned, most of them in Malines. Almost all of them were then deported to Auschwitz in October and deceased by November¹ (Table 9.3).

The ineluctability of the process appears even clearer when we look at transition rates between states. While other states at other times are remarkably stable, transition rates between August and November 1942 describe an almost fatal process: for those who were identified in Lens in one of those four months, the chance of being interned the following month was over 80%; for those interned, the chance of being deported the following month was 86%; and for those deported, the chance was 75% that they were dead the following month.

Classifying and Explaining Trajectories

Observing the individual biographies of each of the 991 Jews of Lens may give the impression of an infinite diversity. Yet, modelling these biographies into sequences allows to uncover underlying forces that tie many of these biographies together. It indeed turns out that the ten most frequent sequences, i.e. less than 10% of the observed different sequences, make up almost half of the individual trajectories through the War. On its own, the most frequent sequence (“Free” until August 1942, “Interned” in September, “Deported” in October and “Deceased” by November 1942) represents 18.7% of the trajectories of Lens Jews through the War.

¹ Here we must recall that this mass extermination of Lens Jews actually happened in a much shorter period of time, between 11 September and the end of the month, and that the extension of the process to October is an artifact due to adopting monthly intervals as units for sequence analysis.

Table 9.3 State distribution table and transition rates between August and November 1942

	State distribution				Transition rates				
	Aug-42 (%)	Sept-42 (%)	Oct-42 (%)	Nov-42 (%)	Identified	Interned	Deported	Deceased	
Identified	34	0	0	0	Identified	0.01	0.81	0.01	0.00
Interned	6	36	3	2	Interned	0.00	0.13	0.86	0.00
Deported	6	6	35	6	Deported	0.00	0.00	0.19	0.75
Deceased	1	4	8	37	Deceased	0.00	0.00	0.00	1.00
Other	53	54	54	55					
Total	100	100	100	100					

The existence of patterns of sequences authorizes recourse to classification procedures that help distinguish their principal types. In that perspective, we mobilized optimal matching methods to separate War trajectories (i.e., between 1939 and 1945) into clusters grouping those who most resembled each other, with substitution costs calculated from transition rates between states (Lesnard 2010; Macindoe and Abbott 2010). The optimal matching procedure separated five different clusters (Table 9.4): two clusters (1 and 5) of trajectories leading to extermination, with cluster 1 grouping trajectories where extermination is preceded by identification, and cluster 5 grouping trajectories where it is preceded by freedom; two clusters (2 and 3) of survival trajectories, with cluster 2 containing a significant amount of gaps (unknown states), and cluster 3 grouping survival trajectories that implied going underground at some point; and eventually one cluster (4) grouping right-censored trajectories, i.e. trajectories that sooner or later end up with losing track of individuals.

Analysis of the relations between types of trajectory and the socio-demographic characteristics of individuals reveals a certain number of particularities. Thus, members of large families (more than four persons) had a much higher risk of undergoing a trajectory of extermination, but not just any of these: those (cluster 1) in which extermination was preceded by a long period of identification and surveillance by the authorities. Close to half (46.3%) of members of families of six persons or more experienced this type of trajectory, as against less than a third of members of families of less than four persons, whom we also lose trace of far more often than others. On the other hand, members of large families did not experience more often than others trajectories of persecution preceded by a long period of freedom (that is, absence of identification).

Do Pre-War Trajectories Explain Persecution Trajectories?

Yet, the explanation of trajectories in terms of socio-demographic characteristics continues to raise a certain number of problems relating to causal relations: did the socio-demographic specificities (such as family size or nationality) we have brought to light “cause” persecution trajectories, or were they simply indicators of other causes that still remain to be determined? If it is true that “the determining cause of

Cluster	1	2	3	4	5	Total
	From identification to extermination	Survival w/ gaps	Survival w/ clandestinity	Right-censored trajectories	From freedom to extermination	
Total	34.1	29.3	12.1	16.6	8.0	100.0
Sex						
Female	35.7	27.5	9.8	18.0	9.1	100.0
Male	32.7	31.1	14.3	14.9	7.0	100.0
Status						
Head of household	29.8	31.8	13.1	18.0	7.2	100.0
Spouse	36.4	30.2	9.3	15.1	8.9	100.0
Child	36.0	27.5	12.8	16.0	7.8	100.0
Family Size						
1 or 2	28.8	26.4	12.9	20.2	11.7	100.0
3	30.9	28.4	14.4	20.6	5.7	100.0
4	31.2	37.2	12.3	14.3	5.0	100.0
5	38.5	25.2	10.4	12.6	13.3	100.0
6+	46.3	21.1	9.5	15.0	8.2	100.0
Parenthood outside household						
0 family link	30.7	26.9	11.8	22.4	8.2	100.0
1 family link	40.6	31.6	13.1	7.4	7.4	100.0
More than 1	35.8	34.0	11.7	10.5	8.0	100.0
Nationality						
French	28.7	32.7	14.3	19.7	4.5	100.0
Polish	36.2	32.3	12.2	10.0	9.3	100.0
Other	51.3	17.1	11.8	17.1	2.6	100.0

Table 9.4 Five clusters of trajectories through persecution

a social fact must be sought among previous social facts" (Durkheim 1894), may we not make the hypothesis that the determining cause of a part of a trajectory must be sought in previous parts of that trajectory? It is precisely the hypothesis that we wanted to test, which is why we also applied a procedure of automatic classification to the portions of trajectories that preceded the War. In that purpose, we focused on the 1920s and 30s, and on the individuals who were already born by 1920, thus being at least 19 years old in 1939; we discriminated "free" states according to residence, by distinguishing between foreign residence, France, and Northern France. That pre-War clustering procedure revealed three distinct types: a first cluster grouped individuals who combined earlier arrival in France and earlier settlement in Lens (N=177); a second cluster of similar size (N=180) grouped Jews who also arrived earlier in France but settled later in Lens (i.e. closer to the beginning of the

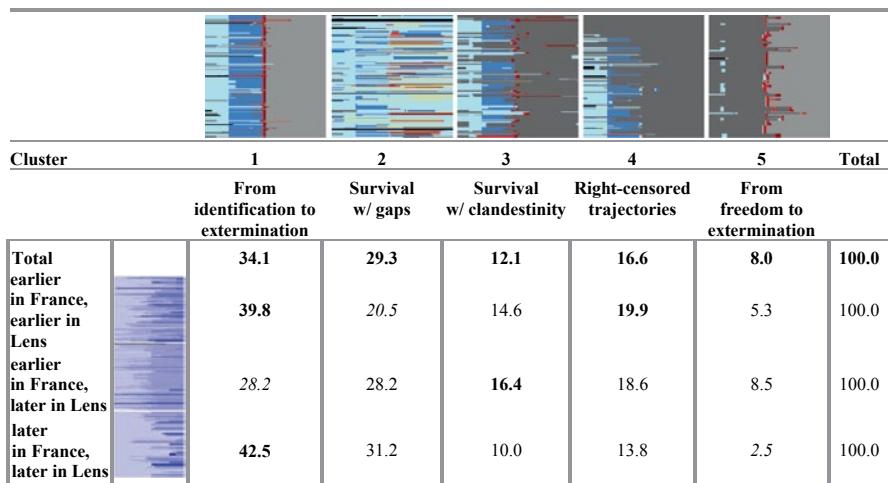


Table 9.5 Pre-War trajectories and persecution trajectories

War); and a third smaller ($N=80$) cluster grouped people who combined later arrival in France with an equally more recent settlement in Lens² (Table 9.5).

The result was striking, establishing relatively clear relations between trajectories of persecution and pre-War residential and migration trajectories. It thus appears that with a similar date of arrival in France, a later settlement in Lens reduced the risk of experiencing a trajectory of extermination preceded by identification (cluster 1); on the other hand, when arrival in France occurred later, the risk of experiencing this type of trajectory was very high: for individuals who had lived for a long time in France before settling in Lens, only 28.2% experienced this kind of trajectory, whereas this was the case with 42.5% of those who arrived in Lens only late, whether directly from Poland or after shorter stays elsewhere in France. Two categories of Jews are distinguished by an early identification: Lens residents of long date, who no doubt thereby displayed a certain trust in the country that had accepted them, where they were often born or granted naturalization, and the most recent immigrants, who lacked the resources to try and escape the administrative injunction. As for survival trajectories, these also took different paths as a function of pre-War residential trajectories: Lens inhabitants of longer date tended more than others to “disappear from the radar” (cluster 4), whereas Jews who had lived elsewhere in France before moving to the North may have had more opportunities to go underground (cluster 3). In total, we may reasonably believe that as well as the traditional socio-demographic factors envisaged above (age, nationality, family size, etc.), it is now necessary to add to the analysis the biographical trajectories themselves as possible explanatory elements of persecution experiences.

² Two remaining clusters were discarded from analysis and are not displayed in Table 9.5, as they mostly grouped poorly documented trajectories, either right-censored or containing important amounts of “biographical gaps”.

Conclusion

As we have seen, the formatting of data collected in a prosopographic perspective into trajectory data is neither simple nor self-evident. It does however produce innovative results. If the systematic application of the tools of sequence analysis to our data constitutes only a first indication of the existence of connections between pre-War trajectories and persecution trajectories, this is undeniably an invitation to explore further the possible factors of biographical coherence that they bring to light: were the residential histories of the Lens Jews before the War endowed with more or less in the way of specific resources (relational, material), which were subsequently more or less readily mobilizable in the face of persecution? Should we also see here an effect of the hysteresis of habituses (Bourdieu 1980; Bourdieu 2002), with dispositions acquired in the course of migratory experiences continuing to play a role, far beyond their point and time of arrival, in chances of survival in the face of persecution? Should we pursue the exploration further, particularly in the direction of analysis of social networks, inasmuch as certain trajectories may constitute opportunities of accumulating a social capital of relations that could be mobilized in the face of persecution, and which could in turn be effects of structures of relations (Mercklé 2011)? Whatever the answers given, which still remain largely to be constructed, there can be no doubt in any case that the formalizing of biographies in the form of trajectories allows us to formulate these questions in an operatory fashion, questions that it would not be possible to raise without the mobilization of the tools of sequence analysis.

In conclusion, despite these difficulties, one of the results of this text is to promote a new way of working on the Holocaust process, using all the traditional methods of historians. We are convinced that the methods of the social sciences can be applied to objects of research that due to their exceptional character are also objects of intense debate and contention. There is no reason to write the history of this period with different tools than those used by other historians and social scientists.

References

- Abbott, A. (1990). A primer on sequence methods. *Organization Science*, 1(4), 275–392.
- Abbott, A. (2001). *Time matters. On theory and method*. Chicago: University of Chicago Press.
- Anders, E., & Dubrovskis, J. (2003). Who died in the holocaust? Recovering names from official records. *Holocaust and Genocide Studies*, 17(1), 114–138.
- Bourdieu, P. (1980). *Le sens pratique (Le sens commun)*. Paris: Ed. de Minuit.
- Bourdieu, P. (2002). *Le bal des célibataires. Crise de la société paysanne en Béarn*. Paris: Seuil.
- Browning, C. R. (2010). *Remembering Survival: Inside a Nazi Slave Labor Camp*. Chicago: W.W. Norton & Co.
- Croes, M. (2006). The holocaust in the Netherlands and the rate of jewish survival. *Holocaust and Genocide Studies*, 20(3), 474–499.
- Desrosières, A. (2001). Entre réalisme métrologique et conventions d'équivalence : les ambiguïtés de la sociologie quantitative. *Genèses*, 43, 112–127.

- Durkheim, E. (1894). *Les règles de la méthode sociologique* (rééd. 1988 ed., Champs). Paris: Flammarion.
- Fussell, E. (2005). Measuring the early adult life course in Mexico: An application of the entropy index. In R. MacMillan (Ed.), *The structure of the life course: Standardized? Individualized? Differentiated?* (pp. 91–122). Amsterdam: Elsevier.
- Gabadinho, A., Ritschard, G., Mueller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Gross, M. (1994). Jewish rescue in holland and France during the second world war: Moral cognition and collective action. *Social Forces*, 73(2), 463–496.
- Klarsfeld, S. (2012). *Mémorial de la déportation des Juifs de France*: Paris, Fils et Filles des Déportés Juifs de France (FFDJF).
- Lemercier, C., & Zalc, C. (2008). *Méthodes quantitatives pour l'historien*. Paris: La Découverte.
- Lesnard, L. (2010). Cost setting in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3), 389–419.
- Macindoe, H., & Abbott, A. (2010). Sequence analysis and optimal matching techniques for social science data. In A. Bryman & M. A. Hardy (Ed.), *Handbook of data analysis* (pp. 387406). London: Sage.
- Mariot, N., & Zalc, C. (2010). *Face à la persécution. 991 Juifs dans la guerre*. Paris: Odile Jacob. Fondation pour la Mémoire de la Shoah.
- Mariot, N., & Zalc, C. (2012). Destins d'une communauté ou communauté de destins? Approche prosopographique. In T. Brüttmann, I. Ermakoff, N. Mariot & C. Zalc (Ed.), *Pour une micro-histoire de la Shoah* (pp. 73–96). Paris: Seuil.
- Mendelsohn, D. (2006). *The lost: A search for six of six million*. Scarborough: HarperCollins.
- Mercklé, P. (2011). *Sociologie des réseaux sociaux*. Paris: La Découverte.
- Pollak, M. (1990). *L'expérience concentrationnaire. Essai sur le maintien de l'identité sociale*. Paris: Métailié.
- R Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. URL: <http://www.R-project.org/>. Accessed on 17 october 2013.
- Tammes, P. (2007). Jewish Immigrants in the Netherlands during the Nazi Occupation. *Journal of Interdisciplinary History*, XXXVII(4), 543–562.

Chapter 10

A Contextual Analysis of Electoral Participation Sequences

François Buton, Claire Lemercier and Nicolas Mariot

Introduction

This chapter presents the first steps in a collective research program, PAECE, that aims at qualifying mainstream models in electoral studies by considering electoral participation not merely as an individual act, but also as the product of social environments (CEPEL 2009). In order to observe such environments, small spaces of analysis have been chosen by each research team in PAECE, in the spirit of ecological analysis. In these French local contexts, it has been possible to use records of turnout in order to analyze behaviors, not discourses; this difference in sources is important, as abstention is under-reported in opinion polls and interviews (Braconnier and Dormagen 2007). We had of course no access to the contents of the vote, but we know whether registered voters participated in each election, thanks to signature lists. One of these teams, that of François Buton and Nicolas Mariot, built a longitudinal dataset: for one polling station, situated in a residential part of a town in the Paris region, it records 29,756 traces of acts of turnout or abstention, over three decades (1982–2008) and 44 ballots. It also includes limited but useful information on voters and especially on their living together in households.

Such a dataset requires some sort of sequence analysis. Of course, it is always possible to reduce turnout trajectories to one number by computing individual participation rates. We have done it and it has produced interesting results. However, patterns in the trajectories themselves also deserve to be described and interpreted. We have begun to do so in a paper that has been published in a mainstream journal in the field (Buton et al. 2012), despite or thanks to its methodological peculiarities. Interested readers should refer to this paper to learn more about the setting of the

C. Lemercier (✉)

Center for the Sociology of Organizations (CSO), CNRS, Paris, France
e-mail: claire.lemercier@sciencespo.fr

F. Buton

Center for the Political Study of Latin Europe (CEPEL), CNRS, Montpellier, France

N. Mariot

European Center for Sociology and Political Science (CESSP), CNRS, Paris, France

case study, which we will only briefly sum up here. Our first results are also presented there. They show an extremely high homogeneity of voting patterns inside what we call “electorate households”. Registered voters who live together tend to vote or abstain at the exact same moments. Not taking into account this specific social dimension of participation thus leads to misinterpret it. This chapter provides a more methodological reading of the first stages of this research, those that led to the 2012 paper and those that happened afterward.

The second part presents the ways in which our research design allows us to contextualize the analysis of electoral participation. This design is replicable and in no way specific to France; we hope that it will be an inspiration for other researchers. The third part sums up our published results on the correlation of participation behaviors inside households and the correlation between position in the household and participation. It shows why sequence analysis was essential in producing these results and how a slightly different, less standard modeling can help to better interpret them. The fourth part opens an even more temporal question: do parents vote more when their children become potential voters? We show how visualizations help to tackle this question. The fifth and final part presents tentative results that differentiate ballots according to the type of election, hence introducing a third type of contextualization, after the synchronic household context and the diachronic context of individual voting trajectories.

A Contextual Analysis

Electoral studies generally belong to the realm of standard methods considering a “general linear reality”, in the words of Abbott (2001). Information on voting behavior is gathered through large-scale surveys of representative samples of non-related voters. This behavior is explained by more or less standard attributes of voters, e.g. gender, age, occupation, religion. Voting is seen as a conscious decision, isolated from social contexts as well as isolated in time: even recent panel studies have rarely taken more than four ballots into account. Hence, turnout is often considered as the output of an individual, general tendency that can be, e.g., related to an interest in politics and correlated with variables such as age or education.

However, pioneering studies taking into account more than one ballot have abundantly shown that consistent participators and consistent abstainers were a minority (Lancelot 1968; Subileau and Toinet 1993). Our own data makes this point even clearer: if we concentrate on the 1,665 voters who were registered for at least three ballots, we only find 16% of consistent participators and 7% of consistent abstainers, most of the latter being registered eight times or less. Turnout thus has its patterns: trajectories are rarely plain, but they are not random either—hence the interest of exploring them.

We study them with hypotheses that are rather different from the standard model. We consider turnout as an institution, in the sense of something that individuals tend to take for granted, not to reflexively question: a sort of routine that is collectively

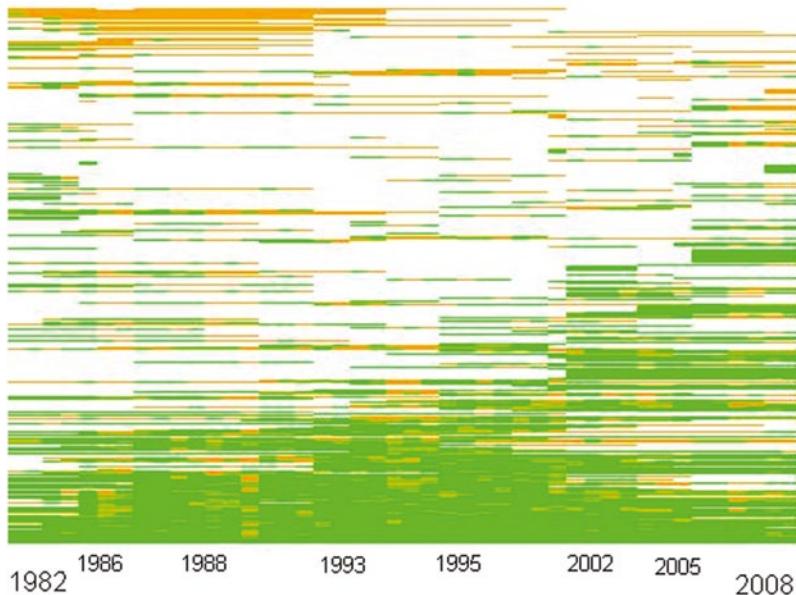


Fig. 10.1 A visualization of participation trajectories (index-plots). Sorting based on multi-dimensional scaling on the optimal matching distance defined below. Participation: *Black/Green*. Abstention: *Gray/Orange*. Not registered: *White/White*

performed, in which imitation, acting like other voters, plays an important role. Participation is not performed by everybody, for abstention happens; but it happens neither randomly nor necessarily as a product of an individual conscious decision. In some cases, abstention can also be a routine. This view of electoral participation calls for a contextualized study, in order to test the importance of various contexts: do voters act like voters close to them, especially in their own household (social context)? Do they act as they used to (temporal context)? Do they act like other voters at the same time, whatever the personal connection (political context)?

Three Contexts of Turnout

A very simple visualization of our trajectories, provided, like the rest of our analysis, by the R package TraMineR, clearly illustrates this point (Fig. 10.1; for R, see R Development Core Team 2013). Each line in this index-plot represents a voter, and lines are sorted according to their pairwise similarity, as defined by a distance that we will present below. The scarcity of consistent behaviors is visible. There are few completely black or green lines, describing consistent participators, and almost no completely orange or gray lines, denoting consistent abstention.

Many trajectories include large white patches: these are associated with ballots when the voter was not registered in this polling station. He could have been too

young to vote (the minimum age is 18 in France) or deceased at that time, but most of these non-registrations are related to moves in space. In France, registration is automatically performed by the administration only for young people who turn 18; they are then registered at the polling station where they live. When a voter moves to another place, hence a different polling station, it is up to her to state it and make a new registration in her new domicile. If she has just moved to our polling station, her colored line begins at that time; if she has moved from our polling station to another, it ends.

Patterns in our trajectories are not only created by registration. Vertical orange/gray lines can be noticed in the dominant green/black landscape: while a majority of registered voters participated in most ballots, some are characterized by overwhelming abstention. In addition to these horizontal or vertical lines, a more subtle pattern is also apparent: there is more green in the lower-right corner, more orange in the upper-left one. It thus seems that those who became registered voters late in our period often tended to vote, contrary to those who were only registered in the early years. What is even more obvious, finally, is that much diversity remains to be described and explained, apart from these general patterns.

The aim of our research is to provide such a description, which is not an easy task, if we want to take the timing and not only the aggregate number of acts into account. In addition, we will consider possible mechanisms that would help to make sense of this description. They relate to three types of contextualization: social, temporal and political. They will play a prominent role in the next parts of this chapter; let us briefly list them here.

First, voting does not appear as an individual decision, because people who live together tend to go, or not to go, to the polling station together. If we group the lines in Fig. 10.1 according to household, new, very sharp patterns appear that capture a large part of its diversity. It is our most important result, and taking timing into account allowed us to better assess it, as we will show in the third part of this chapter.

Secondly, although consistent participants/abstainers are rare, it is reasonable to think of “participation careers” in which voters more or less consciously build on their previous actions, acting consistently in one way or another—even if consistency consists in regularly alternating between vote and abstention. In a side comment, Abbott (2005) clearly stated that in his view, standard electoral studies were wrong because they underestimated the fact that votes are mostly based on individual memories (he did not consider our two other types of context). Our dataset allows this sort of temporal contextualization, although we have not yet focused our investigations on it. The general idea, here, is to focus on the horizontal lines of Fig. 10.1. Parts three and four of this chapter will show our first steps in this direction of a genuine longitudinal study.

Third, each moment in time, here each ballot, also can be considered as a specific context, in that the registered voters are more or less exposed to the same information from the mass media; to the same, narrow or wide, range of political parties, etc. It means that vertical lines in Fig. 10.1 are not to be ignored. Attributes of the ballots therefore should also be taken into account in our analysis—a dimension of context that we will briefly begin to address in part five. While we have not yet

brought all these dimensions together in a comprehensive description, let alone a model, our point is here to show that our research design, simple as it seems, allows us to tackle all of them.

A Specific Research Design

We chose this design because our background was quite distinct from that of specialists of electoral studies, which deserves a few comments. First, we learned quantitative methods in France, a country where specific, more descriptive and often more contextual methods than in the English-speaking world have been developed for applications in sociology and history and have become dominant in some fields—such as, for example, multiple correspondence analysis (Greenacre and Blasius 1994). Using formal methods in micro case studies that allow researchers to carefully build their own datasets, including non-standard variables, has been more routine there than elsewhere (see Lemercier and Zalc (2008) for a handbook and Mariot and Zalc (2010) for an example). Therefore, many French scholars have read Andrew Abbott's methodological work with enthusiasm (Demazière and Jouvenet *forthcoming*)—some of them are in fact co-authors of this very book. Secondly, we share ambiguous disciplinary allocations, somewhere between political science, sociology and history, which is related to the label “*socio-histoire*” (Buton and Mariot 2009). For present purposes, it translates into an interest in processes and configurations, in the spirit of Norbert Elias, and into the aim of dealing with social science questions by using “historical” sources, in the sense of traces, longitudinal data and data that is situated in time and space.

In this context, it was quite normal for the two of us who gathered data to input them ourselves from the source, during long days in the polling station. It allowed us to get a better, somewhat ethnographic knowledge both of the place and of the source, hence to devise better coding schemes and research questions: this type of data input can be considered as fieldwork. For example, one of the main drawbacks of our source is the lack of any direct indicator of education, wealth, occupation or social status. Personally knowing the place allowed us to decide on a coding of neighborhoods from addresses that offered reasonably defined clusters in terms of type of housing, hence indirect indications on social status. The specific history of the polling station also matters for our interpretation: initially a small village, it was included in the urban planning of the Paris region in the 1970s, and grew rapidly, with changing occupations, during our period. Finally, inputting very tedious data on turnout allows us to observe some of the standard and exceptional patterns in trajectories, which helped us to better mine a dataset that is at the same time very simple and very complicated. This research design, although produced, in our case, by a specific French academic socialization, could easily be adopted by other researchers. It would probably produce interesting results on electoral participation. More generally, it could benefit the methodology of sequence analysis by expanding its application to different types of data.

Unusual Sequences

Our dataset is indeed peculiar, as compared to what has until now been the bread and butter of sequence analysis: life-course and/or occupational trajectories, often gathered from large surveys that do not allow much re-coding of states or covariates. It is not only the political theme that creates this difference, but mainly the fact that our 1,799 individual trajectories are not statistically, or substantively, independent. On the contrary, the very purpose of our case study is to discuss whether similarities between trajectories can be related to ties between individuals, be it because of direct interactions in political discussions or because of shared exposure to the same micro-environment. As is classically the case in network analysis, in order to study such questions we have to forsake the advantages of large representative samples and turn to an ecological case study.

The problem of boundaries hence becomes complicated. Working on signature lists led us to concentrate on one polling station that provided us with a large enough number of trajectories. However, the perimeter of our polling station changed during the decades that we study. We chose to concentrate on a fixed perimeter in terms of addresses, considering all the voters who lived there and those alone. More importantly, as has already been seen in Fig. 10.1, people moved into and away from the part of town where our polling station was situated: only 9 % of the population was registered during the whole period, while 10 % was only registered for one ballot.

To deal with this issue, we had to accept a trade-off between two types of contextualization that we were interested in: on the one hand, that of voters among fellow voters registered in the same, small place (there were 500–800 simultaneously registered voters); on the other hand, that of voting acts among past and future voting acts by the same persons. Following voters from their first registration onwards, in their successive polling stations, would be completely impractical in terms of sources, as it would require us to look for records in many different places. Were it possible, it would certainly help us better understand how previous votes shape later votes and how this interacts with specific political contexts (prominence of the election in the media, number of candidates, etc.) and micro-environments (behavior of the parents, spouse, neighbors, etc.).

While our dataset is very local, historical and not very large, it is quite different from those built by Abbott and Hrycak (1990) in an early analysis of historical data, or by Blanchard (2012) in a study of multi-dimensional mobilization careers. In both cases, as in ours, the trajectories are those of named individuals who constitute the whole of a population; each line of the dataset is potentially interesting per se and in its relationships with others, and individuals are perceived as more than a collection of variables. We are not studying a representative sample of the French population. We are attempting a different sort of generalization. Our aim is to find interesting patterns in data on turnout and covariates and to hypothesize some of their generative mechanisms; it is these mechanisms that are offered as candidates for a role in other case studies.

This goal is probably shared with Abbott and Blanchard. However, as compared with their datasets, ours might appear ridiculously simple. We only study three possible states for each ballot: voting, not voting, not being registered. Although the third option creates methodological problems, coding, color-coding, etc. seem trivial in our case. In addition, our short list of variables was entirely built on the basis of the sex, name, maiden name, address, date and place of birth and date of first registration of the voter: this was all the information available in our source.

However, our ethnographic approach to data input and our intention of taking all the dimensions of contextualization seriously led us to consider these admittedly poor data as potentially rich in meaning (as advocated by Rosental (1999) for similar data on migration). For example, translating our data into a representation such as that in Fig. 10.1, a more or less standard operation in the non-standard field of sequence analysis, in fact involves at least two implicit decisions. First, any act of turnout is the same as the others (the same color, here). This is in fact not self-evident: voting when most of the population does not, for example, is admittedly a specific behavior—in the same way as working when most people do not (a problem which led Lesnard (2010) to devise a specific sequence dissimilarity metric). We had many discussions on how to take this point into account when computing individual participation rates; we finally chose flat rates for our 2012 paper and plain colors for our graphs, but we here present reflections on this topic in part five. Secondly, time is treated in a peculiar way in our sequences, which are sequences of ballots. There were sometimes four ballots in an interval of 8 days (two separate elections on the same days, with the first and second ballots for each), while we had an interval of 3 years without any ballot from 1989 to 1992. We chose to consider only one act when two ballots happened on the same day, as almost all voters had consistent behaviors; but this leaves us with intervals ranging from 1 week to 3 years in calendar time, and with ten separate ballots for 1988–1989. Finally, we obviously do not know much on how voters considered their own “voting careers” and their timing, if they did at all. Our homogeneous time scale based on ballots, not years, is therefore a methodological choice in itself, to be questioned and possibly changed in future stages of this investigation.

Households as a Context of Participation

This simple way to consider participation trajectories as sequences of 44 successive states of participation, abstention or non-registration already led us to interesting results in terms of social contextualization. Had we not used sequence analysis ideas and software, these results would certainly have been less convincing. However, the fact that we published them in a mainstream journal has much to do with the recent evolution, led by TraMineR authors, toward including some forms of tests, thanks to permutations of data, in sequence analysis. These tests are useful even for a case study like ours which deals with non-independent trajectories—both as rhetorical devices that help to convince specialists in quantitative methods and, more

importantly, as tools that proved fit to answer our main question: do people who live together also go and vote together?

In spite of the dominance of the standard models that we briefly sketched above, a few studies of voting and/or participation had already taken correlations among couples into account; one of them had even chosen households as a unit of observation, although with shorter observed trajectories (see e.g. Huckfeldt 1986; Verba et al. 2005; Braconnier and Dormagen 2007, and especially Johnston et al. 2005). We confirmed that this dimension should be taken into account in any analysis of turnout, as the magnitude of correlations in our case was very large. This magnitude could partly be due to the French schedule of elections, as they always happen on Sundays, when time spent together by household members is likely to be higher, on average, than on a standard working day. International comparisons would be required to test this hypothesis.

In order to assess the correlations among voting trajectories, we defined an “electorate household” as a group of registered voters who live together. Often, but not always, they are also married couples, or parents and children, or they share other kinship ties. The residential area that we study mostly includes detached houses, which helped us to define households on the basis of names, addresses and on-site observation. It should be noticed that “electorate households” are distinct from households in that we only consider registered voters. A registered woman with a foreign partner and young children, for example, would be counted as an “electorate household” of one. This captures ties of co-residence that empirically, in most cases, are also kinship ties. Ties with kin outside of the household, which arguably also play a role in electoral participation, could not be observed, and arguably were very few in the small space that we studied. What we are therefore able to assess is whether voters who are also coresident, and generally coresident kin, vote at the same moments. What we are interested in is the very collective character of this behavior. As in other studies of collective and especially family behavior, we are not able to determine if it is created by some sort of internal influence (from parents to children, or vice versa, etc.) or by exposure to a shared micro-environment (on this alternative, see e.g. Palloni et al. 2001).

Correlations in Several Dimensions of Voting Trajectories

On the basis of this definition, we found a correlation within households in terms of overall participation rates. We used ANOVA to show its significance. In addition, a multilevel regression that also included individual attributes such as the date of birth or that of registration demonstrated that the similarity within households had a significant, autonomous and very important effect on participation, representing a majority of the variance. However, this only proves that members of the same “electorate household” have similar aggregate tendencies to vote or not to vote, whatever the ballot.

We added to this result in two ways, by taking the sequential character of our data rather more seriously. First, we devised two alternative ways to measure participation, in addition to aggregate rates: a Change of Behavior Index and an optimal matching distance.

The Change of Behavior Index simply computed the number of changes from participation to abstention, and vice versa, as a proportion of the number of registrations, for each individual trajectory. This allows us to differentiate between two types of voters with similarly medium aggregate participation rates. The first type includes voters who turned from consistent participators to consistent abstainers, and vice versa. A proportion of them were in fact people who had moved away from the polling station, but neglected to change their registration. This phenomenon, called “misregistration”, is common in France: especially when the new domicile is far away, voters who have not bothered to change their registration often do not make the trip back in order to vote, hence abstain. The second type is made up of voters who regularly alternate from vote to abstention, possibly according to the national or local character of elections, the available candidates, etc. Voters of the second type have a high Change of Behavior Index, contrary to those of the first type. This index also showed a very significant correlation within households, indicating similarities in the shape of participation careers, not only in aggregate participation rates.

In addition, we wanted to more directly demonstrate that people who lived together tended to participate at the exact same moments. This is essential for our institutional, routine interpretation of the act of voting: it is not simply some abstract tendency toward participation that is shared in an “electorate household”, but the specific fact of participating at the same moment. This was done using a computation of dissimilarities between sequences based on an optimal matching distance. Had our data not included non-registrations, we would simply have counted the exact matches between sequences. The existence of this third state made optimal matching useful. We chose the lowest possible substitution costs between participation or abstention and registration (0.51) as compared to those between participation and abstention (1) in order to minimize the effect of non-registration on distances. In addition, we used very high indel costs (5), so that similar sequences were those where vote or abstention happened at the exact same moments—according to our research question. These choices of course do not totally prevent sequences from being considered close just because voters were registered at the same moments, but they minimize this effect as compared to those that were more substantively interesting for us.

This choice of costs allowed us to characterize each household by a level of internal variance between sequences and to compare such levels to that of the population as a whole. Pseudo ANOVA calculations implemented in TraMineR (Studer et al. 2011), although originally designed to test homogeneity of sequences in larger groups (such as women or the youngest), proved very useful in assessing the significance of similarity in temporal patterns of voting inside our 469 households that included more than one voter, and inside the 210 that included at least three. Changing scales to a visual inspection of the large households that showed an especially

high or low internal variance helped us to better understand and describe our data. While tests added strength to our demonstration, such comings and goings between the population and examples, between numbers and visualizations, helped us to better grasp and describe our dataset.

Participation and Position in the Household

After having extended the study of correlations inside households to temporal patterns, we added a second dimension to the literature. Our data allowed us not only to define the perimeter of “electorate households” but also to assign a position in the household to each individual. After having tried more complicated coding schemes, we classified these positions as follows. “Couples” (27% of individuals) are those who live as spouses (married or not) with another registered voter, but without any child registered as potential voter. “Children” (21%) have at least one “parent” (22%) present in their “electorate household”, and vice versa. “Others” (4%) live with at least one other registered voter, but do not fall into the previous categories. Finally, “isolated” voters (26%) are the only registered voter in their household.

Our hypothesis was that households mattered not only because some sort of alignment of behaviors happened between their members, but also because they produced social roles that could influence participation, which would lead to differences *within* households. In our multilevel regression, which also included the household effect and other individual attributes, with the overall participation index as the dependent variable, position in the household indeed proved significant, with a hierarchy from “isolated” (who participated less) to “others”, “children”, “couples”, and “parents” (who participated more). This effect is not that of age, as some “children” were older than “couples” and as the date of birth was also included as a covariate. It hints at a specific type of social integration that seems to affect turnout: not the existence of social or family ties per se, but that of ties with other registered voters of the same polling station.

While we did not publish it previously, both because of a lack of space and of the non-standard character of the method, we also tried to test the effect of position in the household, along with that of other individual covariates, on the exact timing of participation. To do this, we used the dissimilarity tree method implemented in TraMineR (Studer et al. 2011), that produced Figs. 10.2 and 10.3. This method performs a tree structured discrepancy analysis of objects that are described by their pairwise dissimilarities: here, participation trajectories described by their optimal matching distances. The procedure iteratively splits the data. At each step, it selects the variable and split that explains the biggest part of the discrepancy, i.e. the split for which we get the highest pseudo R² in TraMineR’s pseudo ANOVA procedure. The significance of the retained split is assessed through a permutation test. This method accommodates qualitative as well as quantitative variables; in each case, the algorithm decides on the best possible way to split the observed values into two groups. It is a generalization to sequences of the induction tree method of Breiman et al. (1984).

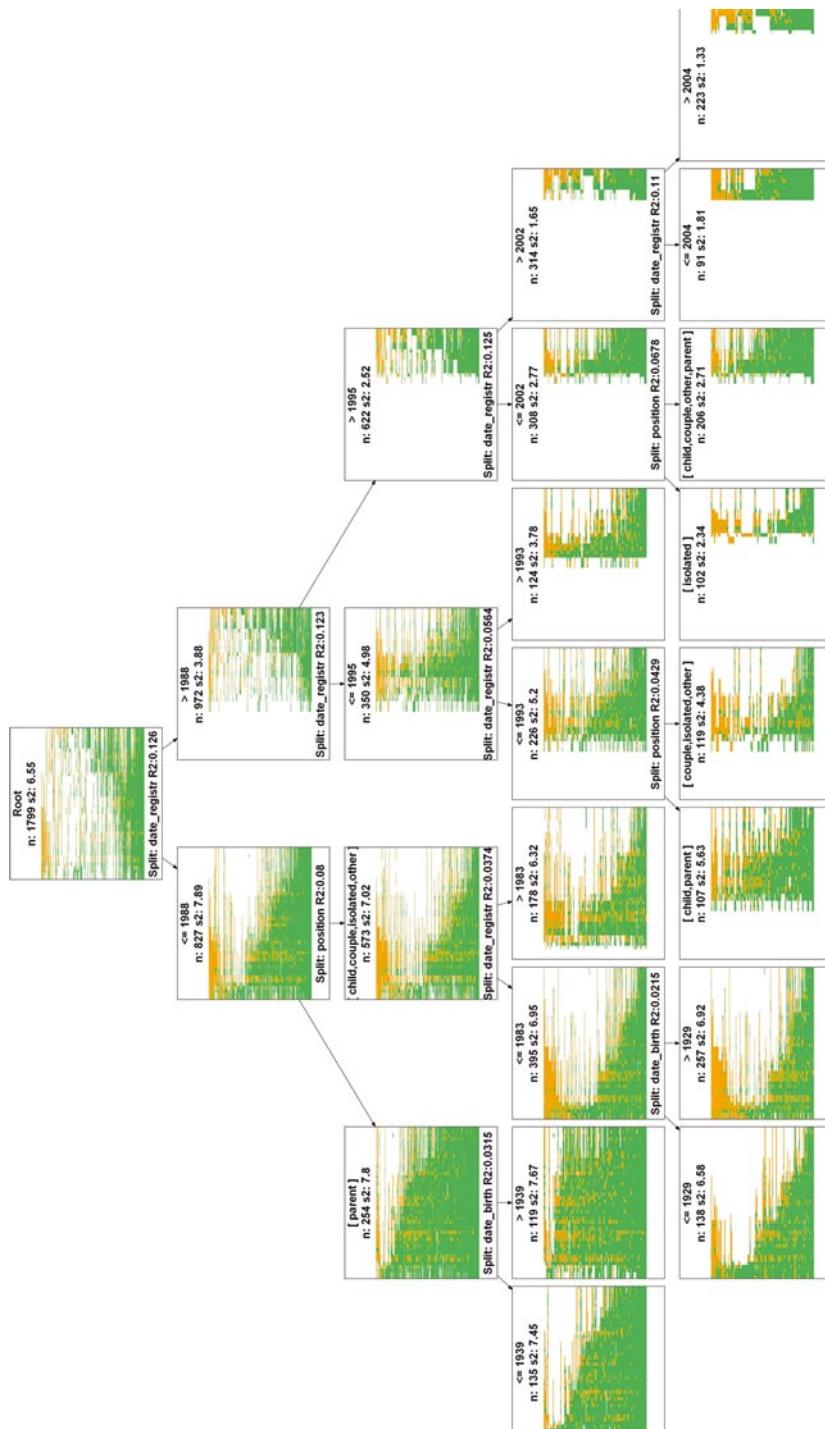


Fig. 10.2 Tree diagram based on an optimal matching distance-index-plots. Covariates that were included but do not show up in the results: sex, place of birth, neighborhood, and distance from address to the polling station (as in the multilevel regression included in Buton et al. 2012). Color code: Participation: Black/*Green*. Abstention: *Gray/Orange*. Not registered: *White/White*

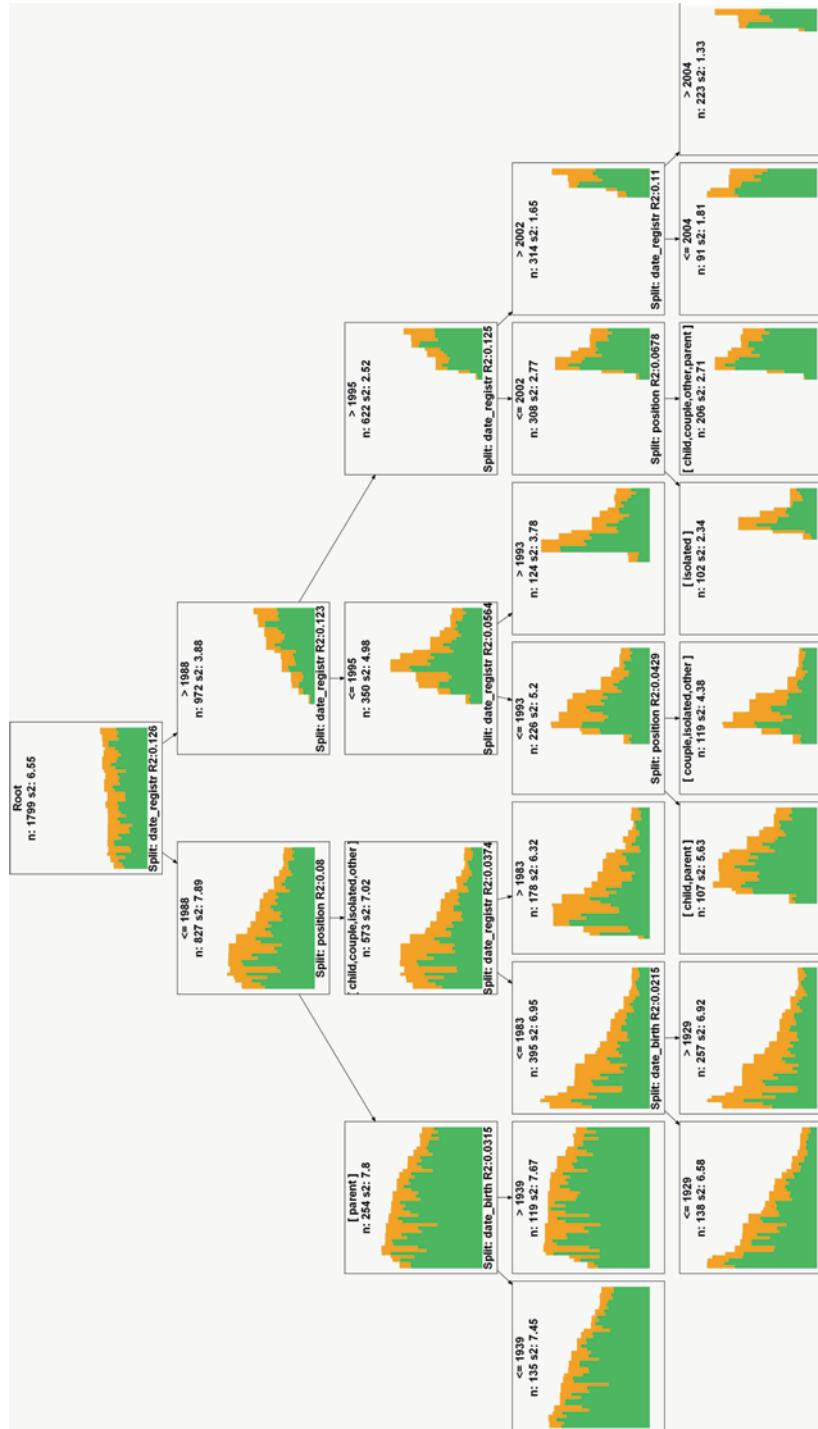


Fig. 10.3 Tree diagram based on an optimal matching distance-transversal distributions. Legend: see Fig. 10.2

This method has important limits in our case. First, due to the large number of different households, it is not possible to use multilevel modeling, including homogeneity inside each household along with individual covariates, while we have shown that it was an important dimension in our data. What we model here is thus something like a residual: what happens apart from this homogeneity. Secondly, what drives the model is the search for groups that have the lowest possible internal variance, in terms of distance between sequences as defined by optimal matching; despite our choice of costs, it is mostly in terms of date of registration, on the one hand, and of overall level of participation, on the other hand, that these groups appear quite visibly homogeneous. However, we can be confident that if something important sets them apart in terms of exact timing of turnout, it should be taken into account in the diagram.

This being said, this analysis is helpful to better understand our data in several ways. First, as compared to regression coefficients, the visualization helps to assess the magnitude of each variable and especially not to overestimate the homogeneity of behaviors among individuals that are similar according to a few significant covariates. For example, the 254 “parents” who registered before or in 1988 are quickly singled out by the model; but the internal variance in this group is higher than that of the total population (due to the fact that generally they were registered for a long time), and visual inspection reminds us of differences among them. Secondly, while the prominent part played by the date of registration in this model is admittedly partly due to the effect of non-registration on distances, what is not trivial is the way in which the procedure defined classes of dates, with significant breaks in 1988, then 1995, then 1993 and 2002. It is interesting to use such an inductive way to build classes, rather than using arbitrary thresholds; the same is true for dates of birth. Thirdly and more importantly, what a tree diagram does and a regression does not do is consider and put forward interactions among variables, not the independent action of those same variables. This is very much in the spirit of Abbott and of French methodologies, and we consider it an important complement to our regression.

As regards position in the household, the tree diagram very much validates its significance. It is the only covariate that appears in this model along with the date of registration and that of birth: the small effects of the place of birth and distance from the polling station that we found in the regression on overall participation rates do not show up in this model on the exact patterns of participation. Trajectories differ primarily according to the date of first registration, before or after 1988, which first splits the tree into two branches. For the early registrations, the difference between “parents” and all the other positions in the household causes the second significant split. For the later cohorts, it is still the date of registration that is most discriminating, with successive splits. However, position in the household also appears significant in those right-hand branches of the tree, except for voters who were only registered after 2002. Although its significance is higher for voters who were registered in 1988 or before, it is present almost everywhere.

While our regression showed a hierarchy of positions in terms of participation, the tree diagram adds that this hierarchy also applies to the duration of registration:

“parents” tend both to be registered for a longer time and to vote more than the others and especially than the “isolated”. Social integration in “electorate households”, local integration in terms of registration (being registered, and not moving away) and electoral participation are strongly correlated. Finally, there are slight differences between branches of the tree in terms of which positions exactly are different from the others, as regards exact participation patterns. This is food for thought for our future investigations, for example as regards the difference between the “isolated” and all members of “electorate households”: this difference is very strong for registrations happening between 1996 and 2002, but not for the other groups. Conversely, it is “parents” that are singled out in the first registration cohort. However, these small differences remain roughly consistent with our continuum of positions, from “isolated” to “parents”. This intriguing effect of social integration thus deserves further investigation.

Getting More Sequential: from Couples to Parents

We present here a very preliminary investigation in this respect. It begins with a critical return to our coding scheme for positions in the “electorate households”. Our definition of households and positions is static: each individual trajectory is related, in our database, to a given household, which is deemed to have, for example, four members; and each individual is assigned a position, e.g. that of “parent”. However, in the simple case of a couple with two children who became voters during our period of observation, the number of voters in the household in fact increased from two to three, then four, while the “couple” became “parents”. However convenient for a first step of analysis, it is thus somewhat odd to classify trajectories according to static labels.

In the case of positions in the households, there is a more substantive reason to take this matter seriously. How come that “parents” vote more than “couples”? “Couples” may or may not have children; what differentiates them from “parents”, in our coding scheme, is that the children of “parents” are also registered voters in the same polling station. Why would this lead to higher participation? Do parents want or feel compelled to set an example by voting more, once their children are also registered? Or do “parents” participate more just because this position happens to be correlated with other variables, without having any specific causal effect? We have no final answer to this question, which will of course also require comparisons with other case studies and a more qualitative inquiry. Yet we can begin to tackle it, thanks to our database and a few visualizations.

What can we learn by observing those “couples” who become “parents” in our dataset? Before talking about the “effect” of a “variable”, it is important to observe the individuals who experienced a change in the modality of this variable (Blossfeld and Rohwer 2002). Here, while it rapidly leads us to manipulate small sub-samples, it provides food for thought and future testing. Let us consider the sub-sample of 120 “children” in our database who came to be registered after our first date of

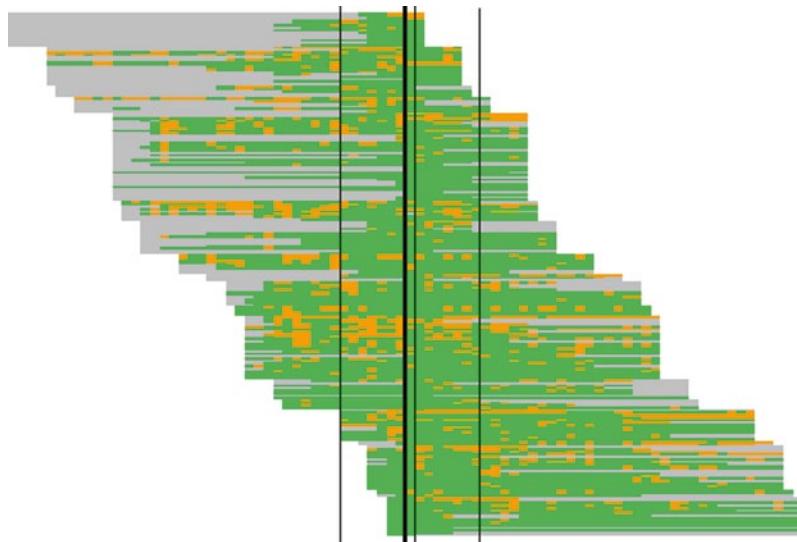


Fig. 10.4 Participation patterns of “parents” before and after the first registration of a child. Trajectories sorted according to registration date of the parent. Participation: *Black/Green*. Abstention: *Dark Gray/Orange*. Not registered: *Light Gray/Gray*

observation, along with their 215 parents. In fact, we only considered the first child in each household who was registered: the one who caused his parents to become “parents” in our coding scheme. We also excluded children who did not have at least one parent who was registered before them. Let us, in a first stage of investigation, ignore the fact that among these 215 parents, some are part of the same couple.

What we did was simply to align an index-plot not on calendar time, but on the date of an external event: the first ballot when a child of the individual was registered in the polling station. This apparently simple representation required the use of the piece of code devised by Denis Colombi and Simon Paye (see their chapter in this volume on aligning sequences on events). Figures 10.4 and 10.5 therefore have a quite different timeline from the other figures in this chapter: reading them requires some concentration, but this unusual representation offers new insights. Each line represents the trajectory of a parent. The ballot in the center of the figure, that between two close vertical lines (white lines in the black and white figure, bold lines in the color version), is the ballot when a child of the parent was first registered. According to our coding scheme, it is the time when a “couple” became “parents”. This time occurred in various years: there is a calendar time scale for each line, as the sequence begins in 1982 and ends in 2008, but no common calendar time scale for the whole graph. Its scale is organized around the first ballot of the child, with “first ballot minus one, minus two, etc.” on the left half and “first ballot plus one, plus two, etc.” on the right half. The two additional vertical lines (black lines in the black and white version, thin lines in the color version) define the seven ballots before and after the first registration of the child.

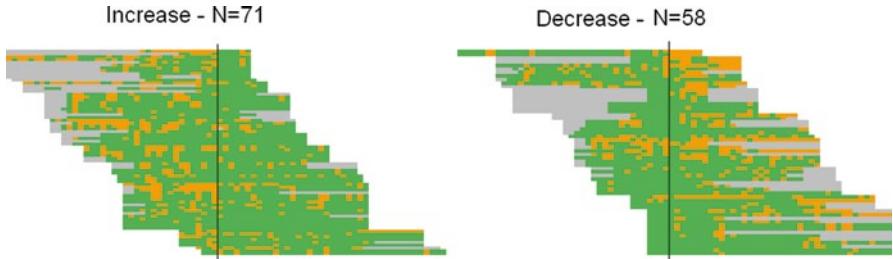


Fig. 10.5 Participation patterns of “parents” before and after the first registration of a child. Decrease vs. increase in overall participation rates. The only trajectories considered here are those that include at least four ballots before and after the event. Trajectories with equal rates before and after are also excluded. Color code: Participation: *Black/Green*. Abstention: *Dark Gray/Orange*. Not registered: *Light Gray/Gray*

No clear pattern emerges from Fig. 10.4, which is in itself a result: our event is not a turning point; it apparently does not have any lasting effect on trajectories. Its one noteworthy impact is transient: for the precise ballot when their first child becomes a registered voter, parents tend to vote, whatever their usual behavior. 100% of the few who were themselves not registered for the previous ballot participated; 94% of those who had participated did the same, as well as 71% of those who had abstained. The idea of setting a good example, as a kind of strategy of socialization into the voters’ world, would thus apply. In our institutional understanding of participation, it would mean that, at a fairly macro level, the thing to do is to vote for the first ballot of your child. However, what is even more important is that this only holds true for one ballot: no difference is to be found in mean participation rates between all ballots before and after the event, or between the seven ballots directly before and directly after. Alternatively, it is possible that many children only registered when they and their parents intended to vote for the next ballot—when they were, in one way or another, especially interested in it. Empirically, ages of registration are far from always aligned on the minimum age of 18; they are often much higher. In any case, something interesting happens at the level of households, not only for children but also for parents, at this time when our “couples” become “parents”; but this effect does not seem to be enduring. It cannot explain the general difference in participation between “couples” and “parents”.

However, careful visual exploration and successive sortings of our small subsample led us to complementary results. We were puzzled by the fact that participation *decreased* almost as often as increased when “couples” became “parents”. The longitudinal study of change seemed to directly contradict our results based on static labels, thus questioning the very category of “parents”: did they in fact vote more because this label was correlated with some other, hidden but more important variable? We therefore visualized differences between trajectories with an increase in overall mean participation after the event and those with a decrease, in order to better make sense of the latter (Fig. 10.5).

Figure 10.5 gives a hint as to what happens in cases of decrease, as we see that many “parents” in this case eventually ceased to be registered at the polling station. We checked that it was not caused by their death—at least, their age was not especially high, or higher than in the other group. Spans of abstentions often happened just before this non-registration began. This hints at two possible, non-exclusive interpretations which will lead us to put new questions to our dataset. First, the decrease in participation is partly the effect of “misregistration”, i.e. the abstention spans that appear when a voter has moved, but neglected to change registration. It is an arguably important and under-researched (see however Denni and Bon 1978; Braconnier and Dormagen 2007) dimension of abstention in systems with voluntary registration, like the French one, that we want to more generally take into account by measuring its frequency and discussing its association with various covariates.

Secondly, in other cases, an increased tendency toward abstention is directly followed by non-registration, without such spans. On the contrary, those “parents” whose participation increased after the event rarely ended up in a state of non-registration. This could indicate that a strong tie to the place, in the sense of not leaving and not planning to leave, is correlated with the increased participation of “parents”. It is all the more interesting as most “parents” with an increased participation rate had early dates of registration: they belong to the specific group that we had individualized in the left hand of the tree diagram (Figs. 10.2 and 10.3). These early registration cohorts, representing the original population of our zone—those who were born when it was a village, not the educated middle- or upper-class members who came later—generally had low participation rates, as discussed in Buton et al. (2012). These “parents” were an exception to this rule and even tended to participate more and more. Within this cohort, those who were firmly anchored in the place and whose children also became voters there seemed to have much more participatory behavior than others. This result per se is certainly specific for the place that we study, but it points to more general mechanisms as regards relationships between an enduring local anchoring at the level of the household and electoral participation. We will thus look for the same type of mechanisms, if not for the exact same results, in other case studies investigated in our research program.

Finally, another correlation with this decrease or increase in participation is worth mentioning. Small numbers call for caution in interpretation, but according to a rough classification of the neighborhoods in our zone into two groups, the more and the less wealthy, decreases happened more often in the latter group. A chi-square test is barely significant, but it is the only covariate that exhibits some sort of correlation. A provisional conclusion could be that the effect of “couples” becoming “parents” leading to an increase in participation only happens when the “parents” are wealthier and/or have a more longstanding relationship to the place. This is little more than a hypothesis, but it gives us directions for future qualitative and quantitative investigations. In addition to the two groups that we roughly defined in our first paper (villagers voting less and newcomers voting more), we are able to individualize a group of villagers who remained registered at the polling station during our period, who had children, and whose children became registered voters at their parents’ address; they tend to live in the wealthiest neighborhoods and to participate more than others.

A Vote is a Vote is a Vote?

Finally, in the same exploratory spirit, we will present reflections on our third type of contextualization: political contextualization. How can we take into account types of elections and especially their prominence in the media, i.e. the political context? All the ballots were considered in the same way, represented with the same color, etc. in our previous analyses, be they the second ballot of the 2002 presidential election, with vibrant calls to participate in order to minimize the score of far-right Jean-Marie Le Pen, or the 1988 referendum on self-determination in the overseas territory of New Caledonia, with a national participation rate of less than 40% (on participation in France in diverse types of elections, see Héran 2004). Regarding the act of turnout as one and the same in all cases is debatable: different mechanisms could be at play in different types of elections, and social and temporal contexts could especially have different roles.

Our first idea in this respect was to differentiate between local and national elections. There are no consistent differences in overall turnout rates between the two categories, as the latter include, along with the presidential elections that are generally considered the most important events in French political life, referenda and European elections, which usually generate less turnout. However, reasons to participate in local and in national elections could be different. It is generally believed that the latter have more to do with parties and their programs and with national media, while the former leave more space to face-to-face relationships and local issues. We therefore wondered whether correlations inside households were as strong for each type of election. In addition, in our case study, we already found different patterns of participation across cohorts, both in terms of date of birth and date of registration. We interpreted them as proxies of more general social differences between inhabitants who had known the place as a village and middle- and upper-class baby-boomers who had settled later and had, on average, the highest participation rates. We could expect different behaviors in these two groups as regards local and national elections.

At the level of overall individual participation rates, there is some correlation between turnout for local and for national elections in our population; but, with an R^2 of 0.55, there are also many voters who regularly participated in national elections but not so much in local ones (the reverse case being much scarcer). Linear regressions with individual covariates as independent variables and the two separate participation indices as dependent variables show that the date of birth, place of birth and distance from the polling station had a significant effect only for local elections. Counter-intuitively, it is in local elections that the difference in participation between newcomers and other is most significant, with newcomers voting more. It calls for a careful study of patterns of turnout for these elections, avoiding common sense equivalences between local anchorage and interest in local politics—as already shown by more qualitative studies (Girard 2011). Participation in national elections is much more difficult to model accurately than that in local elections—perhaps because it is more correlated with variables that we cannot observe, such as wealth, occupation, education or political commitment.

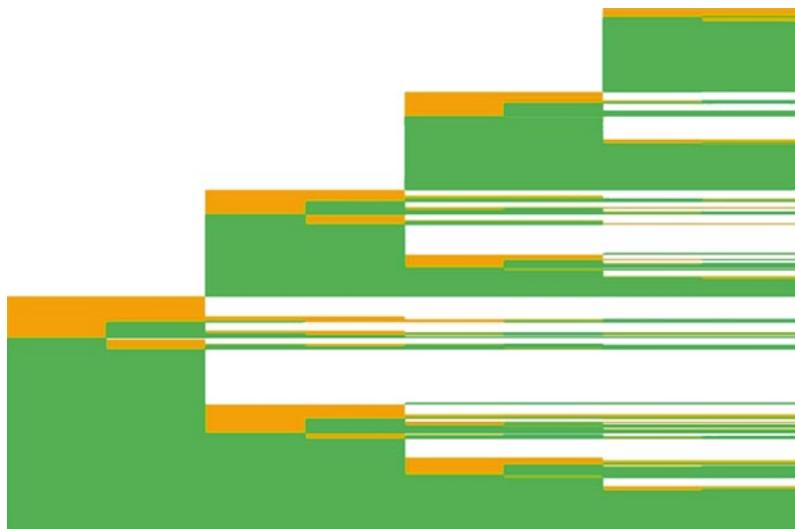


Fig. 10.6 Participation trajectories, presidential elections of 1988, 1995, 2002 and 2007. Trajectories are sorted according to the behavior at the first, second, etc., ballot (non-registration, then abstention, then participation). The individuals who were never registered for a presidential election are excluded ($N=1,447$). Color code: Participation: *Black/Green*. Abstention: *Dark Gray/Orange*. Not registered: *White/White*

However, the effect that we are most interested in, that of the position in the household, is significant in both cases. In addition, ANOVA shows that correlations inside households are extremely strong and significant in both cases. This legitimizes the definition of trajectories that we have chosen in Buton et al. (2012) and in the first parts of this chapter. Considering participation trajectories as a whole, including a mix of local and national elections, is certainly a simplification and should not prevent us from studying, for example, the group of voters who participated more in national elections; however, the effects of cohorts of registration, household homogeneity and position in the household seem to have been at play whatever the type of election.

In order to assess this more precisely, we have just begun to analyze the specific case of presidential elections. Due to the prominence of these elections, it makes sense—from the point of view of the researcher, and possibly even if we think of more subjective “voting careers”—to extract our eight presidential ballots, for four elections, and to study them as specific trajectories (for a similar attempt at a macro level, see Bélanger et al. 2012). Figure 10.6 gives one possible representation of these quite simple but potentially interesting trajectories. It shows that there are few abstentions and even fewer constant abstentions for this sort of ballots. Voting only for the first or for the second ballot is not very common either. 3–7% of voters only participated in the first, 4% to (in 2002) 15% only in the second. It is an interesting pattern, as strategic discussions on voting often provide good reasons to do so, due to the too wide or too narrow political offer in each case. Abstention at only one of

the two ballots of a presidential elections thus could be thought of as more “political”, less “social” than other abstentions (on these concepts, see Subileau and Toinet 1993).

However, if this distinction makes sense, our results show that “political” decisions on turnout are a minority. For 77% of our population, presidential voting patterns only included a mix of non-registration and participation and, in a few cases, abstention at the two ballots. In such cases, looking for other contextual effects on participation, be they social or temporal, is pointless.

We will however focus our further investigations on the other 23%, those who sometimes changed behaviors between ballots. This rarely seemed to be a consistent behavior: the specific context of the moment of one election seems to have been more at play here than individual longstanding preferences or habits begun in previous elections. 88 voters participated only in the first ballot, only on one occasion, while 19 did it several times without doing the reverse or abstaining for both ballots; 169 voters participated only in the second ballot, only one time (half of them in 2002), while only six did it more consistently. If there is some sort of political strategy behind such behaviors, it is not independent of the specific context of each election. What seems much more promising is thus to look into household patterns in presidential election trajectories. A cursory look at them shows that some of the specific inter-ballot patterns are indeed found consistently in some households, where we could hypothesize the existence of shared, maybe even collectively discussed strategies; but there are also cases when voters in the same household have opposite behaviors for the same election.

The presidential election is seen by many as the case when voters are most likely to make conscious choices according to political information provided beyond the borders of their household, due to the wide mass-media coverage. It is therefore a limiting case for us, as we are primarily interested in the weight of more local contexts. Our preliminary investigation does not point in the direction of decontextualized and/or purely “political” behaviors. For most voters, voting in a presidential election mostly appears as a taken-for-granted behavior; for smaller numbers, abstaining has the same status. One of our next steps will be to focus on household patterns in presidential election participation and to try to make sense of them, in order to better understand the minority of non-standard patterns.

Acknowledgments We thank all the participants in the LaCOSA conference for valuable suggestions and Richard Nice for an excellent editing of our English. This research has been funded by the French National Agency for Research, PAECE program.

References

- Abbott, A. (2001). *Time matters: On theory and method*. Chicago: University of Chicago Press.
- Abbott, A. (2005). The historicality of individuals. *Social Science History*, 29(1), 1–13.
- Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians’ careers. *American Journal of Sociology*, 96(1), 144–185.

- Bélanger, É., Cautrès, B., Foucault, M., Lewis-Beck, M. S., & Nadeau, R. (2012). *Le vote des Français de Mitterrand à Sarkozy: 1988-1995-2002-2007*. Paris: Presses de Sciences Po.
- Blanchard, P. (2012). Analyse séquentielle et carrières militantes. Research Report. <http://hal-archives-ouvertes.fr/hal-00476193>. Accessed 13 June 2013.
- Blossfeld, H.-P., & Rohwer, G. (2002). *Techniques of event history modeling: New approaches to causal analysis*. Mahwah: Lawrence Erlbaum Associates.
- Braconnier, C., & Dormagen, J.-Y. (2007). *La démocratie de l'abstention. Aux origines de la dé-mobilisation électorale en milieu populaire*. Paris: Gallimard.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Chapman & Hall.
- Buton, F., & Mariot, N. (Eds.). (2009). *Pratiques et méthodes de la socio-histoire*. Paris: PUF.
- Buton, F., Lemercier, C., & Mariot, N. (2012). The household effect on electoral participation. A contextual analysis of voter signatures from a French polling station (1982–2007). *Electoral Studies*, 31(2), 434–447.
- CEPEL. (2009). ANR “PEACE”. <http://www.cepel.univ-montp1.fr/spip.php?article231&lang=en>. Accessed 13 June 2013.
- Demazière, D., & Jouvenet, M. (Eds.). (forthcoming). *La sociologie d'Andrew Abbott*. Paris: Éditions de l'EHESS.
- Denni, B., & Bon, F. (1978). Population électorale, population électorale potentielle, population totale dans la région Rhône-Alpes. *Revue française de science politique*, 28(6), 1055–1066.
- Girard, V. (2011). Quelles catégories de classement pour l'analyse localisée de la représentation politique? Le cas des techniciens élus au sein d'un territoire industriel. *Terrains & Travaux*, 19, 99–119.
- Greenacre, M. J., & Blasius, J. (1994). *Correspondence analysis in the social sciences: Recent developments and applications*. London: Academic Press.
- Héran, F. (2004). Voter toujours, parfois... ou jamais. In B. Cautrès & N. Mayer (Eds.), *Le nouveau désordre électoral. Les leçons du 21 avril 2002* (pp. 351–367). Paris: Presses de Sciences Po.
- Huckfeldt, R. (1986). *Politics in context. Assimilation and conflict in urban neighborhoods*. New York: Agathon Press.
- Johnston, R. J., Jones, K., Propper, C., Sarker, R., Burgess, S., & Bolster, A. (2005). A missing level in the analyses of British voting behaviour: The household as context as shown by analyses of a 1992–1997 longitudinal survey. *Electoral Studies*, 24(2), 201–225.
- Lancelot, A. (1968). *L'abstentionnisme électoral en France*. Paris: Armand Colin.
- Lemercier, C., & Zalc, C. (2008). *Méthodes quantitatives pour l'historien*. Paris: La Découverte.
- Lesnard, L. (2010). Cost setting in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3), 389–419.
- Mariot, N., & Zalc, C. (2010). *Face à la persécution. 991 Juifs dans la guerre*. Paris: Odile Jacob.
- Palloni, A., Massey, D. S., Ceballos, M., Espinosa, K., & Spittel, M. (2001). Social capital and international migration: A test using information on family networks. *American Journal of Sociology*, 106(5), 1262–1298.
- R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Rosental, P.-A. (1999). *Les sentiers invisibles: Espaces, familles et migrations dans la France du 19^e siècle*. Paris: Éditions de l'EHESS.
- Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, 40(3), 471–510.
- Subileau, F., & Toinet, M.-F. (1993). *Les chemins de l'abstention. Une comparaison franco-américaine*. Paris: La Découverte.
- Verba, S., Lehman Schlozman, K., & Burns, N. (2005). Family ties. Understanding the intergenerational transmission of political participation. In A. S. Zuckerman (Ed.), *The social logics of politics. Personal networks as contexts for political behaviour* (pp. 95–114). Philadelphia: Temple University Press.

Chapter 11

Governance Built Step-by-Step: Analysing Sequences to Explain Democratization

Matthew Charles Wilson

Introduction

A significant portion of political science research is devoted to understanding the causes and consequences of democratization. The drive to do so is compelled, in part, by the need to uncover causal mechanisms that explain why democracies do so well. Democracies are less likely to fight international conflicts but more likely to be victorious (Bennett and Stam 2000). They are also bastions of domestic freedoms and guarantees when it would otherwise seem disadvantageous to observe them (Acemoglu and Robinson 2005).

Despite a number of generalizations about when and why democracies emerge, however, research that might explain it is held back by methodological limitations stemming from unique forms of time-dependence (Grzymala-Busse 2011; Mahoney 2001, Pierson 2004; Page 2006). Grzymala-Busse (2011) and others have called attention to the need to highlight specific forms of time-dependence in the development of models of political outcomes, such as tempo, duration, and order. A critical endeavor on the political science agenda is to think more critically about ways to account for order. In the study of democratization, such endeavors might shed light on previously undetected ways in which past transitions affect the emergence of democracy.

This chapter addresses the nature of political sequences by conceiving of prior or regime histories as principal determinants of the timing of democratization. In support of differentiating unique temporal mechanisms, I provide an application of sequence analysis to regime type. Using techniques of visualization and comparison that are standard to the approach, I demonstrate how one might draw on a country's institutional history to explain the emergence of stable democracy. In so doing, I exemplify ways that one might go "beyond duration" to explain the time-dependent nature of democracy and make a substantial contribution in political science literature.

M. C. Wilson (✉)
The Pennsylvania State University, Pennsylvania, USA
e-mail: mcw215@psu.edu

Theory

Though different notions of democracy exist, democracy is principally understood to be a system of representative government in which the right to contest and participate in elections is open to virtually all members of society (Dahl 1971). Transitioning to democracy requires the creation of rules, values, and leadership necessary to install participatory and competitive institutions. In non-democratic settings, such a process can be lengthy and protracted, provoking questions regarding *how long* it takes democracy to flourish and what are the necessary preconditions. Suggested mechanisms by which democratization takes place include changes in political culture (Almond and Verba 1963); modernization (Lipset 1959); elite unity (Brownlee 2007); socioeconomic structure (Moore 1966); civil society (Almond and Verba 1963); institutions (Shugart and Carey 1992); path dependency (Collier and Collier 1991) and strategic choices (Acemoglu and Robinson 2005). Theories about the emergence of democratization fall into three related streams centered on the economy, political institutions, and bargaining.

The first view emphasizes economic conditions that favor democracy and political arrangements that follow from economic decisions. On the one hand, economic development is the cause of democracy—leading to a more equal distribution of assets and education and increased capital and political mobility (Huntington 1993; Lipset 1959). On the other hand, politics changes occur in tandem with a developing economy but do not directly result from it—a diminished agrarian influence, the breakdown of traditional roles, and the rise of the working class form simultaneous pressures for democratic representation (Anderson 1983; Boix 2003; Levi 1989; Moore 1966; North 1981). The difference between hands separates an endogenous explanation from an exogenous explanation for democracy (Przeworski 1991; Przeworski et al. 2000).

Institutional theories of democratization explain democracy as the conception of a particular set of institutions. Based on Dahl's (1971) *minimal* definition of democracy—contestation and competition—three pathways mark the transition of a fully closed authoritarian regime to democracy. Dahl asserted that political liberalization followed by heightened citizen participation was best for the establishment of successful democracy, but that this path was no longer a viable option for many authoritarian regimes which constitute hegemonic but broadly inclusive regimes. A similar argument can be found in Huntington (1993), the primary thesis of which is that instability is the product of the rapid mobilization of new groups coupled with the slow development of political institutions.

If certain institutions can engender democracy, they may also halt the transition process—either by undermining democracy or by serving a different intent. For example, it has been argued that neopatrimonialism has given transitions in sub-Saharan Africa a unique character. Presidentialism, clientelism, and the use of state resources are three “albeit informal... stable, predictable, and valued” institutions in African neopatrimonial regimes (Bratton and Van de Walle 1997). Suspicions that residual institutions can forestall democracy are also validated by the transitions

between military regimes in Latin America, where coups seem to beget more of the same (Nordlinger 1977). What is more, the emergence of seemingly democratic institutions may merely provide a basis of support for “hybrid regimes” or “electoral autocracies” (Levitsky and Way 2002; Gandhi 2008; Magaloni 2007; Haber 2006). It is thus not fully clear what roles specific institutions play in the emergence and consolidation of democracy, as *opposed* to autocracy.

Lastly, bargaining theories assert that democracy emerges as a result of a set of struggles over how resources ought to be distributed. Boix’s (2003) model of redistribution, for example, is based on assumptions that the poor cannot commit to lower the level of taxation and that the rich cannot avoid revolutions by promising to redistribute in the future. Because *de facto* power is transitory, the poor use their bargaining position to secure institutional guarantees from the elites that ensure future gains. Democracy emerges to relegate the future allocation of power (Acemoglu and Robinson 2005). The lack of real choice between competitors holds exit at bay in the form of passive loyalty, making the use of voice a more attractive option (Hirschman 1970). In turn, the effectiveness of voice depends on the credibility of exit. Under different circumstances, citizens can be expected to choose to exit or to voice their dissatisfaction over a regime, with implications for whether and when the government accedes to demands for reform.

Competing theories that explain the development of democracy are neither fully exclusive, nor are they inclusive. Endogeneity is a central concern of empirical tests of democratization. Scholars nevertheless disagree about proximate causes of democracy. Meanwhile, a number of scholars have speculated about how the *order* of such factors affects the timing of democracy (Dahl 1971; Huntington 1993; Moore 1966; O’Donnell and Schmitter 1986). There are also abundant references to complex sequences in political science (Mahoney 2001; Mansfield and Snyder 2007; Carothers 2007).

Despite having different theoretical foundations, various explanations that might account for the emergence of democracy share in common a focus on role of history and intervening factors. What some of the best qualitative scholars have identified is the unique role of specific events and *their interaction* with historical events in shaping the long-term political prospects of a country. To this end, political science research is primed for sequence analysis.

There are several innovative methodological approaches to general forms of time-dependence, though they are not well-suited for detecting order. One is to include one or multiple lags for the explanatory variable according to some length of time. This approach is limited insofar as it is often unclear how long a purported effect should last. Testing for and including higher-order lags can also exhaust the theoretical and methodological limits of a model that is contingent on discreet, time-dependent states (Wilson and Butler 2007). Another approach is to construct an event-history model, although the outcome in a duration model is a function of the duration of time rather than of order and sequence. A third approach is to aggregate the accumulated number of state occurrences and divide it by the span of time between changes in the dependent variable. This quickly becomes complicated if multiple events are in question.

In short, many existing approaches to time dependence account for time dependence, but not for the effect of order and sequence on the distribution of errors. They are better suited for uncovering mechanisms that occur within *path* dependent processes than for *path* dependent processes (Page 2006). Sequence Analysis, which involves calculating a measure of similarity, or distance, between pairs of sequences, lends analytical insights to such processes.

Sequencing Political Events

Operationalization

As a first-pass for analyzing political sequences, I consider whether historical sequences of regime types constitute antecedents of democracy. A regime is herein defined as a unique set of procedural institutions—formal or informal rules—that determine political access and which are accepted by major political actors (Gasiorowski 1996; Kitschelt 1992; Munck 1996). Such configurations include, but are not limited to, elections, militaries, and royal ascendance. In the authoritarian and limited-democratic setting where the government is more intrinsically connected to its leader(s), regime types are also good proxies for the type of leader and the conditions under which they and their opposition bargain (Gleditsch and Ward 1997; Bratton and van de Walle 1997). The sequential history of transitions between regime types is therefore a suitable addition to competing models of regime change.

To capture the role of authoritarian eras as precursors to democracy, I rely on a discrete classification of regime type. The Cheibub et al. (2010) coding scheme was based on the dichotomous classification of democracies and dictatorships introduced in Przeworski et al. (2000). It was a classification of all independent regimes for the post World War II period, covering approximately 202 countries over the years 1946–2008 (or the date of leader death or regime change). The authors characterized democracies as contested elections which occur at regular intervals, the outcome of which is not known prior and the winner of which actually assumes office. The authors relied on four rules: (1) the executive must be chosen by a popular election; (2) the legislature must also be popularly elected; (3) more than one party must compete in the election; and (4) alternation in power under electoral rules must occur (2010, p. 69).

Of the regimes that satisfy the four criteria for a democracy, Cheibub et al. (2010) distinguished between presidential, mixed, and parliamentary democracies. They coded non-democracies as monarchial, military, or civilian regimes. Monarchies are regimes based on family and kin decision-making, in which the executive bears an imperial title and legitimizes a hereditary successor or a predecessor. Military regimes are characterized by the leadership of a current or previous member of the armed forces, although they do not include regimes borne out of guerilla movements. Civilian regimes represent a “residual” dictatorship category in which the leader wields neither hereditary nor military power.

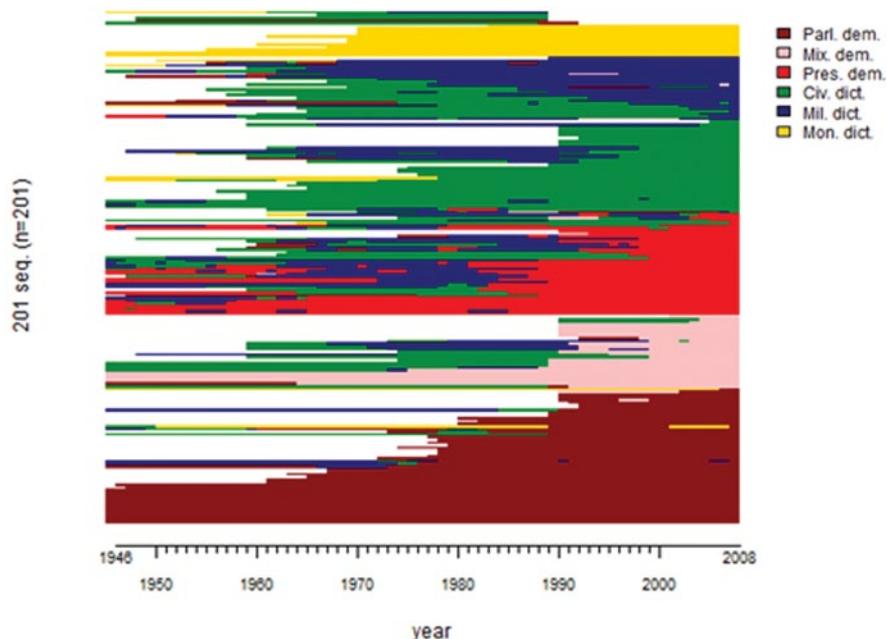


Fig. 11.1 Sequence Index plot: regime type by year

Figure 11.1 shows regime types by year over the period 1946–2008 for 201 countries and territories. The sequence index plot is sorted by regime type at the end of the observation period. Each horizontal line in the plot refers to the regime history of a particular country. As can be seen in the index plot, there are several distinct periods in which independent countries entered the international system. A number of countries that had an authoritarian spell existed before 1946 and lasted until 2008. There was also a wave of countries that emerged in the 1960s, many of which began as civilian dictatorships and occurred as a result of decolonization. The 1990s also saw an influx of new countries, including a new cohort of civilian regimes.

Of the 201 countries and territories, 84 entered the sample already a democracy. Sixty-seven countries were consistently democratic across the window of observation, while 17 experienced eventually transitioned back to autocracy. Of the 134 non-left-censored countries, nearly half (69) experienced a democratization episode between 1946 and 2008. This comprises 98 democratization attempts, 58 of which lasted five years or more. As the figure shows, most of the consolidated democracies that did not enter the sample already a democracy are a product of the mid-1980s and 1990s. Before then, new democracies were likely to last a shorter period of time than their post-1990 counterparts.

Though difficult to visually discern from the sequence index plot, mixed democracy did not emerge until the last two decades, which also saw an increase in

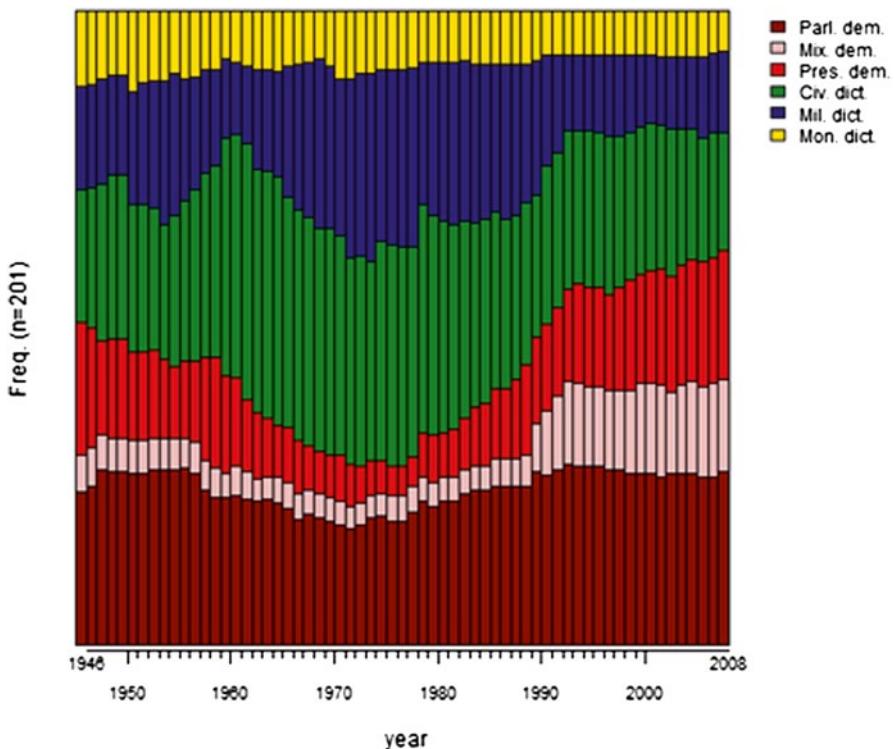


Fig. 11.2 Density plot: regime type by year

presidential democracies. As an alternative visualization, Fig. 11.2 shows how the density of states (regime type) changes by year. Among countries that were not already stable democracies, parliamentarism—the most common type of democracy in the left-censored cases—experienced a slight dip between 1965 and 1990. Presidentialism exhibited a similar pattern over the same period, but was much more dramatic.

Mixed forms of democracy were almost non-existent until the 1990s, after which it became more abundant. As it regards consistently democratic countries, mixed democracy declined regularly until 1980. Among autocracies, civilian dictatorships are highly consistent, experiencing decreases in the late 1980s and less-pronounced decreases in the late 2000s. Military dictatorships occurred in two distinct waves, one in the 1950s and one between 1960 and 1990. Monarchies remained relatively stable over time, but with a slight decrease. Overall, civilian dictatorship was the most common regime type in the sample of democratizers. There is considerable cross-national variation in regime type, but the sequences are not so distinct so as to suggest that dominant patterns cannot be found in the data. They are thus particularly suitable for sequential analysis.

Alignment

In comparing sequences, I make several assumptions about the nature of my data. First, I assume that although political transitions occurred simultaneously on the world stage, the sequences that they make up occurred independently of one another. That is to say, transitions between regimes are more likely to stem from conditions specific to the country in question than to neighborhood or ‘diffusion’ effects (Bratton and Van de Walle 1997). Nevertheless, one could account for diffusion by controlling for geographical regions or political linkages.

I also do not require the political sequences to correspond to a standard length. Blanchard (2005) standardizes sequences to the same length following a proportionality rule. Though distance calculations can reflect the uneven distribution of sequence length, such non-proportionality can be theoretically useful. The effect of uneven sequence length serves to distinguish newer countries from older ones, as the democratization process may be affected by the time-period in which a country emerged (Huntington 1993).

The core of Sequence Analysis lies in the algorithm for distance comparison. Optimal Matching (OM) generates edit distances that are the minimal cost, in terms of insertions, deletions and substitutions, for transforming one sequence into another.¹ The insertion/deletion (‘in-del’) cost is a single value; the substitution cost can be a single value (a constant) or a set of weighted values. I selected an in-del cost of 1 and a constant substitution cost of 2. These are the default OM values used by TraMineR, and values for which the OM distance is the same as LCS distance (Gabadinho et al. 2010, pp. 98).

Among the many applications of the distances calculated by the OM algorithm, clustering methods can be used to aggregate the sequences into a reduced number of groups. I clustered the data based on complete linkage, which sets a threshold based on the largest dissimilarity between points in two adjacent clusters and is also called the furthest-neighbor method². One question concerns how many clusters are appropriate for the data. Figure 11.3 illustrates the nature of cluster assignment based on complete-method clustering as I increase the number of clusters from 1 to N/2 (100). Grayscale changes in the plot denote changes in cluster assignment based on complete linkage. As the figure shows, complete-method clustering isolates the most distant sequences into separate clusters. It is thus particularly conducive to identifying outliers. This unique property is demonstrated by entropy (overlaid), which indicates the ease of identifying cluster membership for a sequence randomly selected from the population. The marginal impact of increasing clusters is initially large and diminishes in a logarithmic fashion.

I grouped regime histories into six clusters, based on the six regime types recognized by Cheibub et al. (2010). This allows me to ascertain whether countries’

¹ The Optimal Matching (OM) edit distance was first proposed by Levenshtein (1966) and has been popularized in the social sciences by Abbott (Abbott 1995, 2001).

² Documentation for the ‘Cluster’ package in R explains clustering methods in greater detail. There are several programs and packages created for Sequence Analysis; the analysis presented herein utilizes the TraMineR package in R 2.13.0 (Gabadinho et al. 2010).

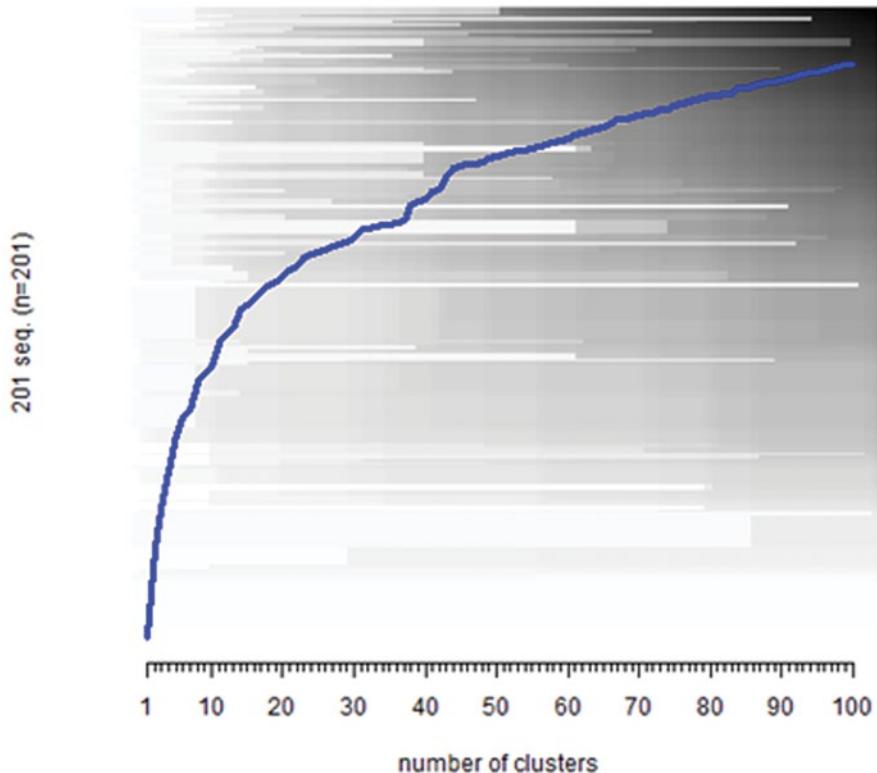


Fig. 11.3 Cluster assignment by cluster number, with entropy (complete linkage)

similarities are dominated by the type of regime that they “know best,” historically speaking. Though exploratory, my use of six clusters appears to be an appropriate choice—beyond six clusters, the marginal impact of an additional cluster affects fewer than $N^{1/2}$ (the square root of N) sequences.³

Results

Figure 11.4 shows the results of clustering the regime histories into six groups using the complete (furthest neighbor) method. Of the regime types that Cheibub et al. (2010) identify, five dominate in specific clusters. Cluster 1 is predominately

³ Taking the square root of a sample is a well-known rule-of-thumb for calculating the optimal number of bins in a histogram. Hypothetically, each bin would therefore contain a number of observations equal to the square root of the sample. The observation that the marginal impact of an additional cluster affects fewer than this number suggests that—just as an additional bin would be unnecessary—an additional cluster is unnecessary.

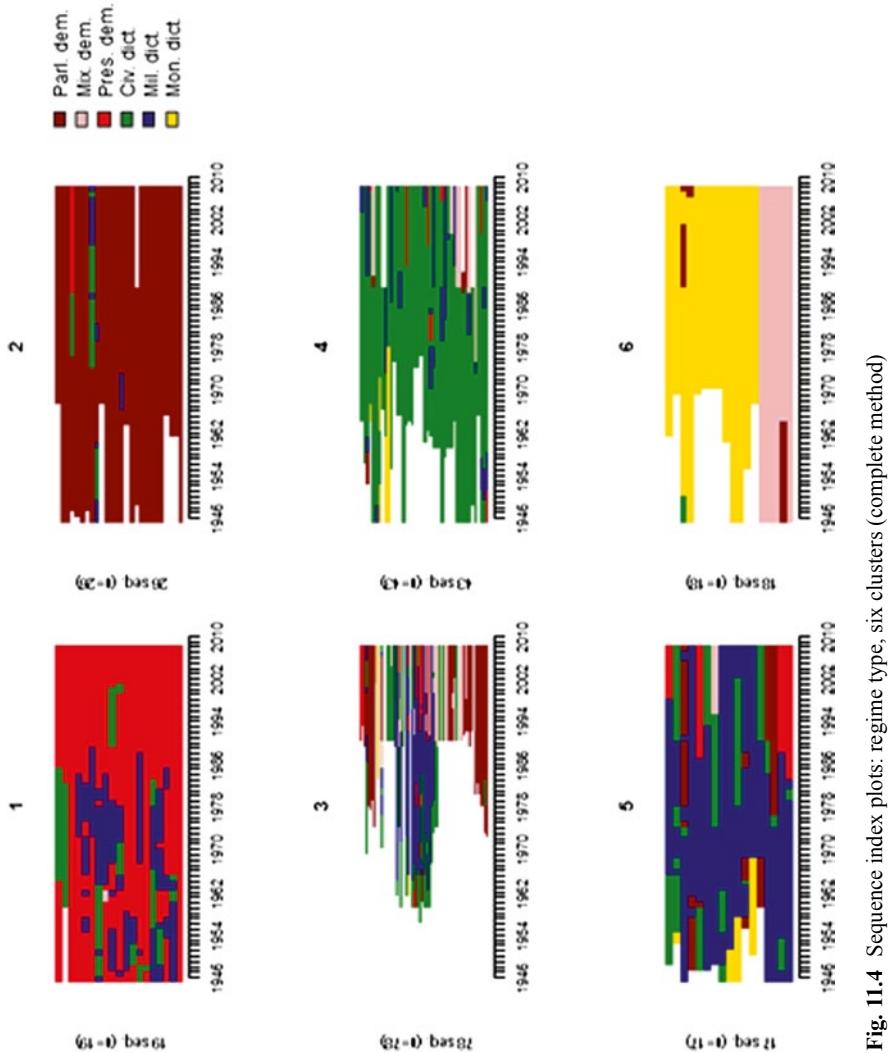


Fig. 11.4 Sequence index plots: regime type, six clusters (complete method)

Table 11.1 Regime transition rates, 1946–2008

	Parl. Dem	Mix. dem.	Pres. dem.	Civ. dict.	Mil. dict.	Monarchy
Parl. dem.	0.987	0.002	0.000	0.001	0.009	0.000
Mix. dem.	0.005	0.985	0.002	0.003	0.006	0.000
Pres. dem.	0.000	0.001	0.968	0.006	0.026	0.000
Civ. dict.	0.004	0.004	0.007	0.959	0.026	0.001
Mil. dict.	0.006	0.007	0.024	0.029	0.934	0.000
Monarchy	0.004	0.000	0.000	0.008	0.005	0.982

composed of presidential democracies, most of which were interrupted by military dictatorship. Cluster 2 is primarily parliamentary democracies. Cluster 4 is contains long-term civilian autocracies, and cluster 5 durable military dictatorships. Mixed democracy is rare among new democratizers. As a consequence of their infrequency, mixed-democracies are grouped with monarchies in cluster 6. The remaining cluster is reserved for regimes emerging after 1960. Omitting left-censored cases does not change how remaining sequences are organized.

The clusters contain useful information for the study of democratization and regime change. First, they visually confirm that stasis is the norm for regimes. Regimes tend to endure, but to a point. Of the sample of countries which had at least one autocratic spell during the temporal window of my observation, party-based autocracies are statistically the longest-lasting. Though crude, the clusters are particularly informative for settling arguments about which type of democracy is most stable.

There is a longstanding comparative debate over the relative advantages and disadvantages of the constitutional frameworks of parliamentary and presidential democracies, and their long-term performance (e.g., stability). Linz and Valenzuela (1994), for example, argue that presidentialism is more prone to breakdown than parliamentarism because it lacks the confidence vote that allows the government to be removed from office while keeping the constitution intact. They also argue that presidentialism encourages political crises in the first place, namely by aggravating the relationship between the executive and members of the legislature (Carey 2005). Cheibub (2006) explains it as presidential being more fragile because they tend to follow from military dictatorships, and that fragility is not due to presidentialism but to contextual issues.

Strikingly, there are very few presidential democracies that *do not* follow from military dictatorship, which is evident in the way that they cluster. Parliamentary democracies are not more likely to follow from a particular type of dictatorship over another (Table 11.1). The clusters thus support Cheibub (2006)'s argument by demonstrating a dominant *order* in regime change that is unique to presidential democracy but not to parliamentary democracy. What is more, of those democracies that emerged after 1946, presidential democracies are more likely to last than parliamentary democracies (3.945 years versus 8.691 years, respectively).

As a form of case-study selection, I compared sequences in a cluster to the modal sequence, or a series comprised of the most frequent (modal) states at each position. Figure 11.5 shows representative cases for sequences clustered with and without left-censored democracies. It confirms that cluster 1 is best represented by durable

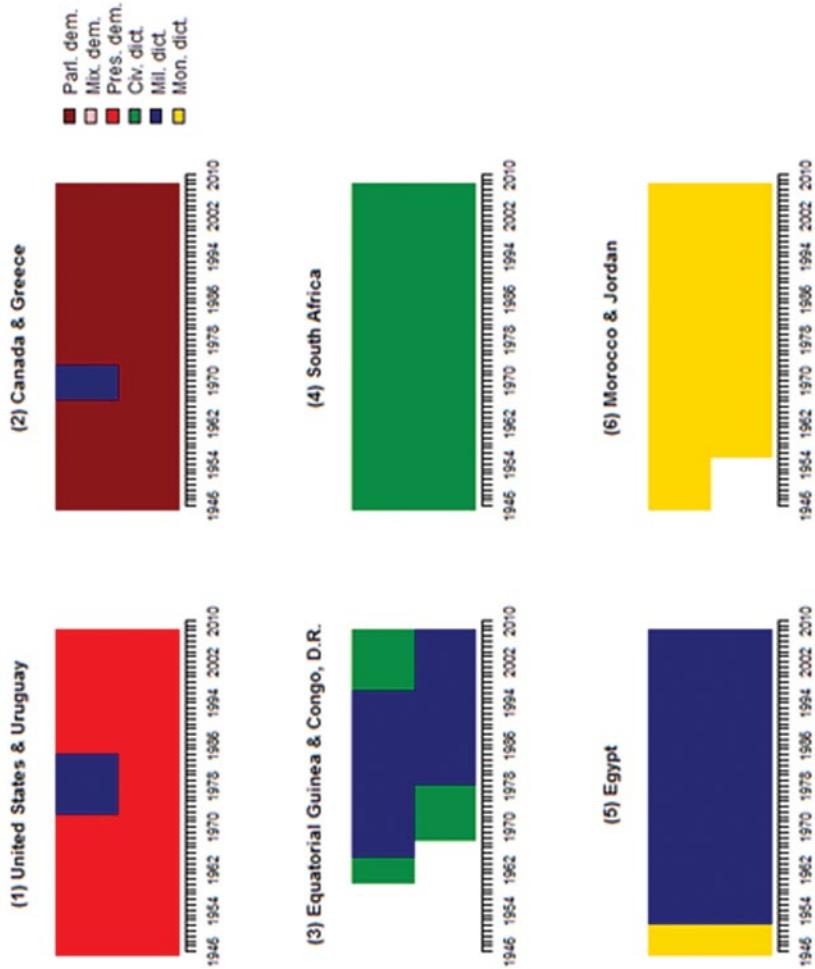


Fig. 11.5 Best-fitting cases, with and without left-censored sequences

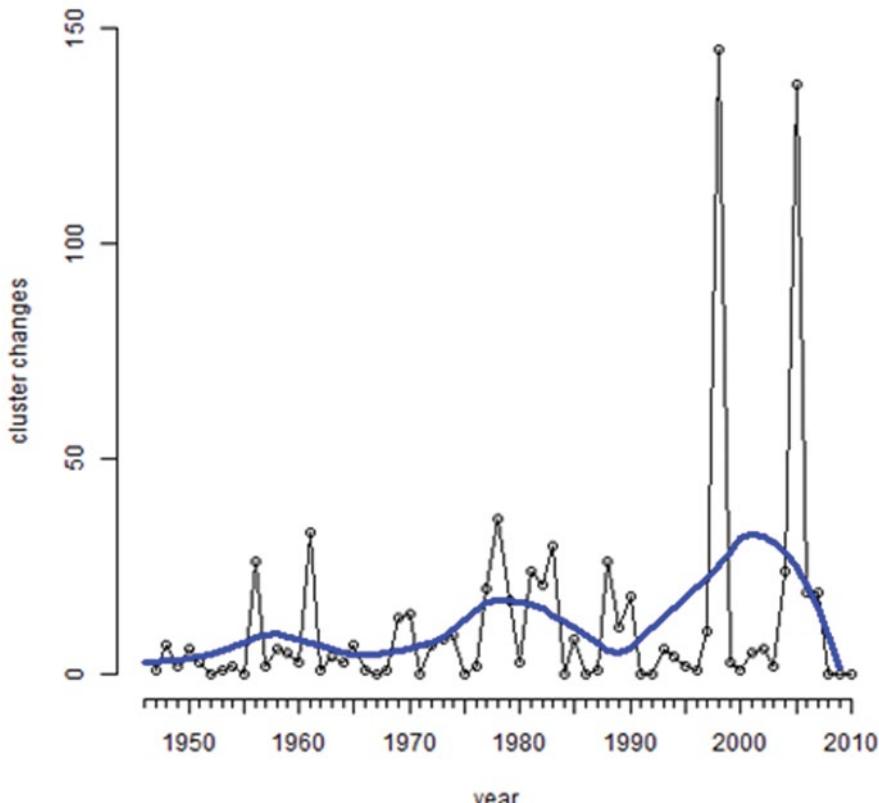


Fig. 11.6 Cluster changes, by year

presidential democracies such as the United States or by interrupted presidential democracies such as Uruguay. Canada and Greece are similar examples based on parliamentary democracy which comprise cluster 2. Clusters 4, 5, and 6 are best represented by durable civilian autocracy in South Africa, durable military dictatorship in Egypt, and durable monarchies such as those in Morocco and Jordan, respectively. Cluster 3 is represented by transitional autocracies such as Equatorial Guinea and the Democratic Republic of Congo.

Lastly, I re-clustered regime sequences by year, allowing clusters to vary over time. This treats clusters as latent classes, insofar as the classes are dominant sequence patterns that become more evident with time. Cluster membership is fairly stable over time. Nevertheless, the switching between clusters over time—shown in Fig. 11.6—supports the notion that democratization has occurred in three “waves” in the last 60 years (Huntington 1993). Three observed “eras of independence” transformed cluster assignment and affected the consequent clusters that represent long-term regime histories.

Discussion

It is beyond the scope of this chapter to enumerate the ways in which sequence analysis enables political scientists to unpack order, nor to fully interpret the impacts of regime history on the timing of democratization and the emergence of democracy. In duration models predicting the time to democracy lasting five or more years, however, there is a statistically distinguishable difference between clusters derived from long-term regime histories (Wilson 2012). In particular, countries with protracted experience with civilian autocracy are significantly less likely to democratize. Short-lived military regimes are much more likely to yield democracy but are also more interruptive. Controlling for past democratization attempts, countries with an extensive history of civilian autocracy are more likely than militarized countries to consolidate (Wilson 2012). The findings suggest that democratization is not a first-order process, but one which is affected by events farther back in the past.

Sequence analysis offers a number of appealing statistical properties that can advance political science research on questions of order and pattern. As an approach, it offers unique ways to quantify and visually discern patterns of discrete information such as regime type. For one, it allows social scientists to incorporate discrete data into empirical models, thereby bridging the gap between qualitative and quantitative research agendas. The logic underlying cluster assignment over time is also amenable to latent class analysis and latent transition models. What is more, the decisions that undergird sequence construction, matching costs, and alignment can be changed without sacrificing degrees of freedom in an empirical model.

There are thus a number of theories about order and sequence in political science that can benefit from expansion of the comparison and visualization tools provided by sequence analysis. In particular, sequence analysis can substantially aid the study of democracy and democratization. Questions that the approach might help to answer include: *Are there common patterns of regime change? What is the most common mode of transition to democracy? Do particular sequences of breakdown and replacement bode well for democracy, or the type that emerges?*

Political science research remains somewhat limited in the ability to empirically test the impact of unique sequences on political outcomes. As a first glance at how political data can be used to shed light on the timing of democracy, this chapter takes up the task of comparing regime histories. Using optimal matching and constant substitution costs on a set of regime-type data for 1946–2008, I clustered countries into six groups based on complete linkage (furthest neighbor method). In combination with standard visualization techniques, the clusters convey a great deal of information about the emergence, duration, and transition of regimes. They also enable one to clearly identify “waves” of independence and democratization in the international system. Furthermore, the data show a striking relationship between presidential democracy and military autocracies. The suggested pathology is relevant to ongoing debates about the institutional stability from different forms of democracy.

Though many assert it as an important research agenda, whether and to what extent “time matters” in political science is a question for individual researchers to answer (Abbott 2001; Grzymala-Busse 2011; Pierson 2004; Page 2006). There are nevertheless promising avenues of research provided by sequence analysis to allow them to delve deeper into nuanced temporal features of politics. As I demonstrate, the approach provides relatively straightforward ways to operationalize order and bring old arguments to the forefront for testing. Quite possibly, there are few areas of political science that could not benefit from the step-by-step application of sequence analysis and methodological extensions.

References

- Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21, 93–113.
- Abbott, A. (2001). *Time matters: On theory and method*. Chicago: University of Chicago Press.
- Acemoglu, D., & Robinson, J. (2005). *Economic origins of dictatorship and democracy*. New York: Cambridge University Press.
- Almond, G., & Verba, S. (1963). *The civic culture*. Princeton: Princeton University Press.
- Anderson, B. R. (1983). *Imagined communities: Reflections on the origin and spread of nationalism*. London: Verso.
- Bennett, D. S., & Stam, A. (2000). EUGene: A conceptual manual. *International Interactions*, 26, 179–204.
- Blanchard, P. (2005). Multi-dimensional biographies. Explaining disengagement through sequence analysis. 3rd general conference of the European consortium for political research, Budapest, 8–10 September.
- Boix, C. (2003). *Democracy and redistribution*. Cambridge: Cambridge University Press.
- Bratton, M., & van de Walle, N. (1997). *Democratic experiments in Africa*. New York: Cambridge University Press.
- Brownlee, J. (2007). *Authoritarianism in an age of democratization*. Cambridge: Cambridge University Press.
- Carey, J. (2005). Presidential versus parliamentary government. In C. Menard & M. Shirley, (Eds.), *Handbook of new institutional economics* (pp. 91–122). Boston: Kluwer Academic Press.
- Carothers, T. (2007). The sequencing fallacy. *Journal of Democracy*, 18(1), 13–27.
- Cheibub, J. A. (2006). *Presidentialism, parliamentarism, and democracy*. Cambridge: Cambridge University Press.
- Cheibub, J. A., Gandhi, J., Vreeland, J. R., et al. (2010). Democracy and dictatorship revisited. *Public Choice*, 143(1), 67–101.
- Collier, R. B., & Collier, D. (1991). *Shaping the political arena: Critical junctures, the labor movement, and regime dynamics in Latin America*. Princeton: Princeton University Press.
- Dahl, R. (1971). *Polyarchy: Participation and opposition*. New Haven: Yale University Press.
- Gabadinho, A., Ritschard, G., Studer, M., Müller, N. S., et al. (2010). Mining sequence data in R with the TraMineR package: A user’s guide. <http://mephisto.unige.ch/traminer/>. Accessed 1 June 2010.
- Gandhi, J. (2008). *Political institutions under dictatorship*. New York: Cambridge University Press.
- Gasiorowski, M. J. (1996). An overview of the political regime change dataset. *Comparative Political Studies*, 29(4), 469–483.
- Gleditsch, K. S., & Ward, M. D. (1997). Double take: A reexamination of democracy and autocracy in modern polities. *Journal of Conflict Resolution*, 41(3), 361–383.

- Grzymala-Busse, A. (2011). Time will tell? Temporality and the analysis of causal mechanisms and processes. *Comparative Political Studies*, 44(9), 1267–1297.
- Haber, S. (2006). Authoritarian government. In B. Weingast & D. Wittman (Eds.), *Oxford handbook of political economy* (pp. 693–707). Oxford: Oxford University Press.
- Hirschman, A. (1970). *Exit, voice, and loyalty*. Cambridge: Harvard University Press.
- Huntington, S. (1993). *The third wave*. Norman: University of Oklahoma Press.
- Kitschelt, H. (1992). Political regime change: Structure and process-driven explanations? *American Political Science Review*, 86(4), 1028–1034.
- Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10, 707–710.
- Levi, M. (1989). *Of rule and revenue*. Berkeley: University of California Press.
- Levitsky, S., & Way, L. (2002). The rise of competitive authoritarianism. *The Journal of Democracy*, 13(2), 51–65.
- Linz, J. J., & Valenzuela, A. (1994). *The failure of presidential democracy II: The case of Latin America*. Baltimore: Johns Hopkins University Press.
- Lipset, S. M. (1959). Some social requisites of democracy. *American Political Science Review*, 53(1), 69–105.
- Magaloni, B. (2007). *Voting for autocracy: Hegemonic party survival and its demise in Mexico*. New York: Cambridge University Press.
- Mahoney, J. (2001). *The legacies of liberalism: Path dependence and political regimes in Central America*. Baltimore: John Hopkins University Press.
- Mansfield, E. D., & Snyder, J. (2007). The sequencing Fallacy. *Journal of Democracy*, 18(3), 6–9.
- Moore, B. (1966). *Social origins of dictatorship and democracy*. Boston: Beacon Press.
- Munck, G. L. (1996). *Disaggregating political regime: Conceptual issues in the study of democratization*. The Helen Kellogg Institute for International Studies, Working Paper no. 228. <http://www-bcf.usc.edu/~munck/pdf/Munck%20Kellogg%201996.pdf>.
- Nordlinger, E. (1977). *Soldiers in politics: Military coups and governments*. Englewood Cliffs: Prentice Hall.
- North, D. (1981). *Structure and change in economic history*. New York: Norton.
- O'Donnell, G., & Schmitter, P. C. (1986). *Transitions from authoritarian rule: Tentative conclusions about uncertain democracies*. Baltimore: Johns Hopkins University Press.
- Page, S. E. (2006). Path dependence. *Quarterly Journal of Political Science*, 1, 87–115.
- Pierson, P. (2004). *Politics in time: History, institutions, and social analysis*. Princeton: Princeton University Press.
- Przeworski, A. (1991). *Democracy and the market: political and economic reforms in Eastern Europe and Latin America*. New York: Cambridge University Press.
- Przeworski, A., Alvarez, M., Cheibub, J. A., Limongi, F., et al. (2000). *Democracy and development: Political institutions and material well being in the world, 1950–1990*. Cambridge: Cambridge University Press.
- Shugart, M. S., & Carey, J. M. (1992). *Presidents and assemblies: Constitutional design and electoral dynamics*. New York: Cambridge University Press.
- Wilson, M. C. (2012). Temporal determinants of democracy: Beyond duration. Paper prepared for the Lausanne conference on sequence analysis, June.
- Wilson, S. E., & Butler, D. M. (2007). A lot more to do: The sensitivity of time-series cross-section analysis to simple alternative specifications. *Political Analysis*, 15(2), 101–123.

Part IV

Visualisation of Sequences and Their Use

for Survey Research

Chapter 12

Sequence as Network: An Attempt to Apply Network Analysis to Sequence Analysis

Ivano Bison

Introduction

What happens if we decide to plot sequences as graphs of a network? Can this approach increase our knowledge of the underlying structure common to the single sequences? Moreover, will this approach bring to the surface the structures, patterns, and careers still (perhaps) hidden to our eyes and knowledge?

At present, there are two main types of sequence representation. The first is based on the visualization of changes in the composition of events over time. Examples are state distribution plots and modal state sequences (Brzinsky-Fay et al. 2006; Widmer and Ritschard 2009; Gabadinho et al. 2011). The second type is the sequence index plot (Bison 1999; Bison and Esping-Andersen 2000; Scherer 2001), which translates a sequence into a colored line.

These two main ways to represent sequences have a series of limitations. State distribution plots and modal state sequences describe macro changes but not individual ones. Both display the changes over time in the marginal distributions of the phenomena under observation. These pictures are useful if one is interested in understanding the change over time in the composition of a given phenomenon, for example the proportion of employed, unemployed and first-time job-seekers in the labor market. Unfortunately, these graphs are inadequate if we are interested in understanding the structure of the career patterns. Put differently, state distribution plots and modal state sequences are one-dimensional graphs that furnish little information about micro changes and even less about how careers develop and about the paths followed by individuals and groups.

By contrast, such information is yielded in the sequence index plot. In this case, the level of detail is the most precise and accurate that can be achieved in the visualization of the sequences. It is possible to represent every single sequence in a single

I thank Jacques-Antoine Gauthier and the participants in the LaCOSA Conference for helpful comments.

I. Bison (✉)

Department of Sociology and Social Research, University of Trento, Trento, Italy

e-mail: ivano.bison@unitn.it

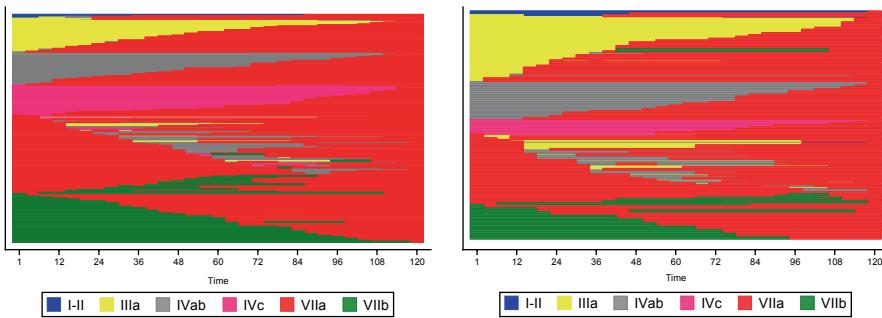


Fig. 12.1 Sequence index plot of working careers that end in class VIIa by gender

graph. This makes it possible to follow, instant by instant, the full unfolding of the sequence over time. Moreover, the physical proximity of each line to the other produces a second result: insight into the possible existence of common underlying patterns followed by multiple actors.

However, these are complex graphics and may mislead the researcher.¹ The main problem is that they are interpreted by means of the senses, which sometimes conceal things or show them in a different perspective. For instance, we may be visually attracted by some elements or some graphic structures or specific configurations of colors and shapes in the graph and neglect others. As a result, the graph is read and interpreted in one way rather than another.

For instance, suppose we want to identify graphically whether there are substantial differences in class careers between men and women. We select from the Italian Longitudinal Household Survey² (ILFI) the subset of men and women who, in the first 10 years of their working careers, changed classes at least once, and who, at the end of the tenth year, were in the urban working class³ (IIIb+V-VI+VIIa, henceforth VIIa). The data yield (Fig. 12.1) two sequence index plots (Brzinsky-Fay et al. 2006) depicting, respectively, the working careers of the Italian women and men who have changed their occupational class position at least once and who after 10 years are in the working class (VIIa).

The two graphs show some differences between men and women. For example, the proportion of women who start in class IIIa and after 10 years end up in class

¹ Interpretation of these graphs should always be combined with other measures that help the researcher to read the content.

² For information about the data used in this article, see Bison (2011).

³ The EGP class scheme used here is the follows: (I-II) Higher and lower-grade professionals, administrators, officials and higher-grade technicians; managers in large and small industrial establishments; large proprietors; supervisors of nonmanual employees; (IIIA) Routine nonmanual employees, higher grade (administration and commerce); (IVab) Small proprietors, artisans, etc., with and without employees; (IVc) Farmers and smallholders; other self-employed workers in primary production; (IIIb+V-VI+VIIa) Routine nonmanual employees, lower grades (sales and services); lower-grade technicians; supervisors of manual workers; skilled manual workers; semi-skilled and unskilled manual workers (not in agriculture, etc.); (VIIb) Agricultural and other workers in primary production.

VIIa is greater than that of men. Conversely, the proportion of men who start in the agricultural classes (IVc and VIIb) is greater than that of women. However, if we exclude the different proportions of classes of the first occupations and rule out that the different thicknesses of the lines are caused by the different sample sizes of men and women, other differences are not apparent.

To be precise, the differences found are irrelevant to our purposes. Furthermore, we could obtain this same information with simple frequency distributions and some tables. Men and women have both the same career patterns and also the same complex patterns, i.e., the patterns whereby respondents moved across multiple classes in the 10 years of observation.

The results from the charts are so clear that they oblige us to draw a single conclusion: except for *differences* at the beginning of their careers, already known in the literature, the career patterns of men and women who end up in the working class after 10 years are similar. The numbers of the subjects exhibit one or the other pattern changes by gender, but there is no change in either the career patterns or the structure and the concatenation of the temporal events.

The problem, however, is whether this conclusion is actually correct. What if it is wrong? Is, for instance, this conclusion owing to our inability to grasp what is depicted by the graph as a ‘whole’? Have we failed to understand that these changes of color from one state to another, apparently similar, describe different patterns for men and women? Furthermore, in addition to a common pattern, do there exist others patterns that we cannot see and that are characterized by a different temporal shape of events? Has the different temporal shape of events generated seemingly equal career patterns that in fact are very different? Finally, is it possible that the two groups are subject to different mechanisms that generate patterns that only apparently are similar but in fact are substantially different as regards timing and shape? These are just some of the doubts that have passed through my mind over the years and that have induced me to look elsewhere for new roads, new standpoints from which to view the object ‘sequence’.

The use of graphic visualization and the social network analysis suggested in this paper have two purposes. One is to find new ways to present results; the other is to gain a new perspective from which to observe sequences. To do this, however, we need to see sequences not as individuals moving from one state to another but as individuals who exhibit common career patterns.

Obviously, the main problem is how to bring out this common pattern. To date, we have used classification techniques. With this approach we have, in fact, continued to operate from a perspective that differs little from the one normally used with any other type of information given the variables. We have, in ways of varying complexity, synthesized the information contained in a sequence into some variable that we have treated in the same way as all the other variables.

In making these changes, however, it has only apparently been possible to capture the entire informational structure of a sequence, its shape and internal timing. We have assumed that these structures can be captured only at a later time, as a result of some aggregation of a series of sequences within a cluster. Yet there is nothing to guarantee that what we have obtained is real and not the result of some

technical mathematical trick (Bison 2009). Additionally, in this case our minds are very clever at finding regularities even where they do not exist. Furthermore, we ourselves are very able to find models or theories that explain some or other pattern *ex post*.

Obviously, there are reasonable grounds to believe that what we have identified through analysis exists and is therefore not a mathematical artifact. What I argue is different. In some ways, in this quest for models that extract information from sequences, we have canceled the sequences themselves. Our focus has shifted from career patterns to the distances between sequences. We have thus lost sight of our object of study; we have hidden behind a number: a distance. We have ceased to observe our research object, the careers, as a whole.

This proposal starts from the intent to find a new way to observe how careers develop over time and thus to capture their dynamic evolution. To this end, we need to give physical form to sequences and their underlying generative processes: that is to say, we must convert sequences into objects—networks graphs—with which it is possible to explore how they evolve.

Visualizing and Studying Sequences as Networks

The approach proposed in this paper is not a new one. There have been several attempts in the literature to combine sequence analysis and social network analysis. There follows a brief overview of the main applications simultaneously involving networks and sequences.

That networks have entered many fields of science should no longer come as a surprise. A network (or a graph) is a collection of nodes (or vertices), and the connections among them are called arcs, ties or edges. Networks are used to describe, model and analyze an enormous array of phenomena, including physical systems, communication networks, social systems such as networks of friendships or corporate and political hierarchies, physical relationships such as residue interactions in a folded protein, or software systems. Some of the various applications of networks relate precisely to the study and depiction of sequences.

Although they have been introduced recently, biology is the scientific field in which the most developed network techniques are applied to the study of sequences. Everything suggests that this approach has not only become a useful tool in the study of sequences but may in the near future lead to considerable progress in biology.

The most important and widespread use of networks to study sequences in biology goes by the name of *phylogenetic networks*. This technique uses a special network representation to model the distance matrix obtained from the deviations between all of the sequences. The specific aim of the technique is to model graphically the mechanisms of virus recombination (Huson and Bryant 2006). It is therefore an attempt to use a graph (network) to represent relationships between the sequences (taxa, allelic profile, etc.). Huson and Bryant (2006) propose defining

a phylogenetic network as *any* network in which taxa (i.e., phylum, order, family, genus, or species) are represented by nodes and their evolutionary relationships are represented by edges.

Not only biology has sought help from network analysis in interpreting relationships between sequences. Even before the development of this interest among biologists, social scientists had begun to test the ground. In particular, historical sociologists in the early 1990s began to use networks to solve their problems in studying sequences of historical events and/or narrative events.

This enabled them to make substantial contributions to the understanding of particular historical problems through the application of network models (Bearman 1993; Gould 1995, 1996; Padgett and Ansell 1993; Rosenthal et al. 1985; Barkey and Van Rossen 1997; Brudner and White 1997; Bearman et al. 1999; Franzosi 2004; Bearman and Stovel 2000; Bearman et al. 2003) to life history data (Butts and Pixley 2004) and to conduct quantitative narrative analysis (Franzosi and Bison 2010). Their applications focused on persons, institutions, lineages, and other elements linked by flows of resources, patronage, joint commitment, kinship, and violence.

Among these scholars, it is mainly Bearman (Bearman et al. 1999; Bearman and Stovel 2000; Bearman et al. 2003) who has developed a new line of inquiry known as a *narrative network* in which a strategy is developed to represent narrative life histories as networks of elements (a graph). In this new approach, the standard network techniques are applied to this representation of the narrative in order to identify some core elements of the process and thus reveal the properties of historical event sequences.

The idea is to translate narrative sequences made up of events and relationships among events into graphs consisting of nodes and links among nodes. In this regard, a sequence is a standardized and orderly narrative in which all of the elements/events/states that compose it are temporally ordered⁴.

From the Sequences to the Network

To obtain this new representation, the first step is to define what the nodes are and what the links between nodes are. We can conceive of a sequence as a recording of the succession of states, observed at regular time intervals, in the same survey unit. To illustrate this process, I shall use as an example the sequences in Table 12.1.

Each of the six sequences describes the trajectories followed by six people through two conditions, coded with the letters (a) and (b). The first sequence, for instance, begins with the state {a} and continues with the state {b}.

Each is different from the others in at least one element, and this difference can be attributed to a different moment in each of the six sequences. The transition oc-

⁴ Here ‘order’ means that each element in the sequence is in temporal relation to what precedes and follows it.

Table 12.1 List of six random sequences of length seven

Id	t_1	t_2	t_3	t_4	t_5	t_6	t_7
1	a	b	b	b	b	b	b
2	a	a	a	a	a	b	b
3	a	a	a	b	b	b	b
4	a	a	a	a	a	a	b
5	a	a	b	b	b	b	b
6	a	a	a	a	b	b	b

curs between two states. Each sequence can be thought of as a monthly recording of the employment positions of six subjects in the first seven months of their careers.

However, nothing prevents us from representing each of these sequences as a directed graph on which the nodes are the states observed and the ties between nodes are oriented according to the temporal relationships between the states. In this depiction, each node is in temporal relation only with the node that precedes it at time $t-1$ and with the node that follows it at time $t+1$. In the first sequence, for instance, because the node (a) at time t_1 is the first one, it only has a link with (b) at time t_2 . The node (b), at time t_2 , instead has two links, one incoming from node (a), which precedes it at time t_1 , and one outgoing to (b), which follows it at time t_3 .

This view, however, adds nothing to our knowledge about the career patterns followed by the subjects. We have only a different way to transcribe the information collected.

Yet on closer inspection, these sequences have several elements in common. For instance, at time t_1 , all of the subjects begin in the same state (a); at time t_2 , five subjects are in state (a) and one (id=1) moves to (b); and at time t_3 , four subjects are in state (a) and two are in state (b): one (id=1) moves/stays in state (b), and one (id=5) moves from (a) to (b). This suggests, for example, that between times t_1 and t_3 , at least four subjects had the same career: one (id=5) has a part of his/her career in common with the other four and a part in common with subject (id=1); and one (id=1) has only the beginning in common with all the other subjects and at time t_3 with subject (id=5). In other words, these subjects share a common pattern for the initial parts of their careers.

On reaching this conclusion, we must shift our focus from the individual sequences to what they have in common and what differentiates them. This enables us to configure a new and more complex structure in which the common and the distinct elements are combined to form a new sequence with characteristics different from those elements that generated it. We have moved from an individual sequence based on the relationships between states to a graph/sequence based on the relationship between states/events. From this new perspective, in fact, nodes are no longer states within a single sequence but become events experienced by one or more actors at time t .

In this process of abstraction/generalization, we start from individual sequences and gradually define a new sequence/graph which reflects the states/events (henceforth, events) and the relationships between events. We have thus moved

Table 12.2 Adjacency matrix of the six random sequences of table 12.1

	01a	02a	02b	03a	03b	04a	04b	05a	05b	06a	06b	07b
01a	0	5	1	0	0	0	0	0	0	0	0	0
02a	0	0	0	4	1	0	0	0	0	0	0	0
02b	0	0	0	0	1	0	0	0	0	0	0	0
03a	0	0	0	0	0	3	1	0	0	0	0	0
03b	0	0	0	0	0	0	2	0	0	0	0	0
04a	0	0	0	0	0	0	0	2	1	0	0	0
04b	0	0	0	0	0	0	0	0	3	0	0	0
05a	0	0	0	0	0	0	0	0	0	1	1	0
05b	0	0	0	0	0	0	0	0	0	0	4	0
06a	0	0	0	0	0	0	0	0	0	0	0	1
06b	0	0	0	0	0	0	0	0	0	0	0	5
07b	0	0	0	0	0	0	0	0	0	0	0	0

progressively from the individual path to the common pattern—from sequences to the narrative.

In technical terms, the solution is to transform the pattern traced by the subjects in their movements between states into the links between events. If we consider the six sequences of Table 12.1 as a matrix of rows and columns, we can see that:

- Each column of the matrix can be interpreted as a variable whose modalities are the events observed in the sample/population at that time as a longitudinal sample. At time t_1 , for example, only one distinct event (a) is detected in the six sequences. At time t_2 , the events increase to two because one subject is in state (b) and five subjects are in state (a).
- Each row of the matrix can be interpreted as the recording of the transitions and the links between points in time. One notes that between time t_1 and time t_2 , subject (id=1) changes from state (a) to (b) while the others remain in the same state (a). Between times t_2 and t_3 , subject (id=5) changes to state (b) while subject (id=1) stays in (b) and the other four stay in (a).

These two different perspectives can be displayed together to show the sequences through a square matrix, called an adjacency matrix or a sociomatrix, which has as many rows and columns (nodes) as the sum of the distinct states observed in each point in time. For instance, at times t_1 and t_7 , there is only one state (a) at t_1 and one (b) at t_7 , and only one node will be defined at time t_1 and at time t_7 . At time t_2 , however, both (a) and (b) exist. In this case, defined at time t_2 will be two nodes, one for state (a) (coded⁵ as 02a) and one for state (b) (coded as 02b). Each cell of the matrix greater than zero (Table 12.2) defines the link between two nodes, and

⁵ To ensure that the temporal order of states does not change when collapsing the individual sequences into the network, we generate a new coding that substitutes the original code of sequences and combines each individual state recorded in sequence with its temporal position. In the example, the new coding of the first sequence will be: {01a, 02b, 03b, 04b, 05b, 06b, 07b}. Here 01, 02, ..., n code the state time positions t1, t2, ..., tn and the letters (a) and (b) are the identifiers of states/events. This encoding allows, during visualization and analysis, identification of the states/events according to the temporal order in which they have occurred.

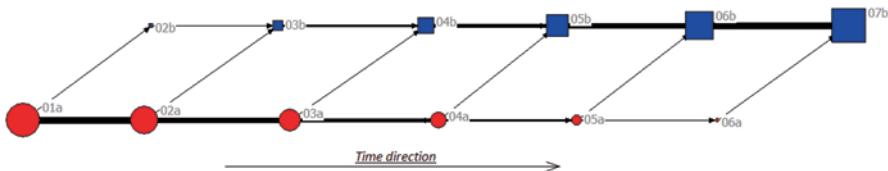


Fig. 12.2 Time sequence network of the six sequences of table 12.1

the cell values represent the force or the weight of the link between two nodes. In our example (Table 12.2), cell (01a, 02a) has a value of 5. This value indicates that between time t_1 and time t_2 , five transitions were observed between (a) and (a). In other words, five subjects in this range moved from event (a) at time t_1 to event (a) at time t_2 .

At this point, the switch to a directed graph is immediately possible (Fig. 12.2). The proximity of the states, in fact, determines which elements are related to each other, and the temporal order gives the orientation of the arc. For example, in the first sequence, the states 01a and 02b are consecutive, which suggests that subject (id=1) moved from state (a) to state (b) between time t_1 and time t_2 . In graphic terms (Fig. 12.2), this transition is depicted visually by an arc between (01a) and (02b) with the arrow pointing to (02b). In contrast, the other five subjects who stayed in (a) are represented by an oriented arc between (01a) and (02a).

There are several programs that enable the production of networks. The most popular is NetDraw (Borgatti 2002). This offers a wide range of ways to manipulate the graph: one of them is the assigning of weights to the links. NetDraw uses attributes—for example, the number of people who follow the pattern—by varying the thickness of the link pattern between the two nodes (Fig. 12.2). It is also possible to plot only the links above a certain weight threshold. Another important opportunity provided by NetDraw is that of attaching specific attributes to each node of the network. NetDraw displays these attributes by transforming them into the color, size and shape of the node. In the example of Fig. 12.2, a shape is given to each node depending on the event to which it belongs—a circle for (a) and a square for (b)—and the size of the node varies according to the number of subjects in the node at time t_i . These features considerably increase the network's readability.

Time is Space

What has been presented is not the only way in which a sequence may be displayed as a network. The proposed method was devised to preserve the causal structure and the temporal order of events. Each of these graphs is a trace in space of the trajectories followed by a group of subjects who moved within the time between events. In fact, people and events do not move through physical space; they do so through time. They are born, grow up and die over time. Movements between events occur in time, not in space. The transition from one social class to another is not a physical

Fig. 12.3 Transition sequence network of the six sequences in table 12.1



move from one place to another; rather, it is a change of role (in time). Thus, these graphs describe the motion of the events and subjects over time. The problem is what would happen if we decided to cancel time—in other words, if we decided to eliminate one of the two dimensions of the first network.

The time sequence network obtained from Table 12.1 is certainly highly distinctive (Fig. 12.2). It takes the form of two closely related patterns adjacent to each other. The direction of the transitions is entirely from event (a) to event (b). What differentiates the two patterns is first, their different timings. Although both patterns have the same length of time, they are temporally shifted. Pattern (a) occurs and ends an instant before pattern (b). The second difference is that over the course of time, all subjects, with constant frequency, move from (a) to (b). The result is that the nodes of pattern (b) progressively increase their weights over time, whereas the nodes of pattern (a) decrease their weights until they disappear. Indeed, at the end of the observed process, all subjects initially in event (a) have passed to event (b).

The question at this point is how the graph would change if it was decided to cancel the time dimension. In other words, what would happen if we focused on the transition between different states?

This would significantly reduce the system's complexity. In this new perspective, our attention is directed only to the succession of events as they arise from transitions between different states within the individual sequences. For example, suppose that the intention is to translate the sequence {aaaabba} into a time sequence network. The distinct nodes of the new graph would be seven in number (01a, 02a, 03a, 04a, 05b, 06b, 07a), as many as the points in time. If we cancel the time, however, and then consider only the transitions between different states, the new sequence will be composed of only three nodes (01a, 02b, 03a), one for each change state. Here, (01, 02, ..., 0n) ranks the changes to different states along the sequence and (a, b, ...) are the codes of the states/events.

With this shift, on the one hand we lose the timing of events but on the other hand we considerably reduce the complexity of the system itself. The point now is to ask what this new object is. What new information emerges from this graph? What factors differentiate it from the time sequence network? What are the limitations and what are the risks of its use?

Let us return briefly to the sequences in Table 12.1 and build our adjacency matrix considering only the changes between different events. In this new definition, which considers the transitions between events, a square matrix with 12 rows and columns in Table 12.2 changes to one with only two rows and columns, on which the cell (01a, 02b) has a value of 6?

In addition, the network is completely different. This new network (Fig. 12.3), which to distinguish it from the previous one, I shall call the ‘transition sequence network’, consists simply of two nodes and one arc. The graph is quite poor and

does not seem to provide much information. Yet it is in many ways much more informative than the previous one.

The two graphs are the two sides of the same coin—the same phenomenon. One is space or structure (Fig. 12.3) and the other is time (Fig. 12.2). The former represents the shape of space described at the point in time; the second represents the shape of time described at the point in space.

The time sequence network is like the sine wave that describes the point that ideally travels the circumference of a circle over time. In space it represents the temporal evolution of the transitions between events. It is the form that changes over time. The transition sequence network is the space described by the point that travels the circumference of the circle. It is the shape of space. It is the elementary underlying generative mechanism that produces the sequence over time. They are two sides of the same career. In some ways they are inseparable: one describes the shape of the space and the other describes the shape of time. Analyzed together, they describe how the space changes over time.

In the case of our last example, what is the information that we will obtain? With the time sequence network, we have come to the conclusion that the career pattern evolves from (a) to (b), so that it is constant in time. With the transition event sequence network, we have concluded that the complexity of the entire graph is produced by a single simple generative mechanism that stems from the transition from (a) to (b). Combining the information from both graphs, we can draw the conclusion that pattern (b) is a function of pattern (a).

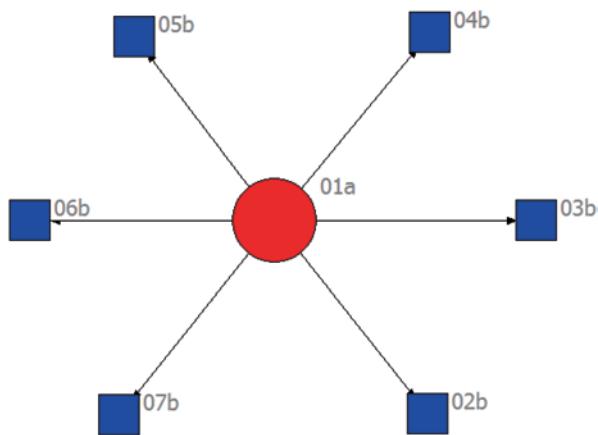
Time and Space

There is a third way to represent sequences as a network, and it combines both of the previous solutions proposed. In this new graph, the nodes are defined as transitions between states over time.

A new node will be observed if and only if an actor has a new event at time t . The tie between two nodes, in this new depiction, is defined as the probability of observing a transition between the state at time t_i and a new state at time t_n , that is, a subject who does not change state from t_i and t_n does not define a new node or new link.

If we apply this new definition to generate nodes and links between nodes to the sequence in Table 12.1, we obtain the network in Fig. 12.4. The network assumes the shape of a star with at its center the starting state (a), which in this case is common to all, whereas the links and nodes are defined by subjects who exit from state (a) to reach state (b) at time t_i . Each node (b) identifies the time t_i (02, 03, ..., 07) when the transition has occurred.

Fig. 12.4 Time event sequence network of the six sequences in table 12.1



The Class Careers of Men and Women

I asked in the introduction a series of questions on the real capacity of the sequence index plots, and the display systems in general, used to graphically capture the structure of the patterns underlying the sequences analyzed.

I wondered whether the utility of graphic tools could extend beyond simple graphic display; whether a graphic approach could become a valuable investigative tool with which to bring out new ideas on the structure, evolution and composition of the pattern; and whether adoption of these graphic tools could provide a new point of view useful for extending the hypothesis to be tested with other tools.

I have tried to answer these questions by proposing the application of network analysis techniques to sequences, and I have suggested three distinct types of visualization. Obviously, what follows is only experimental in nature and does not claim to prove anything. It is just another piece of information with which to test the potential of this approach.

The starting point is the same as in the introduction: I investigate whether the careers of men and women who end in the working class after 10 years of work are similar, as could be inferred from observation of the sequence index plot.

Following the order of presentation, I start with the time sequence network plot. To simplify the graphic representation, the observation window is transformed from monthly to quarterly. Therefore, each node represents the class position occupied by one or more subjects every 3 months in the first 10 years of their careers. The graph's temporal orientation (Fig. 12.5) is from the top-left corner to the bottom-right corner. The shapes of the nodes are based on the six EGP social classes, and the sizes are based on the number/proportion of subjects in each of the six classes in each quarter. The label next to each node reports the time of the observation (01, 04, etc.), and the EGP classes are identified as follows: (a) I-II; (b) IIIa; (c) IVab; (d) IVc; (e) VIIa; (f) VIIb.

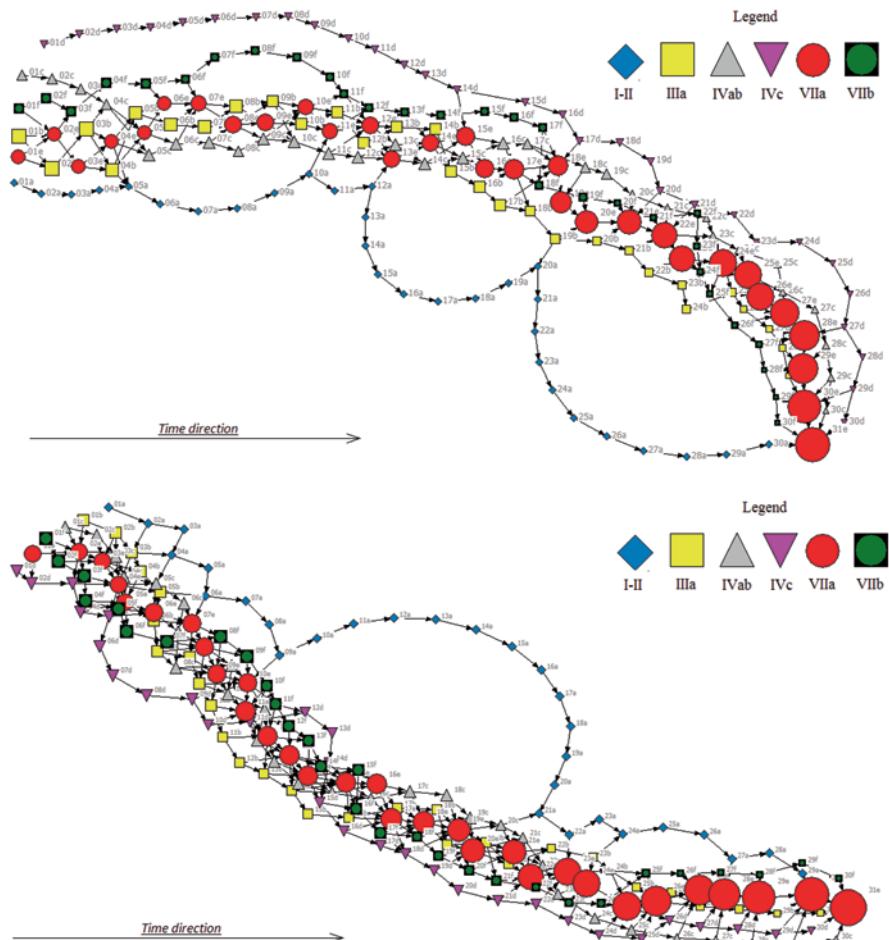


Fig. 12.5 Time sequence network of careers that end in class VIIa after 10 years

In contrast with the sequence index plot (Fig. 12.1), visual inspection of the two graphs (Fig. 12.5) reveals evident differences between the careers of men and women.

The first finding is the different number of ties in the two networks. Women have a smaller average number of links (283) than do men (357). This suggests that women have a career structure simpler than that of men. Women less frequently change from one class position to another and thus have careers that are simpler and less complex than those of men. Having a simpler career structure may also mean having a timing and time shape completely different from that of men, which is exactly what is observed when comparing the networks of women and men.

The two genders share a common central structure formed by the classes IIIa, IVab, and VIIa. These three patterns are closely interrelated and share a high number of ties, which indicates that most of the transitions between classes occur between these three classes.

What differentiates them is the timing with which the transitions from one class to another occur. The first difference is the different timings for men and women who start in class I-II. Among women, the first transition from I-II is observed after 3 years. Thereafter, the few remaining transitions into the other classes occur at intervals years apart and do not seem to display any sort of transition pattern. We could say that, among women, descending to class VIIa after starting in IIIa seems to depend on random factors not related to time.

For men, the structure of the pattern is different. The move from class I-II begins almost immediately and continues with some frequency and regularity in the first 3 years of work. There are no other transitions from the I-II to the VIIa classes for the next 4 years. The outflow, however, resumes with regularity in the last 3 years. Unlike women, apparent in this case is a career pattern with nonrandom characteristics. The pattern for men is a career in which the risk of descending from class I-II is concentrated at the beginning and end of the observation window. This would suggest that men have a high risk of descending early in their careers when the chances of failure are greater, especially for those men without the means to respond to contingencies that may occur in character the earliest years. There is then a period of relative quiet in which the risk of falling diminishes substantially. Finally, before the end of the tenth year of the career, the risk of falling towards the VIIa class starts to increase.

This is not the only difference between male and female careers revealed by the network. Among women, the first transition from IVab is observed after 3 years; among males it begins in the first quarter. The transitions to and from classes IVab, IIIa, and VIIa are also significantly more frequent among men than among women. Nevertheless, although the number of ties is smaller, the whole process leading to the working class seems to finish earlier among women. Especially when we consider classes IIIa and IVab, we note that most of the transitions from these classes are made until the eighth year of the career. Among men, however, there is a continuous flow to and from IVab, IIIa, VIIa until the last quarter, when by definition all have ended in the working class.

The last feature concerns transitions from the agricultural classes IVc and VIIb. This pattern is almost nonexistent among women. The first transition from IVab is observed after 5 years or so, whereas the other transitions follow the random pattern mentioned above. In contrast, among men there is a dense network of incomings and outgoings from the urban working class VIIa. This continuous flow is observed with the same intensity throughout the decade of observation.

Although these observations are not conclusive, they have enabled me to raise a series of issues that if found significant will be tested. However, it is clear that this first application of networks to the study of sequences has already yielded a fair amount of information that, with the instruments used to date, has remained hidden. Consider the possibility of measuring the degree of complexity of the system by its number of links. With this display, we can graphically analyze the evolution of

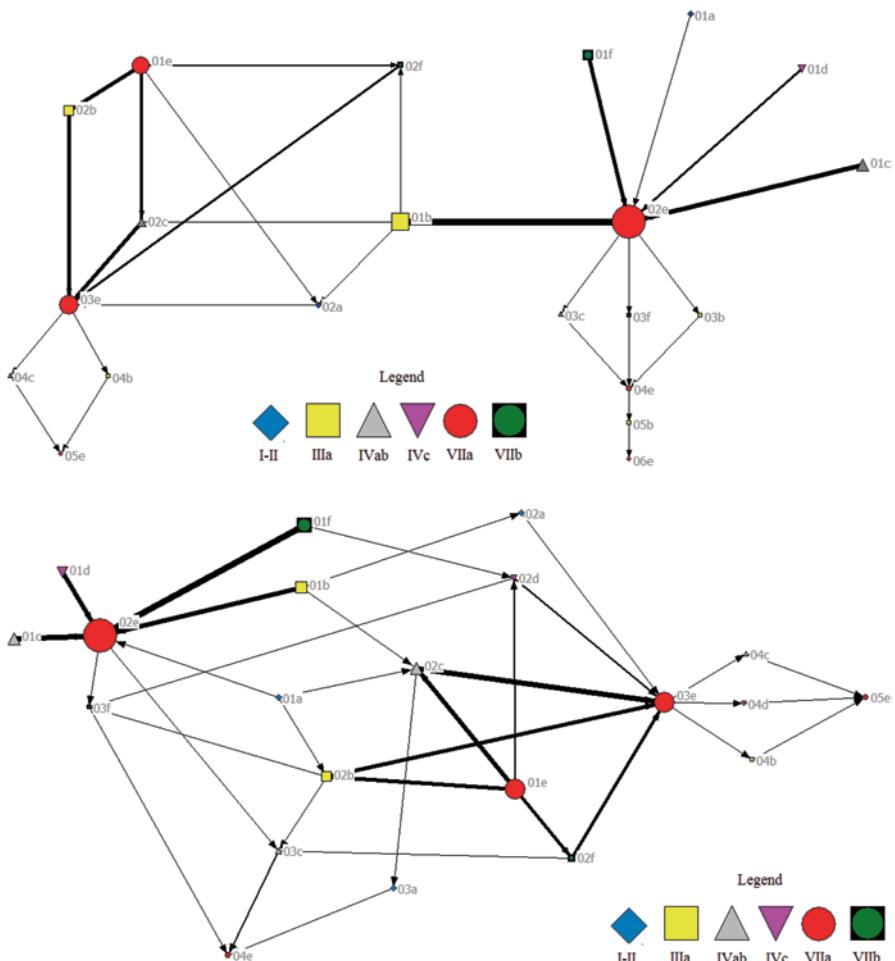


Fig. 12.6 Transition sequence network of careers that end in class VIIa after 10 years

events, capture their timing, and define the time shape in which the patterns unfold (for an example, see Bison 2012).

This, however, is not the only way that we can observe a sequence. A second method is to observe the transitions between events, thereby eliminating the timing of the transitions.⁶ Additionally, in this case the difference between women and men is evident (Fig. 12.6).

Two main patterns are apparent in both networks. The first pattern describes the direct transition from I-II, IIIa, IVab, IVc, and VIIb to the urban working class VIIa.

⁶ The node shapes are exactly the same as those in the previous graph. The changes in this graph are the following: (a) the observation window is monthly; (b) the size of a node is based on the proportion in the total population; and (c) the number of the label next to the node reports the order/succession of transitions (the number of event).

The second pattern describes those who start in VIIa, move mainly into IIIa, IVab or VIIb, and then terminate in VIIa.

What differentiates men and women is the differing complexity of their networks. Women have a simpler network with a lower number of links. Moreover, if we remove the ties followed by only women, we obtain two distinct patterns. These results suggest that among women there are two primary underlying generative mechanisms that operate separately from each other and combine to form the two main patterns of class careers of women who end up in VIIa after 10 years.

The first pattern consists of women who after entering some class then move to the urban working class VIIa. The second pattern consists of women who start in the working class, pass in the majority of cases to the white-collar middle class IIIa, the urban petty bourgeoisie IVab or into the agricultural working class VIIb, and end their journey back in the urban working class VIIa.

The male patterns are more complex. In this case, it is impossible to identify a single central node or some arcs that, if removed, as in the case of women, produce two separate career patterns. To be noted is a significant number of complex patterns with three or more different classes traversed in the first 10 years of the career by a large proportion of men.

Conclusions

There are many reasons that suggest that the advantages of bringing the analysis of sequences into network analysis outweigh the disadvantages. The main reason is that in this way we merge two approaches that in fact pursue the same purpose: that of overcoming the variable-centric vision in the search for common structures.

From the view of social network analysis, the social environment can be expressed as patterns or regularities in relationships among interacting units. (Wasserman and Faust 1994, p. 3)

The second reason is that in this way, we can combine the power of the synthesis of sequences with the power of the representation of the network in order to treat complex structures in a different way.

By representing complex event sequences as networks, we are able to observe and measure structural features of narratives that may otherwise be difficult to see. (Bearman et al. 2003, p. 64)

Adopting the network analysis perspective in the study of sequences is, however, to remain within the fundamental paradigm of sequences analysis, the ultimate purpose of which is to identify regularities and structures and to understand the phenomenon as a whole (Bearman and Stovel 2000).

To adopt the perspective of network analysis is also to adopt tools to represent complex structures such as sequences. It is to bring out the hidden structures with observation. It is to change the perspective. The purpose is to redefine the object by changing the viewpoint from which it is observed. It is to take a different perspec-

tive, one that is certainly new and, for this reason, perhaps able to furnish new and complementary information about the underlying structures and generative mechanisms.

Adopting a network perspective means shifting the focus from the actors to the events and to the relationships between the events generated by the actors. It is to study, in some way, the evolution of the phenomena observed in their individual careers in order to build a “...logical narrative with an inherent telos” (Abbott 1990, p. 141).

This paper has presented the first results from an application of network analysis to the study of sequences. It is an introductory work, and there are many issues that need to be studied and solved. Nevertheless, a number of insights emerge from these initial experiments that bode well for the future.

First, network analysis seems to be a valuable tool with which to visualize sequences. Graphs can bring out career patterns that have never previously been observed, such as the different working careers of men and women (already known) used as examples and briefly discussed here. A sequence takes place in time, and the time sequence network has also provided evidence of the extreme sensitivity of the instrument even in the visualization of the more marginal patterns.

The complexity of interpreting certain networks is certainly a limitation of the method suggested here. This may discourage researchers and induce them to abandon this approach to displaying sequences. I think this would be a mistake. Obviously, the use of these forms of display requires more attention from researchers. They must learn how to separate the patterns that actually exist, even if produced by a small number of subjects, from the ties produced by the noise floor from errors in data collection.

This does not mean that the power of these views is so great as to allow entry into each individual pattern. We can follow the path and its interweaving with other patterns. Through the relationships and transitions from one state to another, we can see the timing and time shape of the career patterns in graphic terms.

Obviously, not all graphs are easy to use. Some are so complex and intricate that they resemble more a plate of spaghetti than a network of sequences. I believe however, that this limitation in the ability to read these networks depends only on our ability to find attributes (indices, measures and selection thresholds) with which to bring out the underlying coherent structures even in these cases.

This is not the only limitation of the method proposed. Among the others, the one that creates the greatest methodological and philosophical problems is the annulment of individual sequences. This may not be a problem if all subjects follow the same pattern, but unfortunately they do not. It may not be a problem if one decides that the past does not affect the future, but this is equivalent to undermining the pillars on which sequence analysis is based. Everything is (con)fused to form a different structure in which the individual trajectories disappear to make space for a ‘mean’ trajectory that describes the transitions between two temporally contiguous points.

What is important in these phases is to not lose sight of individual careers and how they contribute to the realization of the network. A simple method that can be

used for this purpose is, again, to use the attributes to define additional information on the nodes (Bison 2012). This would (visually) check the disconnect between the performance obtained with the network and the actual one observed in individual sequences.

There are many things that remain to be done. Consider all the potentialities that have been only briefly mentioned here, including the biomedical approaches using *phylogenetic networks* and the *splits tree* method (for an example, see Bison 2012).

This article has illustrated a new, network-based strategy with which to represent and analyze sequences, but this is only one aspect of what can be done by combining network analysis and sequence analysis. The real novelty, which is not considered in this paper for reasons of space, is the transition from static to dynamic.

The real breakthrough will come when we are able to move from a static to a dynamic representation of our phenomena, when our patterns begin to take life and shape before our eyes.

The progress achieved in recent years by network analysis is impressive in both methodological and technical terms. There is now an established research area engaged in modeling network dynamics. The frontier in this field is no longer that of representing a structure of relations as a whole. The new frontier is representing a structure of relations as a whole that changes over time and space, which is exactly what we ourselves are trying to do, from another point of view, with the sequence analysis.

Now required are new tools with which to study and graphically model the evolution in time and space of our patterns. I believe, perhaps wrongly, that this opportunity, for now, will be provided by adjusting the tools of network analysis to the study of sequences. I think, in fact, that the shift to a network perspective could provide research tools that will be more useful for the study of sequences.

Finally, I think that when we are able to adopt these tools, the only limitation will be our imagination, our ability to imagine.

Reference

- Abbott, A. (1990). Conceptions of time and events in social science methods. *Social Science History*, 23, 140–150.
- Barkey, K., & Van Rossen, R. (1997). Networks of contention: Villages and regional structure in the seventeenth-century Ottoman Empire. *American Journal of Sociology*, 102(5), 1345–1382.
- Bearman, P. (1993). *Relations into rhetorics. ASA rose monograph series*. New Brunswick: Rutgers University Press.
- Bearman, P., Faris, R., & Moody, J. (1999). Blocking the future: New solutions for old problems. Special issue: What is social science history? *Social Science History*, 23(4), 501–533.
- Bearman, P., & Stovel, K. (2000). Becoming a Nazi: A model for narrative networks. *Poetics*, 27, 69–90.
- Bearman, P., Moody, J., & Faris, R. (2003). Networks and History. *C O M P L E X I T Y*, 8(1), 61–71.
- Bison, I. (1999). *Life-packaging in Italy, paper presented at the POLIS project conference*. Berlin: Max Planck Institute. 17–18 March 1999.

- Bison, I. (2009). OM matters: The interaction effects between indel and substitution costs. *Meth-odological Innovation On Line*, 4(2), 53–67.
- Bison, I. (2011). Education, social origins and career (Im) mobility in contemporary Italy: A holistic and categorical approach. *European Societies*, 13(3), 481–503.
- Bison, I., (2012). Sequence analysis and network analysis: An attempt to represent and study sequences by using NetDraw. *Lausanne Conference On Sequence Analysis (LaCOSA)*, University of Lausanne, June 6th–8th.
- Bison, I., & Esping-Andersen, G. (2000). *Life-packaging: Dynamics of working career and family formation in Italy. Workshop participation of the POLIS project on Globalization*. Madrid 10–11 March 2000.
- Borgatti, S. P. (2002). *NetDraw software for network visualization*. Lexington: Analytic Technologies.
- Brudner, L., & White, D. (1997). Class, property, and structural endogamy: Visualizing networked histories. *Theory and Society*, 26(2/3), 161–208.
- Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with Stata. *Stata Journal*, 6(4), 435–460.
- Butts, C. T., & Pixley, J. E. (2004). A structural approach to the representation of life history data. *Journal of Mathematical Sociology*, 28, 81–124.
- Franzosi, R., (2004). *From words to numbers: Narrative, data, and social science*. Cambridge: Cambridge University Press.
- Franzosi, R., & Bison, I. (2010). Temporal order: Sequence analysis. In R. Franzosi (Ed.), *Quan-titative narrative analysis, QASS (162)* (pp. 118–123). Thousand Oaks: SAGE Publication, Inc.
- Gabadinho, A., Ritschard, G., Muller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Gould, R. V. (1995). *Insurgent Identities: Class, community, and protest in Paris from 1848 to the commune*. Chicago: University of Chicago Press.
- Gould, R. V. (1996). Patron-client ties, state centralization, and the Whiskey Rebellion. *American Journal of Sociology*, 102(2), 400–429.
- Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2), 254–267.
- Padgett, J., & Ansell, C. (1993). Robust action and the rise of the Medici, 1400–1434. *American Journal of Sociology*, 98(6), 1259–1319.
- Rosenthal, N., Fingruttd, M., Ethier, M., Karant, R., & McDonald, D. (1985). Social movements and network analysis: A case study of nineteenth-century women's reform in New York State. *American Journal of Sociology*, 90(5), 1022–1054.
- Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *Euro-pean Sociological Review*, 17(2), 119–144.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.
- Widmer, E., & Ritschard, G. (2009). The de-standardization of the life course: Are men and wom-en equal? *Advances in Life Course Research*, 14(1-2), 28–39.

Chapter 13

Synchronising Sequences. An Analytic Approach to Explore Relationships Between Events and Temporal Patterns

Denis Colombi and Simon Paye

Introduction

Advocates of sequence analysis often emphasise the benefits of its holistic approach, as opposed to event history modelling (Abbott and Hrycak 1990, p. 147; Robette 2010, p. 3). This clearly reflects the fact that sequence analysts have mainly focussed on structural patterns, while devoting less interest to events. Most existing studies based on sequence data have indeed concentrated on measuring resemblance in order to identify typical occupational careers (e.g. Blair-Loy 1999) or life course trajectories (e.g. Bras et al. 2010), whereas most endeavours of event history analysts were geared to uncover and measure causal relationships (Blossfeld and Rohwer 2002).

Despite some pleas for articulating these two statistical methods (Lemercier 2005, p. 17), their development has remained mostly antagonistic. This led to structural patterns and events being studied separately most of the time.

Yet, paradoxically, the interaction between events and larger temporal patterns points to old theoretical problems. Let us consider two of them. The concept of *turning point*, introduced by E. Hughes (1971), is a case in point. The conceptualisation of turning points as transitions between different “probability regimes” (Abbott 1997, p. 92) is still lacking appropriate analytical tools and methods to be operationalised¹. So far, ethnographic methods have proven to be more suited than

¹ Perhaps even more paradoxically, the concept of turning point has largely been overlooked by sequence analysts.

We would like to thank the participants and organisers of the LaCOSA conference, from whom we received stimulating feedback. We are also thankful to Claire Lemercier, from whom we received insightful comments on earlier versions of this chapter.

D. Colombi (✉) · S. Paye
Centre de sociologie des organisations, Science-Po-CNRS, Paris, France
e-mail: colombi.denis@gmail.com

S. Paye
e-mail: simon.paye@sciences-po.fr

quantitative models to identify specific relations between singular events and social processes displaying turning points². They are far less useful when it comes to assessing the depth and the strength of these relations. Quantitative methods could bridge this gap, but they are still poorly equipped to explore the role of specific events in complex biographical or organisational processes. A second core theoretical problem relates to the role of events in processes of social *differentiation* and *homogenisation*. While many authors have shed light on converging and diverging trends in biographical (Dubar 2010) and organisational (Di Maggio and Powell 1983) processes, we know little about the potential triggering role of specific events. One might, for example, wonder whether parenthood is followed by convergence towards conventional employment and residential mobility patterns, or instead by a pluralisation of occupational and residential situations. These two theoretical problems point out the need of a research method that allows analysing singular events and temporal sequences together.

This paper provides a simple analytic solution that enables studying interactions between events and larger temporal patterns. The key operation of this method consists in synchronising sequential data according to idiosyncratic events. After synchronisation, each sequence (e.g. series of job positions) is positioned according to an event that takes place in a particular moment for each individual (e.g. childbirth).

Sequence synchronisation has been used, as far as we know, in very few existing studies (Blanchard 2010; Giudici and Gauthier 2009). By alternatively referring to these studies and our research experience, we will argue that sequence synchronisation (1) opens up potential opportunities to operationalise the concepts of turning point and differentiation with sequence data, (2) invites to reduce the methodological gap between sequence-based and event history approaches, (3) could potentially be replicated in other fields of social science. Hence the need of understanding the theoretical meaning and implications of the operation, as well as clarifying its added value in comparison with other existing approaches.

Two empirical cases taken from our respective ongoing research will illustrate the heuristics of this practice of sequence analysis. The first one shows how the division of academic labour in British universities is tied to career patterns. Sequences of functional mobility of 122 individuals are synchronised at the date of access to the permanent workforce. Sequence visualisation shows that before this date, academic careers tend to converge towards polyvalent jobs, and then diverge anew towards monovalent jobs. Division of labour in academia thus appears to result from a process of functional differentiation. The second case explores the impact of international mobility on social mobility. Class sequences of French individuals from the “*Histoire de vie 2003*” survey (Insee 2003) are synchronised according to the period of expatriation. Sequence visualisation reveals that class mobility reaches a peak during the expatriation period, as compared to those who do not experience international mobility.

² See H. Becker (1963) for a famous example of such approach.

The first section of this paper details the operation of synchronisation and considers a number of practical as well as methodological issues. The next section presents the two illustrative cases and their contribution to their respective research fields. In section three, the added value of sequence synchronisation is discussed with reference to three existing methods that can be used for the same purpose: event history analysis, multichannel sequence analysis and multiple sequence alignment. The chapter closes with a short conclusion that summarises the main methodological arguments and opens up a number of questions that would deserve further attention.

Synchronising Sequence Data: the Operation

Sequence data is, most of the time, either displayed in calendar time (Blair-Loy 1999, p. 1358) or in age-relative time (Robette and Thibault 2009). The first option has the advantage of preserving the historical context, while the second one is useful for analysing life course processes, especially those which display, because of highly institutionalised temporal norms, little age variance (i.e. leaving compulsory education).

A process other than age-based reference can be used by taking as reference a specific event that occurs in a given point in one's career, i.e. an *idiosyncratic event*. Marriage or childbirth are good examples of such events. This option has been followed by P. Blanchard in an analysis of 502 AIDS activists' careers (2010, p. 95), and by Giudici and Gauthier's work on interactions between professional trajectories and parenthood (2009). Only the first author has provided accounts on what synchronisation actually brings forth for the visualisation of sequences. According to him, index plots of synchronised sequences are particularly suited to illustrate two phenomena: (1) individual contrasts in terms of order and duration of commitment experiences into AIDS activism, and (2) the brevity of AIDS activism compared to more lengthy related experiences in other spheres of life, such as sociosexual trajectories (Blanchard 2010, pp. 19–20).

The strengths and weaknesses of sequence synchronisation have not been discussed, nor were they justified on a theoretical basis. Yet, a large number of social processes unfold more according to event-related time patterns than according to age-based time or historical time. Therefore, their study needs specific time references. Childbirth, incarceration or emigration seem to be events influential enough to consider using time axes defined according to the dates on which they happen, rather than calendar or age-based time axes.

Figure 13.1 provides one example of the operation of sequence synchronisation with a sample of five individuals taken from the “*Histoire de Vie 2003*” survey. Sequences of employment status are plotted in three different ways: left-aligned, right-aligned and synchronised according to the date of the first marriage.

Sequence synchronisation allows seizing interactions between the event (marriage) and the whole employment trajectory. Individual one, for example, leaves the labour market as soon as he/she gets married, while individuals four and five

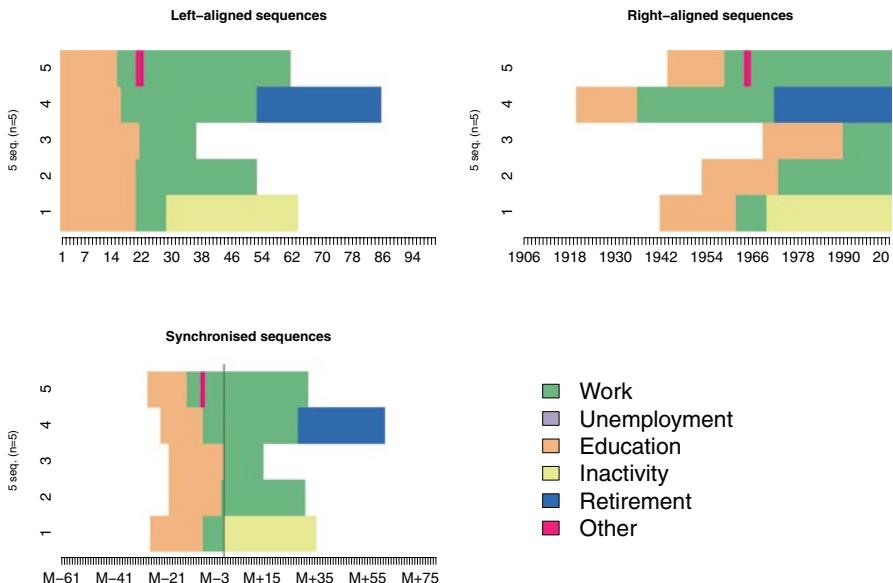


Fig. 13.1 Three types of visualisation of sequences of employment status: *left-aligned* (age-based time), *right-aligned* (calendar time) and synchronised according to the year of first marriage

remain employed. Sequence synchronisation allows observing that none of the five individuals gets married while studying: this could mean, for example, that marriage was “postponed” until education is completed, or alternatively that marital life entailed college drop-out.

In this specific example, the information related to the event (marriage) is provided by an external variable. This may be called *exogenous synchronisation*, as opposed to *endogenous synchronisation*, in which the event used to synchronise sequences is embedded in sequence data. We now illustrate cases of endogenous and exogenous synchronisation using data from 6259 respondents of the “*Histoires de vie 2003*” survey.

Endogenous Synchronisation

Sequence synchronisation can be performed using an event that is endogenous to the sequence alphabet. Figure 13.2 displays 1,454³ employment careers synchronised according to the 1st year of unemployment—information that is embedded in the sequences. Time axes have been left-truncated 30 years before first unemployment episode and right-truncated 15 years after, so as to reduce interpretive pitfalls due to high levels of missing values.

³ Careers displaying no unemployment episode have been removed.

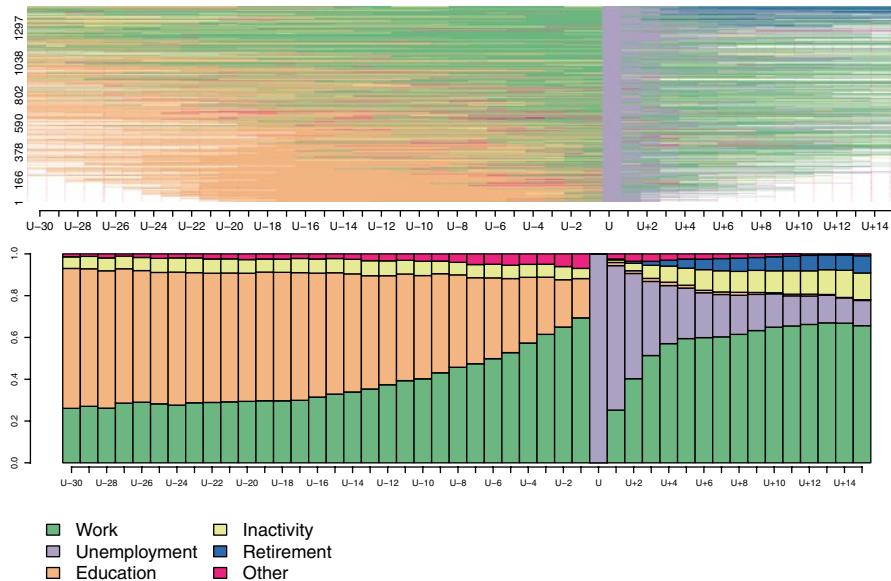


Fig. 13.2 An example of endogenous synchronisation: employment career before and after the first unemployment episode (sequences in the index plot sorted by age, with younger individuals at the bottom)

An interesting outcome relates to trends preceding first unemployment: Most individuals actually occupy a job, and less than 20 % transit directly from education or higher education to unemployment, although this case seems to happen more frequently within the youngest generation. A comparison between men and women's careers (not shown here) reveals that unemployment is more likely to be articulated with episodes of inactivity within the female population.

Exogenous Synchronisation

In the case of exogenous synchronisation, the event is not coded into the sequence alphabet but provided by an external time variable. Figure 13.3 provides index and frequency plots of employment careers synchronised according to the year of first marriage. Time axes have been left-truncated 30 years before first marriage and right-truncated 30 years after so as to reduce interpretive pitfalls due to high levels of missing values.

Marriage clearly corresponds to an important transitory period in occupational careers. Most inactive episodes appear after first marriage. While the level of employment rises continuously before the event, it then stabilises at about 65 % as first marriage occurs.

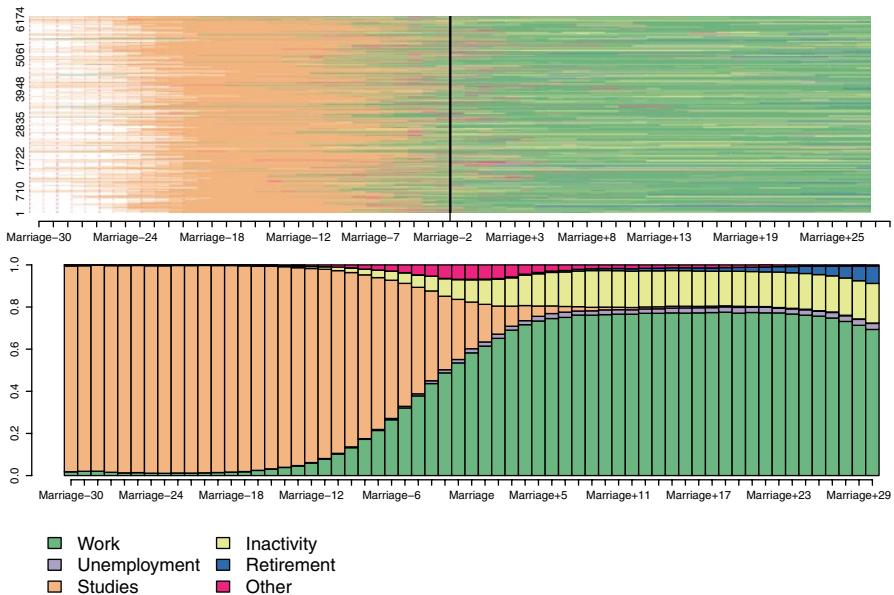


Fig. 13.3 An example of exogenous synchronisation: employment careers before and after first marriage

A comparison between different generations (not shown here) yields further results. Individuals from newer generations are less likely to work before first marriage—perhaps a consequence of longer periods of education. Rather counter-intuitively, marriage seems to become a sharper turning point in people's lives.

Both endogenous and exogenous synchronisation allow exploring interactions between specific events and larger temporal patterns. However, exogenous synchronisation, as it incorporates additional contextual information, offers more analytical opportunities. It particularly allows studying interactions between events and temporal patterns of distinct spheres of individuals' lives, such as work, residence, family life, etc.

Two Illustrative Cases

We now present and discuss two illustrative cases taken from our respective ongoing research. The first one is a diachronic analysis of the division of teaching and research labour within a population of British academics. It is part of a PhD research on the dynamics of career differentiation within the academic profession. The second illustrative case, derived from a PhD research on international mobility and professional careers, deals with the consequences of employment abroad for French high-skilled workers. While the first case illustrates the usefulness of sequence synchronisation for the study of differentiation processes, the second one does so for the study of turning points.

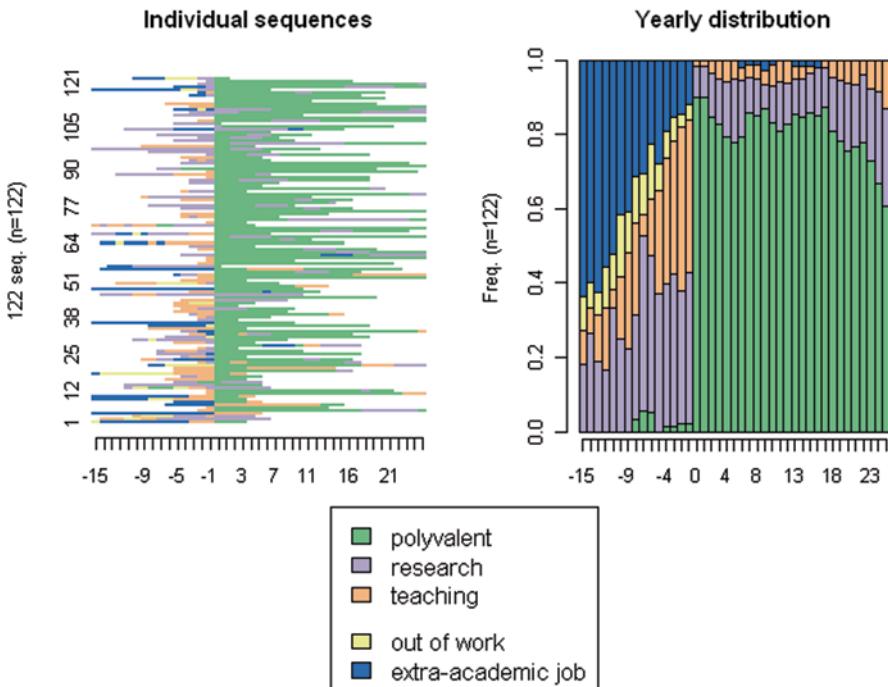


Fig. 13.4 Functional differentiation in academic careers before and after permanentship

Case 1: Internal Job Ladders as Matrix for Functional Differentiation: an Explanation of the Division of Academic Labour

The division of academic labour in British universities is a poorly understood phenomenon. It is mostly accounted for as a consequence of the proliferation of casual jobs (most of them being “teaching-only” and “research-only” positions). However, an important part of monovalent jobs is to be found within the permanent workforce. Division of academic labour therefore needs different explanations according to whether it occurs in an external labour market of temporary jobs or in an internal market in which job allocation follows rules of career advancement (Doeringer and Piore 1971).

To understand how division of academic labour occurs within the permanent workforce, it is proposed to conceive of it as a temporal process of functional differentiation. Synchronising academic careers allows determining whether permanentship (i.e. the event of getting the first permanent position in one’s career) influences research and teaching specialisation. In Fig. 13.4, sequences of job positions have been centred on the year in which individuals got access to their first permanent position.

As index and frequency plots of yearly employment functions show, professional profiles tend to diverge after access to the first permanent position, most of the time towards teaching-only or research-only jobs. This divergence resulting from complex sequences of career transitions has the effect of increasing the overall level of division of labour.

As Fig. 13.5 shows, this research outcome would simply not have been identified without having recourse to synchronisation.

Observability of the phenomenon is highly dependent on the definition of the time axis. Without synchronising sequences, visualisation cannot account for any clear tendency of functional differentiation. Other existing methods such as event history analysis or multichannel sequence analysis would have required much more effort to yield the same result. Synchronisation thus appeared to be particularly well suited to identify career differentiation as a key process of division of academic labour.

Case 2: Expatriation as a Turning Point in Upward Class Mobility

This second application deals with the complex links between international mobility and social mobility. A core question in the related literature is whether international mobility acts as a specific form of capital, just as economic capital or cultural capital (Kaufmann 2001; Lévy 2003; Wagner 2007). Is there such a thing as an “international capital” that generates competitive advantage in people’s social mobility? How does employment abroad interact with individual trajectories across social classes? The case presented here discusses this idea of “international capital” with an analysis of the social mobility of French “expatriates”, i.e. workers who left France to occupy a job abroad and then came back to France. It is geared to illustrate the usefulness of sequence synchronisation in understanding the role of expatriation in class careers.

For this section, we use sequence data stemming from the “*Histoire de Vie 2003*” survey (Insee 2003), which contains yearly information on social, occupational and geographical mobility of 346 French expatriates. Most of them are white-collar workers (30.4%, 115 individuals), classified as “*cadres et professions intellectuelles supérieures*”⁴. Up to 10.5% of this category has experienced international mobility (as compared to 4.2% for the whole population).

The core empirical question is about causation: Did they reach the category “*cadres et professions intellectuelles supérieures*” because they accumulated international capital, or is it because they belonged to that upper class that they were more likely (or disposed) to experience international mobility? Order is crucial in dealing with this question: If international mobility has an impact on social mobil-

⁴ The French socio-professional classification “*Professions et catégories socio-professionnelles*” not only describes occupational situations, but also takes into account social classes. The category “*Cadres et professions intellectuelles supérieures*” roughly refers to managers and knowledge-based professionals, but can be used as a proxy for identifying upper-class individuals.

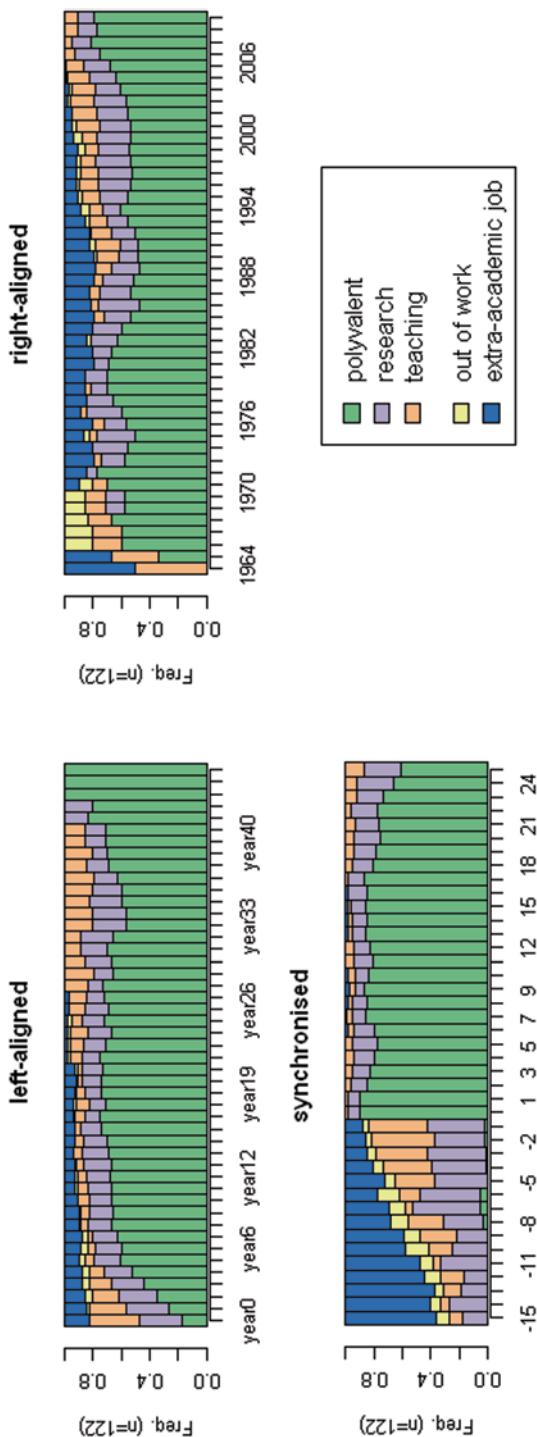


Fig. 13.5 Comparison of frequency plots with different time axes

ity, people should be more likely to join the upper class after expatriation than before. The comparison between expatriates and the reference population is here used to determine whether social mobility is more likely to happen after international mobility.

Figure 13.6 displays class careers using Insee’s “socio-professional groups” scheme. Sequence synchronisation has been slightly customised since it takes into account two “dates” instead of one. Graphs on the top display expatriates’ careers centred on the year of expatriation (left graph) and on the year of return to France (right graph)⁵. The years abroad between those two dates have been deleted. On the bottom, equivalent frequency plots for the whole population are centred on the mean age of first expatriation (left) and on the mean age of first return (right). This synchronisation according to mean ages allows comparing “expatriates” with the reference population⁶.

The proportion of people belonging to the upper-class *before* expatriation is significantly lower than the proportion *after* expatriation. The equivalent difference is less noticeable when looking at the reference population. Expatriates therefore experience more frequently upward class mobility during their migratory episode than does the reference population of the same age range. Expatriation can be seen as a turning point as it operates a clear shift between “probability regimes” of class mobility (Abbott 1997, p. 92). This effect is even more important within the female population (not shown here). Women expatriates are less likely to be part of the upper class before expatriation and more after than is the wider population of upper-class women. There is therefore evidence of a substantial gender sensibility to upward class mobility: Expatriation is a more salient turning point for women. Further analysis is needed to examine whether the anticipation of upward social mobility could explain individuals’ disposition to international mobility.

This short example illustrates the interest of sequence synchronisation for social mobility studies. It allows dealing with a great variety of social trajectories while at the same time taking account of order and events.

Sequence Synchronisation Compared

In this last section, we turn to discuss some of the strengths and weaknesses of sequence synchronisation with regard to three other ways of analysing longitudinal data: event history analysis, multichannel sequence analysis, and multiple sequence alignment.

⁵ Multiple expatriations seldom happened within the population of the dataset: only 49 individuals out of 376 were concerned (13.03 %).

⁶ Comparability is justified if the variance of mean ages (migration and return) is low. Here, 75 % of all expatriations occur before age 28, and 75 % of the returns before age 35. Standard error is 6.6 for departure and 8.5 for return.

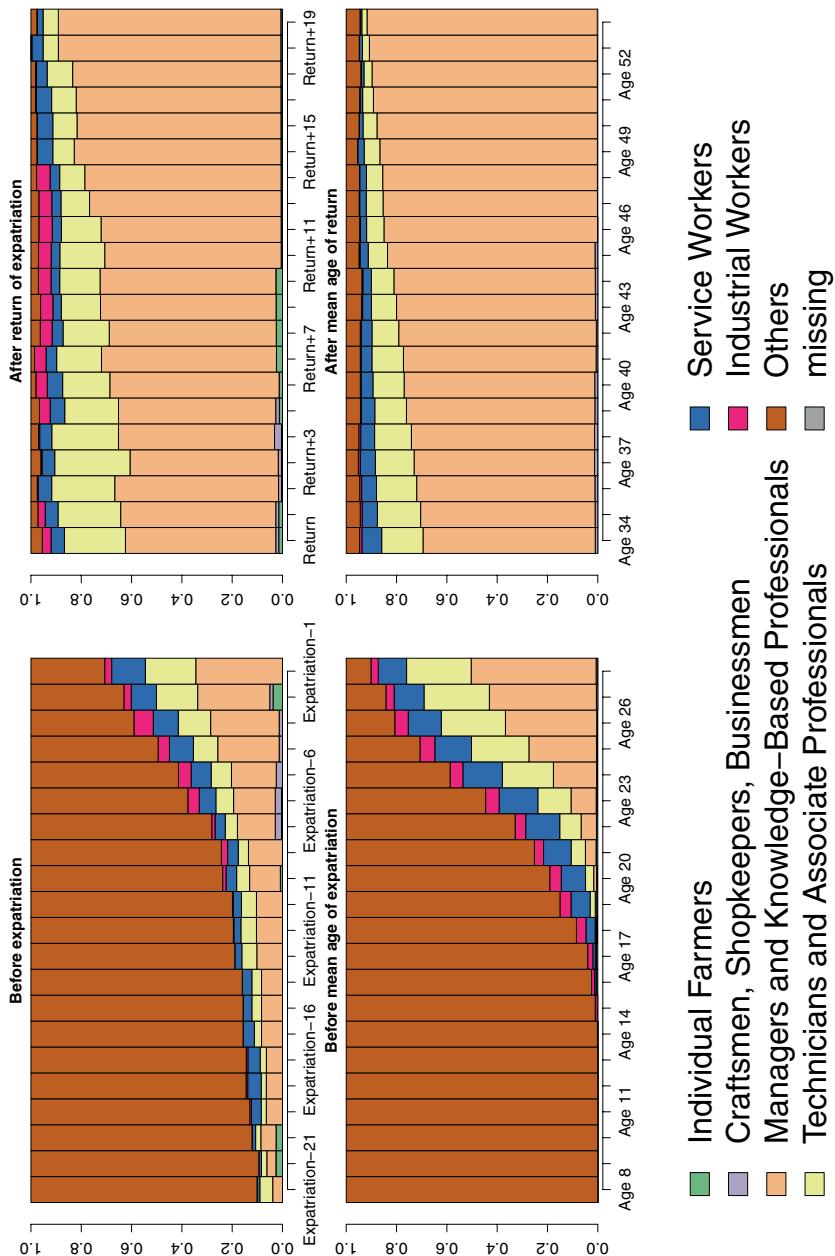


Fig. 13.6 Social mobility compared: upper-class ‘expatriates’ vs. the whole upper-class population

Let us start with the main strength: Synchronising sequences allows comparing between sequences processes that happen before, during and after a specific event. This is useful for treating research questions requiring the identification of transversal patterns around particular events. It allows visualising and computing phenomena of turning points or of divergence or convergence associated with processes of differentiation/homogenisation.

Its major weakness relates to missing data. Synchronising sequences according to idiosyncratic events entails expanding the time axis. The synchronised database therefore contains more missing data on the extreme left and right of the time axis than it did before the operation. Another relative weakness is the fact that an important decision is made prior to analysis. To some extent, synchronizing a life around a given event gives it a centrality that may appear fallacious. It is, nevertheless, difficult to assess the extent to which the choice of an event may artefactually increase its significance.

Strengths and limitations of sequence synchronisation can be put in relation with event history analysis, multichannel sequence analysis, and multiple sequence alignment. We now discuss in detail how each of these analytical approaches deals with the identification of turning points and differentiation/homogenisation processes.

Event History Analysis

Event history analysis is the most obvious solution to study relationships between events. According to Blossfeld and Rohwer's definition, event history modelling is geared to "discover the causal relationships among events and to assess their importance" (2002, p. 3).

One of the main strengths of event history techniques is their suitability for statistical tests of causality assessment. They allow analysing the probability of one specific event (e.g. promotion) occurring after a previous event (e.g. recruitment). Sequence analysis provides more comprehensive accounts since it displays all the transitions at once. This is hardly possible using event history modelling, unless with "competing risks" models. These models, also called "multiple destination models", are designed to deal with one origin state and multiple destination states (Blossfeld and Rohwer 2002, pp. 101–107). However, they do not fully account for all the possible transitions within the state space. Models rarely include more than three competing risks, thereby accounting for three transitions at best. Some sequence databases have an alphabet including more than 20 categories, accounting for no less than 400 theoretically possible transitions. Multiple risk modelling does not help in dealing with such categorical complexity. As C. Lemercier puts it (2005, p. 17, our translation):

If event history analysis can include, in theory, any past event of the individual or any historical event of the past in general, as an element of causality of the movement under scrutiny (thereby taking into account, in a way, the entire trajectory), its primary unit of analysis is the event.

Sequence synchronisation precisely allows studying relationships between an event and entire temporal patterns.

Another advantage of sequence synchronisation is its lesser recourse to hypotheses. Event history models need strong hypotheses about which event counts and which events do not count. This can appear to be very problematic, especially during exploratory stages of research. Sequence synchronisation limits considerably the costs incurred by such *a priori* and assertive selections.

Multichannel Sequence Analysis

By defining different “careers” per individual (e.g. occupational, sociosexual, residential, etc.) and comparing their patterns, Multichannel sequence analyses are a possible way to explore relations between distinct spheres of life (Blanchard 2010; Gauthier et al. 2010; Pollock 2007). They can therefore show how a given trajectory in one sphere (e.g. occupation) can be related to another (e.g. family).

A recent study (Gauthier et al. 2010) has produced insightful frequency plots showing graphical evidence of interactions between family and occupational temporal patterns. These frequency plots do reveal substantial shifts occur between age 18 and 28 in both family and occupational spheres. They do not, however, allow understanding the relation between a given transition in one channel (e.g. from education to full-time employment) with temporal patterns in family trajectories. The examination of the frequency plots can, at best, allow identifying “resonances” between the temporal patterns of the distinct channels. By remaining bound to one class of object (temporal patterns), multichannel sequence analysis does not allow exploring the actual role of specific events in complex time-evolving processes.

Multiple Sequence Alignment

A comprehensive overview of the range of opportunities offered by this method of data representation can be found in J. A. Gauthier’s PhD thesis (2007, pp. 144–167). He defines it as such (2007, p. 144):

[Multiple sequence alignment] aims at aligning not only a pair, but a whole set of sequences at the same time. The multiple sequence alignment procedure attempts the optimal alignment of a set of sequences so that homologous or identical symbols appear together in the same columns.

Its main use is geared to identify (2007, p. 145):

Regions of sequences that are more or less identical. In a sociological perspective, such ‘conserved regions’ may be seen as a configuration where, at some given points of the life course, the same social status (e.g. education, full-time or part-time paid work, housework, retirement) is held by a large number of the individuals.

This method requires using the optimal matching algorithm as a first step to produce a metric that is subsequently used to align the sequences at various points in time (2007, p. 146):

By aligning the sequences, the algorithm optimizes the vertical alignment of identical or homologous symbols, while minimizing the structural distortion of the original sequences.

Amongst the opportunities for sequence visualisation offered by multiple sequence alignment, the one that is of interest to us is what he calls “multi-aligned individual trajectories” (2007, pp. 153–166). According to him (2007, p. 153):

We expect such graphs to highlight some less visible feature of typical individual trajectories, once they have been processed in order to optimize conserved regions among the members of that type.

One major problem is that the obtained index plot has some distorting effect on the time scale, as it inserts gaps (missing values) to some of the sequences. This is the trade-off inherent to multi-align the sequences. If these gaps expand the time scale too much, one option is to “delete columns according to the percentage of gaps they contain” (2007, p. 154). Yet again, this means losing data. Another limit of this method is that it is less useful when applied to sequences showing important variability (2007, p. 164). The dataset has to be particularly stable, and this is far from the case in many social processes.

Gauthier’s thesis displays a wide range of index plots of occupational careers to which multiple sequence alignment has been applied. The examination of these index plots well reflects the fact that many sequences go through the typical sub-sequence education of full-time job—part-time job—retirement. In fact, multiple sequence alignment visualisations emphasise what is common to an important number of sequences (this number increasing as the percentage of missing values raises), but it seems to be less useful in accounting for heterogeneity of the data.

Even if multiple sequence alignment is presented as a means to identify states or subsequences that are shared by many individuals, it also allows uncovering diverging or converging trends. Yet, it tends to distort the time scale and to lose information related to heterogeneity. As regards identifying turning points, it seems less useful as it is limited to the states included in the sequence alphabet.

Conclusion

Sequence synchronisation appears to be a useful analytic approach to explore the articulations between events and temporal patterns. Its main strength is that it takes into account both events and sequences. This allows studying both how events are conditioned by previous sequences of states and how, in turn, events reconfigure the range of possibilities within subsequent temporal patterns. This specificity in regards to existing methods, such as event history analysis, multichannel sequence analysis and multiple sequence alignment, underlies its complementarity. The option of using sequence synchronisation invites researchers to combine events and

temporal patterns in the same analysis, thereby reducing the existing methodological gap resulting from the separate developments of event history and sequential approaches.

As regards theory, the synchronising approach seems consistent with two traditional concepts in sociology and social sciences in general: turning points, defined as probability regime shifts; and differentiation/homogenisation, identified by converging and diverging trends over the life course. The two illustrative cases provide compelling support to this view. The first case on functional differentiation of academic tasks throughout careers exhibited clear diverging patterns. The second case gave evidence of a clear probability regime shift between class careers prior to and following expatriation, especially amongst female white collars.

These illustrative cases, as different as they are, both relate to life course research. The scope of application of sequence synchronization, however, is not restricted to this field. We argue instead that this method can be useful for a variety of research areas, such as time use research, analyses of historical processes, music studies or linguistics.

This chapter only dealt with visualisation techniques. This does not mean that sequence synchronisation is not of interest for other ways of analysing sequences, such as dissimilarity measurements, clustering, dissimilarity trees and so on. One interesting question relates to applying optimal matching to sequences that have been previously synchronised. Unless the Levenshtein II coding scheme is used (Lesnard 2010), the outcome could differ from that obtained with non-synchronised sequence data. In some cases, optimal matching analyses could benefit from synchronisation, because they could lead to clusters that are better defined according to a structuring event, such as marriage, childbirth or incarceration. A second interesting question concerns analytic operations that could follow visualisation of synchronised data. Visualisation is indeed useful for identifying trends but has limited robustness to establish results. Statistical counting techniques and testing could be adapted to the type of hypotheses synchronisation allows to develop. One may, for example, compare transition rates or transversal entropy levels in a sample of sequences before and after a synchronised event (Gabadinho et al. 2011).

Sequence synchronisation entails substituting historical timelines or age-based time axes by event-based time axes. Such operation raises important issues, of which some were not addressed in this chapter. What are the epistemological and ontological implications of “detaching” individual sequences from historical time and/or from age-based time? Getting married at age 20 does not mean the same as getting married at age 70. The same applies for historical periods: getting married in 1900 does not mean the same as getting married today. Are there pitfalls to be avoided? Are there ways to “bring back” historical time in the analysis of synchronised sequence data? Cohort-based analyses might well offer a path to tackle this question.

References

- Abbott, A. (1997). On the concept of turning point. *Comparative Social Research*, 16, 85–106.
- Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology*, 96, 144–185.
- Becker, H. (1963). *Outsiders: Studies in the sociology of deviance*. New York: The Free Press.
- Blair-Loy, M. (1999). Career patterns of executive women in Finance: An optimal matching analysis. *American Journal of Sociology*, 104(5), 1346–1397.
- Blanchard, P. (2010). *Analyse séquentielle et carrières militantes*. Working paper.
- Blossfeld, H. P., & Rohwer, G. (2002). *Techniques of event history modeling: New approaches to causal analysis*. New York: Lawrence Erlbaum.
- Bras, H., Liefbroer, A. C., & Elzinga, C. H. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography*, 47(4), 1013–1034.
- Di Maggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review*, 48(2), 147–160.
- Doeringer, P. B., & Piore, M. J. (1971). *Internal labor markets and manpower analysis*. Lexington: ME Sharpe Inc.
- Dubar, C. (2010). *La socialisation: Construction des identités sociales et professionnelles*. Paris: Armand Colin.
- Gabadinho, A., Ritschard, G., & Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Gauthier, J. A. (2007). *Empirical categorizations of social trajectories: a sequential view on the life course*. PhD thesis, Faculté des sciences sociales et politiques de l'Université de Lausanne.
- Gauthier, J. A., Widmer, E. D., Bucher, P., & Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology*, 40(1), 1–38.
- Giudici, F., & Gauthier, J. A. (2009). Différenciation des trajectoires professionnelles liée à la transition à la parentalité en Suisse. *Revue suisse de sociologie*, 35(2), 253–278.
- Hughes, E. C. (1971). *Cycles, turning point and career*. In Hughes, E.C., *The sociological eye: Selected papers*, 124–131. Chicago: Aldine.
- Kaufmann, V. (2001). La motilité: une notion clef pour revisiter l'urbain ? In M. Bassand, V. Kaufmann, & D. Joye (Eds.), *Enjeux de la sociologie urbaine* (pp. 87–102). Lausanne: Presses polytechniques et universitaires romandes.
- Insee. (2003). *Histoire de vie—2003*. Paris: Centre Maurice Halbwach.
- Lemercier, C. (2005). Les carrières des membres des institutions consulaires parisiennes au XIX^e siècle. *Histoire et mesure*, 20(1/2), 59–95.
- Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3), 389–419.
- Lévy, J. (2003). Capital social. In J. Lévy & M. Lussault (Eds.), *Dictionnaire de la géographie et de l'espace des sociétés* (pp. 124–126). Paris: Belin.
- Pollock, G. (2007). Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1): 167–183.
- Robette, N. (2010). The diversity of pathways to adulthood in France: Evidence from a holistic approach. *Advances in Life Course Research*, 15(2–3), 89–96.
- Robette, N., & Thibault, N. (2009). Analyse harmonique qualitative ou méthodes d'appariement optimal? *Population*, 63(4), 621–646.
- Wagner, A. C. (2007). *Les classes sociales dans la mondialisation*. Paris: La Découverte.

Chapter 14

Graphical Representation of Transitions and Sequences

Christian Brzinsky-Fay

Introduction

The analysis of sequences within sociology originally emerged from the biographical perspective on humans lives (Abbott 1983). Since the 1980s, this qualitative approach shifted toward a more quantitative paradigm of life-course analysis (Aisenbrey and Fasang 2010). Both approaches have a longitudinal notion and are focussed on the investigation of causal effects within processes ordered by elapsing life-time. Whereas the qualitative biographical approach was concentrated on generating hypotheses from single individual histories, the quantitative life-course approach uses large datasets with many individuals trying to apply predominantly hypotheses-testing methods (cp. Heinz 2003; George 2009).¹ Sequence analysis, in its recent developmental stage, is a method which lies in between these approaches with its explorative nature and with its use of a large number of cases. Whereas the earlier applications used only a few cases with short sequences because of limited computer capacities, sequence analysis nowadays is able to deal with large case numbers and long sequences. Hence, it can be used to generate hypotheses and as a pre-stage for the application of hypothesis-testing methods.

Most sequences can be seen as discrete, categorical time-series.² In general, sequential information is complex with respect to three dimensions: first, there is a large number of individuals under investigation; second, studies applying sequence analysis tend to have long sequences, i.e., many time points; and third, they have a certain number of categories. This is what makes the visualisation of sequences

¹ One major objective of the life-course approach is the examination of the embedding of aggregate individual life courses into macro-level contexts and in the linkage between life-courses (cp. Elder et al. 2004).

² This is only true for those sequences based on a time-scale mostly used in sociology. But, there are also studies about sequences that have no time dimension, such as the original application of optimal matching (OMA) on distance or DNA strands.

a challenging task. The exploration of sequential data aimed at inductively deriving ideal types requires a heuristic process in which visualization of the data plays a major role. The visual approach to (graphical) information allows for important insights in relevant structures in the data. Hence, the appropriate visualization of sequential information is important not only for the communication between the analyst and others (as a presentation graph) but also for the methodological research process in itself (as an analytic graph): it constitutes a necessary step of the exploration of sequences. Cleveland (1994) argues with respect to graphical representations that this purpose also constitutes a necessary part of all forms of data analyses: “Numerical reduction methods do not retain the information in the data.” In this view, graphical depiction is a source of new insights: “To regularly miss surprises by failing to probe thoroughly with visualization tools is terribly inefficient [...]” (Cleveland 1994, p. 8). The importance of visual inspection of data is not limited to explorative methods but is also relevant for analytical research strategies, which is wonderfully demonstrated for regression diagnostics by the so-called ‘Anscombe’s quartet’ (Anscombe 1973).

A graph contains encoded information which must be decoded correctly by the reader. If the decoding fails—i.e. if information are obscured or misinterpreted—the whole visualisation process failed (Cleveland 1994, p. 1). Consequently, the creation of graphical displays must be conducted very carefully taking into account the kind of information that has to be transmitted as well as the concrete design of a graph.

This chapter is about the display of sequential information with a focus on socio-logical issues. It presents different kinds of graphical visualisations and discusses their typical advantages and drawbacks. This assessment is mainly based on a model of graphical perception as well as on elementary principles of graph construction and clear understanding developed by William S. Cleveland (1994).

The rest of this chapter is structured as follows: The next section briefly explains Cleveland’s model of graphical perception while concentrating on those principles that are relevant for graphical sequence representation. After that, different kinds of sequence visualisation are discussed, such as sequence index plots, status proportion plots, modal plots, parallel coordinates plots, and transition plots. The main objective is to provide researchers and readers with a deeper understanding of the appropriate application of these graphs.

Cleveland’s Model of Graphical Perception

Based on different kinds of psychological studies and perception tests, Cleveland developed a model of graphical perception that should help to explain why particular graphical realisations are better than others. Accordingly, he compiles a couple of principles that should help researchers enhance the potential of their graphs re-

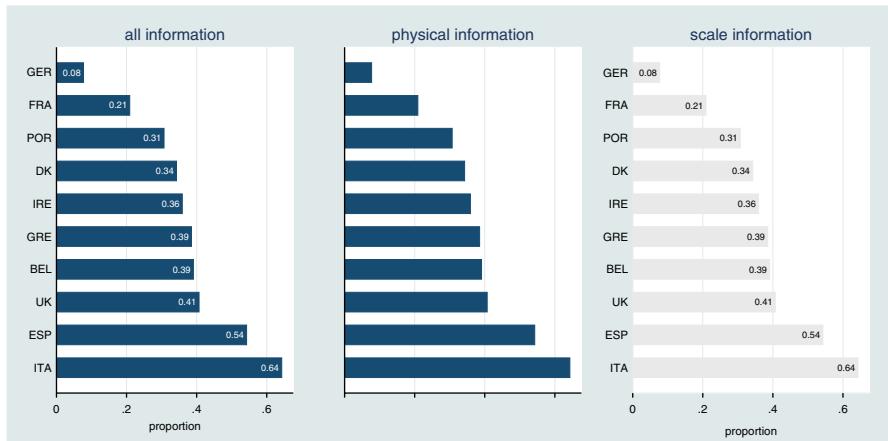


Fig. 14.1 Physical and scale information of a graph

garding information transmission.³ Those principles that are important for sequence visualisation are summarised in this section.

In general, the construction of a graph means encoding information, whereas graphical perception means the decoding of this information. The whole visualisation process fails if the decoding of the information fails (Cleveland 1994, p. 221 f.). All the Information of a graph can be distinguished into scale information and physical information (Fig. 14.1). Scale information is in the name of the category, while physical information is the information that is left when all labels, tick marks, and category names are removed. Physical information refers to the geometry of graphical elements used in a graph.

Scale information is decoded by the so-called ‘table look-up,’ while the visual decoding of physical information is the ‘pattern perception’. For both, efficiency is determined by the speed and accuracy in which the operations are carried out, and fast operations tend to be more accurate. Pattern perception consists of three operations, namely detection, assembly, and estimation. The detection refers to the recognition of the geometric aspects, such as the bars illustrated as having different length, and they are reaching from the left to the right. Assembly is the operation which groups the detected elements together, such as the bars that are showing the same values for different countries. Estimation relates the detected and assembled information by discriminating, ranking, and rationing, such as the longer bars in Fig. 14.1 are at the bottom of the graph. The table look-up also consists of three operations: scanning, interpolation, and matching. Scanning refers to the first recognition of a fix point on the scale, while interpolation describes the estimation of

³ There are many trials of assembling rules for graphical display (Tufte 1983; Wallgren et al. 1996; Blasius and Greenacre 1998; Good and Hardin 2003), but I decided for Cleveland’s principles, because his work is most comprehensive, that is that other rules are included in his aggregation.

the distance between the fix point and a neighbouring tick mark. The matching connects the estimation of the scale distances to the different data points.

These visual operations can be used to gain important information about how graphs are decoded by providing a couple of ideas regarding how to construct them appropriately. Cleveland himself uses them to derive a couple of principles for graphical display and understanding (see Appendix), of which some are of direct relevance for sequence visualisations.⁴ The first principle of graph construction—“Make the data stand out and avoid superfluity” (p. 25)—constitutes already a little problem for the purpose of graphical sequence representation, which is discussed in the respective sections. A second principle is the advice to “use visually prominent graphical elements to show the data” (p. 29). Very important for sequence index plots are the principles that refer to overlapping (p. 50) and to reduction and reproduction (p. 53). The requirement that calls for “superposed data sets [to] be readily visually assembled” (p. 51) is important when comparing sequences of different sub-groups. A basic principle of clear understanding—“Put major conclusions into graphical form” (p. 55)—is, of course, valid for all graphs but constitutes a challenge for explorative procedures. This principle, to a certain extent, contradicts two principles of general strategy, namely that “a large amount of quantitative information can be packed into a small region” (p. 110) and that “graphing data should be an iterative, experimental process” (p. 112). These principles are discussed in the following sections of the respective plots.

Sequence Index Plots

The most intuitive and widely used graphical format for the visualisation of sequential information is the sequence index plot. The three dimensions of sequence data (observations, time points, and categories) are plotted in one single, colourful graph. In its basic version, the horizontal axis represents the time dimension starting point on the left end of the scale and finishing on its right end. The observations are represented on the vertical axis so that each line shows the sequence of one single observation (mostly in reference to individuals). Each point in this coordinate system shows a particular category of a particular observation at a particular time point, while the categories themselves are reflected by a respective color. According to Cleveland (1994, p. 210), there are only two useful applications of colour in graphs, one of which is rendering different categories. Color allows for the quick perception of clear-cut differences in the distinct values of a categorical variable;

⁴ The majority of the principles are already implemented by the most statistical software packages, such as Stata. For example, the principle that the data rectangle should be smaller than the scale-line rectangle (Cleveland 1994, p. 31) is followed by Stata for every graphical command.

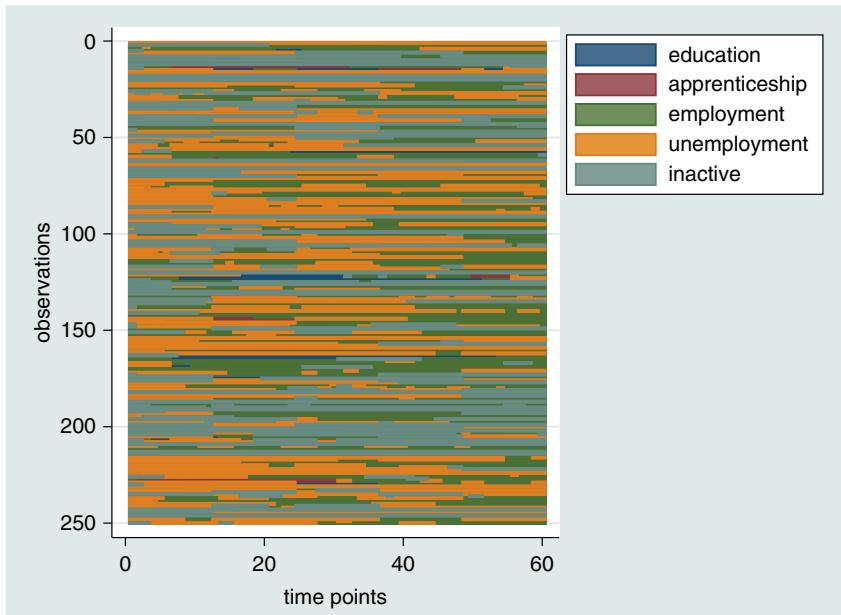


Fig. 14.2 Basic form of sequence index plot

therefore, it is clearly suitable for the indicating different statuses of sequences. An example for a sequence index plot is show in Fig. 14.2.^{5,6}

The advantage of this graph is its intuitive perception: time is on the horizontal axis, individual sequences have to be read from the left to the right, and the differentiation of the categories is clear because of contrasting colors. In a sequence index plot, the data stand out and they are visually prominent; at the same time, a large amount of information is packed into the graph. Therefore, it serves the purpose to visually detect structures in the bulk of data, and is useful within the iterative, experimental process of sequence exploration. This kind of graph clearly helps to detect unknown characteristics and structures, because “our mind’s eye frequently does not so a good job of predicting what our actual eyes will see” (Cleveland 1994, p. 112). The graph can be modified for the purposes of result communication by changing the colors in a way that similar categories look similar because of lower contrasts.

However, there are also some clear drawbacks of this kind of visualization that arise from the increase in the number of observations or in the number of categories.

⁵ The graphs of this chapter are generated with the software Stata and the user-written sq-commands (Brzinsky-Fay et al. 2006); the data are mainly the same as in Brzinsky-Fay (2007).

⁶ Figures 14.2 and 14.3 show monthly employment status sequences of 250 randomly chosen school leavers from a sample of around 4,000 school leavers in ten European countries. The sequences consist of 60 months (12 years) and are from the European Household Panel (ECHP).

A considerable increase in the complexity in one or both of these two dimensions leads to the graph becoming cluttered and not interpretable. This circumstance refers to a violation of the principles in that overlapping symbols must be distinguishable and visual clarity must be preserved under reduction and reproduction. Most of these problems can be solved by modifying the graph in an appropriate manner.

In a sequence index plot, each line represents a chronological sequence of statuses for a given analytical unit. With the increasing number of observations, the lines that have to be displayed in the graph (and whose width is calculated) become thinner. Technically, this constitutes a problem as soon as the calculated line width is smaller than the smallest possible pixel of the graph. In this case, earlier drawn lines are overplotted by later drawn lines.⁷ As a result, some of the sequences are not or not fully visualized and the information in the graph is distorted, because it reflects not the full information anymore. Mainly, there are two strategies to solve this problem of overplotting: (1) reducing the number of observations displayed in the graph, and (2) subtly influencing the order of a sequence.

The first strategy is the reduction of the observation number, which can be achieved in different ways. One can select a random sample from the sequences and display the sequences in a standard sequence index plot. A sequence index plot is able to show between 300 and 400 lines in each graph without overplotting. This strategy clearly inhibits overplotting because with a lower number of cases, the single lines become thicker. However, even a random selection of the sample has a certain likelihood of not reflecting all properties of the original dataset, so that characteristics can be blurred because of sampling. Hence, the sampling should be repeatedly conducted in order to control this effect. Alternatively, the number of observations can be easily reduced by creating one sequence index plot for each category of a particular breakdown variable (see Fig. 14.5). Another possibility is the display of those sequences which are most frequent in the data.⁸ The disadvantage with this variant is that it disregards a lot of information of the data, and it does so systematically by using the frequency as a selection criterion. Whereas in a random sample of the sequences each sequence has the same probability of being displayed in the graph, more frequent sequences have a higher probability. In social science applications of sequence analysis, very often the less frequent sequences are of high interest.

The second strategy refers to the order of the sequences. The negative effect of overplotting can be diminished if the overplotted sequence is equal or similar to the overplotting sequence. This occurs if similar sequences are plotted beside each other. The respective order must be based on a variable that accounts for similarity between the sequences. While in Fig. 14.2 the sequences are displayed in random order, in Fig. 14.3, they are ordered according to the membership in a sequence cluster, which is the result of the application of optimal matching and clustering

⁷ In Stata, the graph command plots the graph category by category, so that categories (statuses) with a higher number overplot those with a lower number.

⁸ This is implemented in the TraMineR package (Gabadinho et al. 2011a; b) for the software R under the name of ‘sequence frequency plots’.

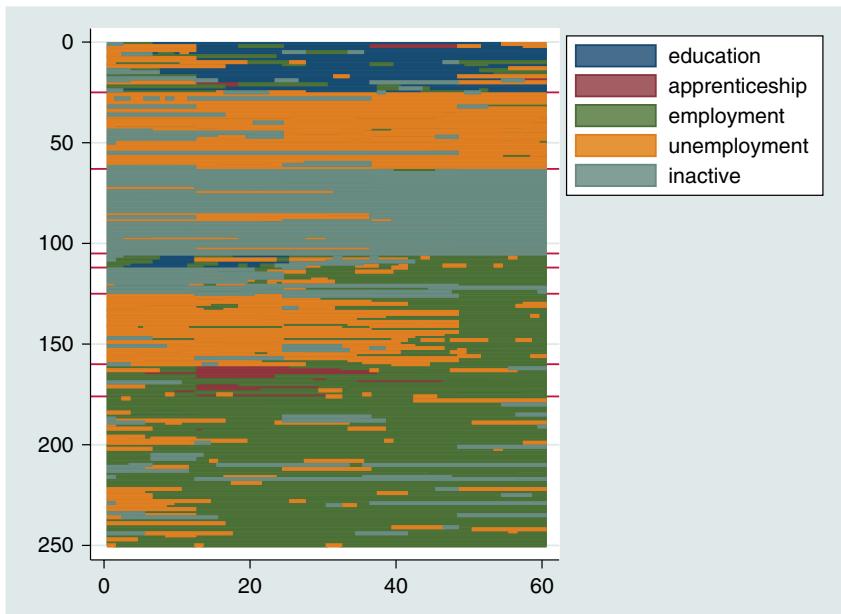


Fig. 14.3 Sequence index plot with ordered sequences

procedure. The observations are the same in both figures, but the clarity regarding the transition types is much higher in Fig. 14.3. Alternatively, the scores of a multi-dimensional scaling may be used to order a sequence index plot (cp. Piccarreta and Lior 2010; Gabadinho et al. 2011a; Fasang and Liao (Forthcoming)).

An interesting combination of reducing the number of observations and changing the order is the so-called representative sequence plot proposed by Gabadinho et al. (2011a). Based on criteria of sequence similarity, the only sequences that are displayed are those representative of multiple sequences from the dataset. Similarly, Piccarreta (2012) provides smoothing techniques, which summarize similar sequences by creating artificial sequences that, while not in the data, are representative of a group of similar sequences. Both techniques solve the problem of overplotting quite well. However, they require pre-conducted comparison procedures and the definition of the selection criteria are quite complex and, to a certain degree, remain arbitrary. Apart from that, the visualized information deviates remarkably from the original data.

The problem of underplotting occurs if a sequence index plot with too few observations is drawn—i.e., if the number of lines (sequences) is much smaller than the number of lines available within a graph. In Fig. 14.4⁹, three sequence index plots are drawn according to the breakdown variable ‘country’. The number of ob-

⁹ Figures 14.4 and 14.5 show monthly employment sequences from school leavers in Belgium, France, and Spain.

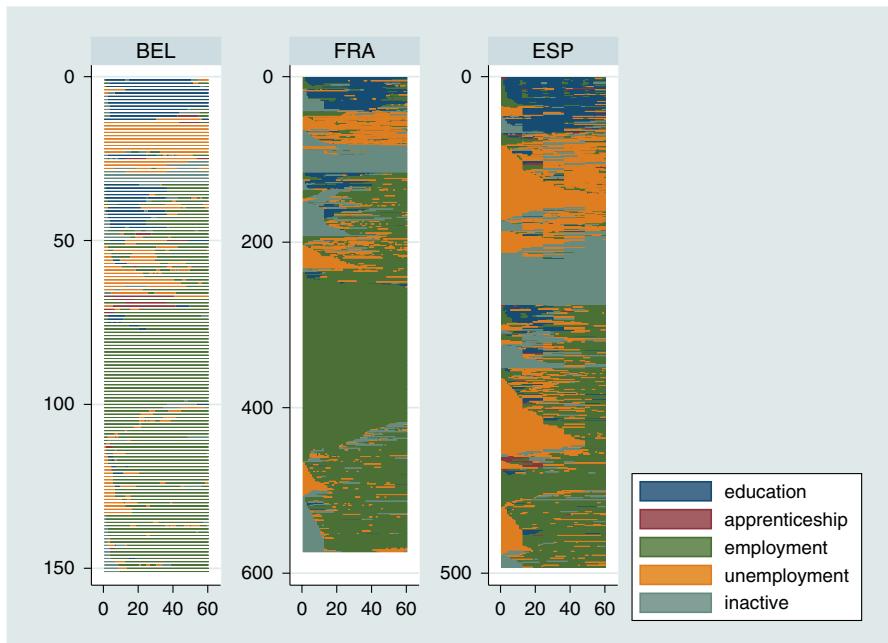


Fig. 14.4 Example for underplotting

servations and the number of lines differ between each of the categories (Belgium, France, Spain). The gaps between the Belgian sequences make the graph nearly unreadable, but they can be changed easily by drawing wide bars instead of thin lines (Fig. 14.5).

Despite customizing the sequence index plot in order to enhance its quality, some of Cleveland's graphing principles remain unsolved: first, only to put major conclusions into graphical form, and second, the preservation of visual clarity under reduction and reproduction. In regard to major conclusions, the violation of the principle stems from the explorative nature of sequence index plots which requires the display of more information than necessary in order to find hidden structures in the data. Apart from this issue, the limitation to major conclusions seems to apply to a larger extent to presentation graphs rather than to graphs serving an iterative research procedure. For the latter case, conclusions have yet to be drawn and the visualisation is needed to find results.

Preserving the visual clarity under print reproduction continues to be a challenging task. The reason can mainly be found in the production and printing process of articles and book chapters (such as the chapter at hand). While the researcher creates graphs on the screen using a particular software, the printing process occurs at a different location in a separate organization using different soft- and hardware. This frequently leads to incompatibilities between different graph formats or at least limits the necessary iteration among constructing the graph, screen view, printing

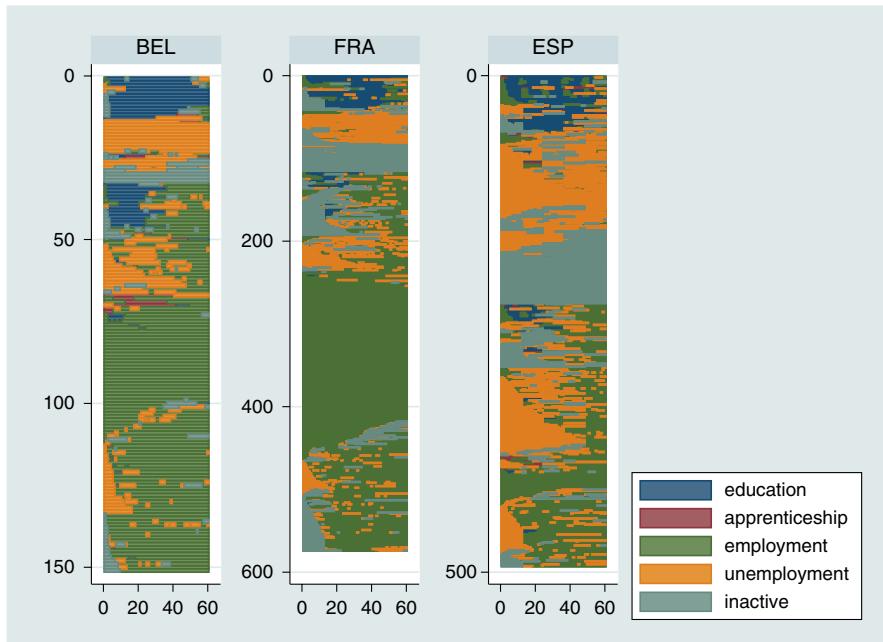


Fig. 14.5 Solution for underplotting

and changing the graph. Technically, it is recommended to produce graphs in either Portable Document Format (.pdf) or PostScript (.ps,.eps) in order to avoid severe damage in the appearance of sequence index plots.

Status Proportion or State Distribution Plots

In reference to one of the principles of clear understanding, namely that only major conclusions should be put into graphical form, the information of a sequence index plot can be condensed for a clearer interpretation of a particular characteristic. A status proportion or state distribution plot displays the relative proportion of each category (status) for every time point. On the horizontal axis, time is displayed, whereas the vertical axis is a percentage scale. Figure 14.6 shows a status proportion plot that uses the same data as Fig. 14.5. It is concentrated only on two of the three dimensions—time and categories—while giving up the observation (individual) dimension. Thus, the construction of a status proportion plot means one step beyond the principle of displaying a large amount of information to the principle of putting only major conclusions into graphical form. The major conclusion here is that in France and Belgium, the share of employment statuses across time is quite similar, whereas in Spain, the monthly share of these statuses differs clearly. The

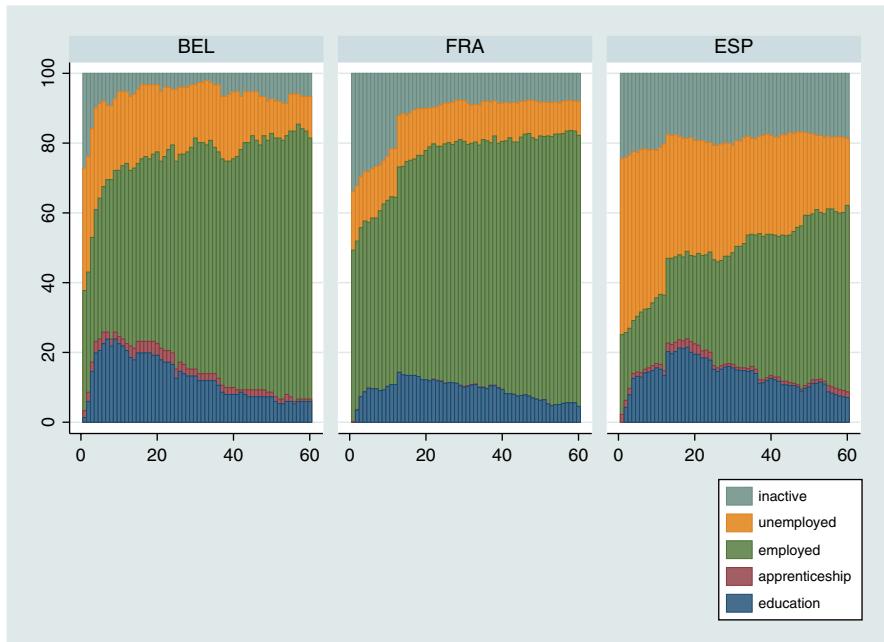


Fig. 14.6 Status proportion plot

reader has to keep in mind that there is no information in this graph concerning individual sequences. The horizontal lines in the graph do not have any meaning in this kind of graph, whereas in the sequence index plot, each horizontal line represents one observation (individual).

Status proportion plots have the advantage of avoiding the violation of the principles of overlapping and preservation of visual clarity under reduction and reproduction. Over- and underplotting does not play a role in this graph. If the colors are chosen in order to provide high contrast¹⁰, changes in size or conversion into black-and-white do not harm the readability. However, this comes at the price of losing the individual information.

Figure 14.7 shows a status proportion plot of the same data used in the other figures but, in this case, sorted differently. Here, the panels of the graph are not showing different countries; rather, they indicate different clusters that are the result of a sequence comparison with optimal matching and a subsequent clustering procedure. The sequence differences between clusters were maximized whereas

¹⁰ In Stata, the standard colors are chosen in order to provide more contrast on color screen and when converted to black-and-white. However, apart from the explorative research procedure being an iterative, experimental process, the choice of the best color scheme is also iterative, and the proof of the pudding is in the eating. For more discussion on the choice of colors for representation of categorical data, see Zeileis et al. (2009).

the differences within clusters were minimized. Therefore, each panel represents a kind of ideal type composed by very similar sequences. Again, the status proportion plot does not display individual sequences; it shows status proportions at each time point: months, in this case. The principle that superposed data sets must be readily and visually assembled is followed by the display of a status proportion plot for each category of the breakdown variable in its own panel.

Modal Plots

A further step into the direction of higher abstraction and information reduction (i.e., showing only major conclusions in the graph) is enabled by constructing modal plots. The modal plot is a summary of the information depicted in a status proportion plot but in a simplified manner. For each time point, the most frequent category (the mode of the distribution) of the sequence is plotted with a horizontal stacked bar chart. In other words, each panel of Fig. 14.7 has now become a horizontal stacked bar in Fig. 14.8. The information of the other categories is discarded because it ignores the variance of categories (statuses) on each time point. But, this graph has the advantage of presenting the clearest possible graphical form of what an ideal type looks like. Bar charts already conform to most of the principles. The principle of putting major conclusions into graphical form is maximized with this kind of graph. Similar to status proportion plots, there are no problems with over- or underplotting, and the visual clarity is maintained under reduction and reproduction, provided that the contrast of the chosen colors survives the conversion into the black-and-white graph.

Sequence index plots, status proportion plots, and modal plots represent different kinds of graphical displays, but they constitute a continuum at the same time. While sequence index plots contain the highest amount of information with a relatively low degree of clarity, modal plots, at the other side of the continuum, contain the lowest amount of information but show the highest clarity. Status proportion plots are located between these two graph types. However, it is not recommended to use only one of these plots because each plot emphasises a focus on different properties of sequential information (Fig. 14.9).

Parallel Coordinate Plots

While the last three graph types are interconnected and are very much part of an integrated explorative research process, parallel coordinate plots constitute a kind of ‘stand-alone-display.’ The parallel coordinate plot depicts an aggregation of individual sequences and can be used to answer questions for the most frequent se-

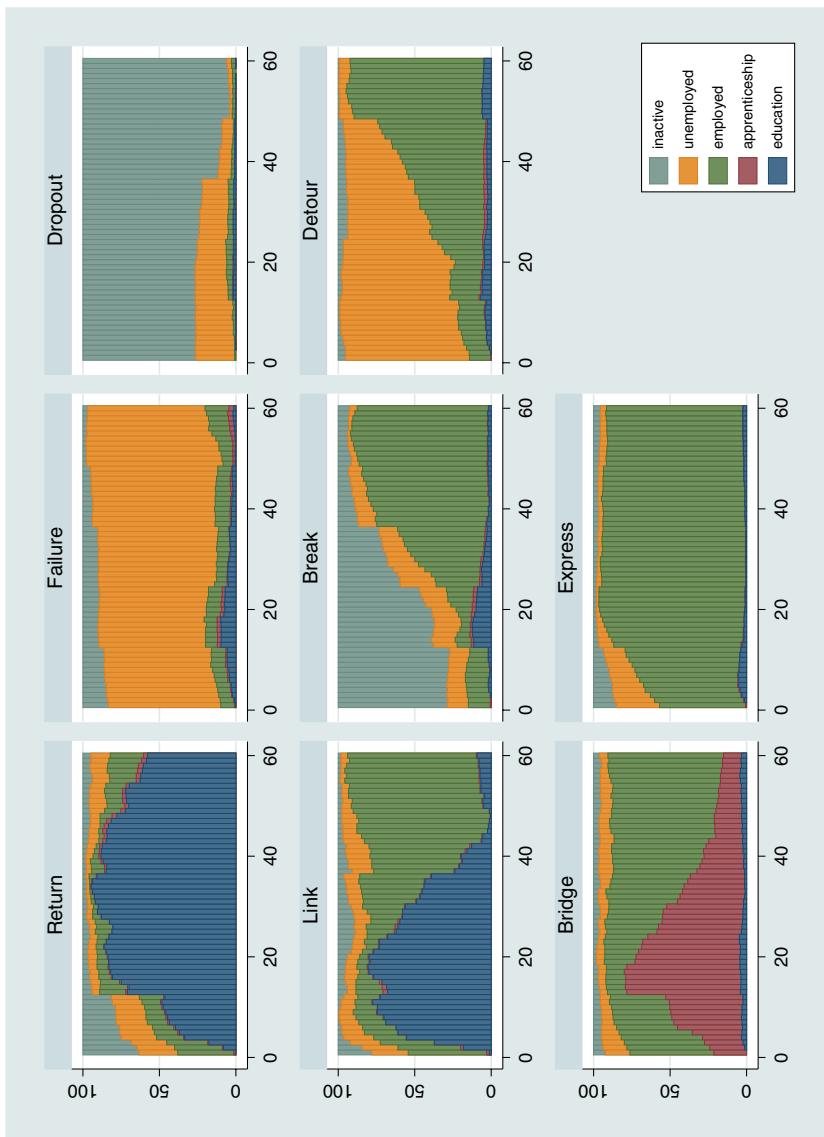


Fig. 14.7 Status proportion plot of sequence types

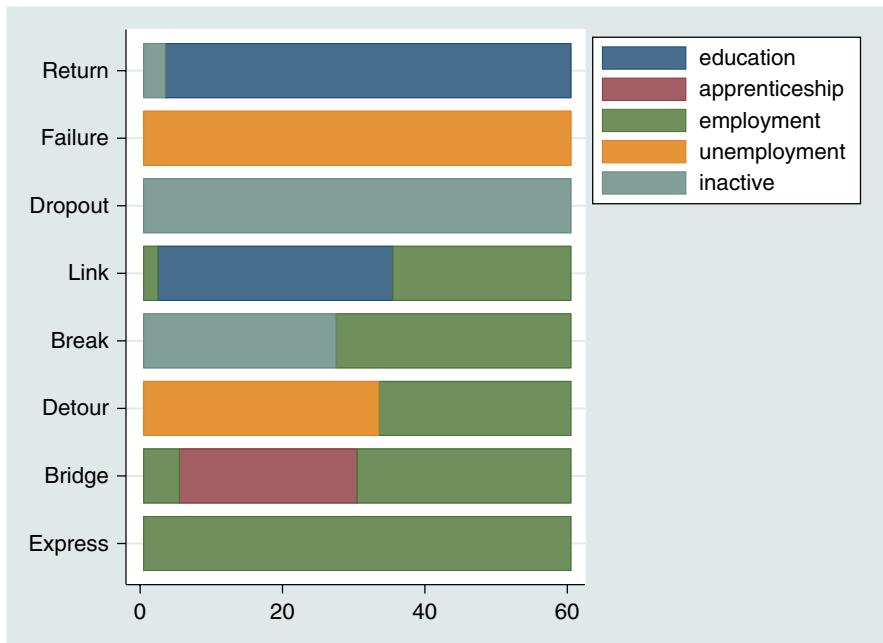
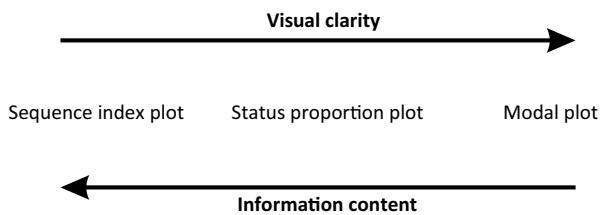


Fig. 14.8 Modal plot of sequence types

Fig. 14.9 Visual clarity and information content of sequence index, status proportion and modal plots



quences. This information can be also gained by simple tables.¹¹ Parallel coordinate plots originally were invented in order to display multidimensional information as an alternative for the Cartesian coordinate system (cp. Theus 2006). They are suitable for the display of sequential information because they can handle more than two dimensions.¹² An example for the data of the Figs. 14.2–14.8 is given in

¹¹ The user-written package TraMineR available for the software R (see footnote 8) offers the possibility to produce so-called sequence frequency plots (Gabadinho et al. 2011a). These plots also show most frequent sequences, and have the form of a weighted bar chart.

¹² However, the original application differs a bit regarding the structure, because each of the vertical axes refers to another dimension, whereas the usage in sequence analysis mixes the Cartesian system with the ‘real’ parallel coordinates in a way that the time points are represented by the horizontal axis.

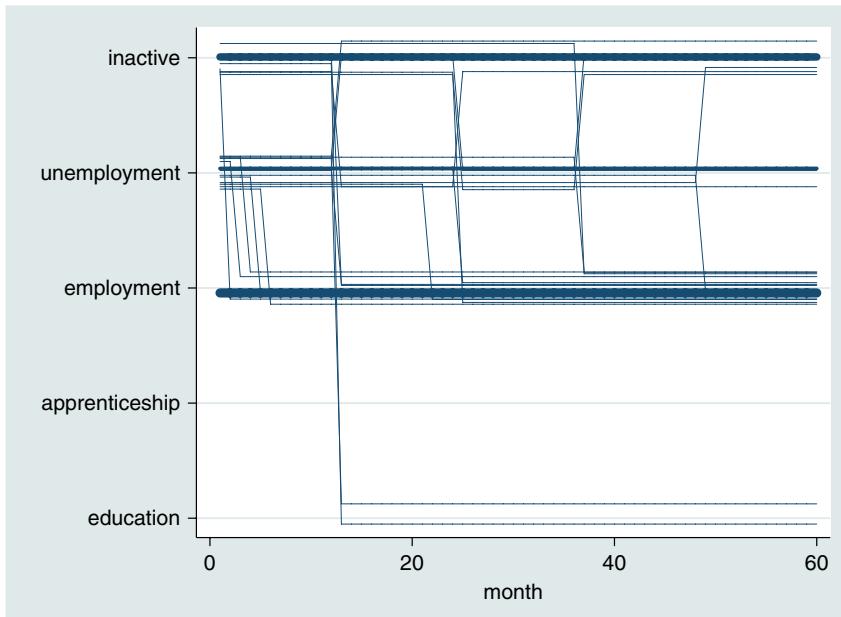


Fig. 14.10 Parallel coordinate plot

Fig. 14.10. On the vertical axis, we find the categories of the sequence states, while the time points are represented by the horizontal axis. The third dimension—the observations—is shown by the lines in the data rectangle. The concession with complexity is the fact that the lines only represent the most frequent sequences in the data set. The number of lines can be defined by the researcher, and their thickness represents the relative frequency. In other words, the thickest lines show the most frequent sequences.

The advantage of this form of graphical display is that the data clearly stand out and the graphical elements (i.e., the lines) are simple and visually prominent. Researchers can easily avoid dispensable information by limiting the display of the lines. Overlapping lines do not constitute a problem because they can be drawn in a jittered way. Major conclusions can be put into graphical form appropriately. Since no colours are necessarily to be used in this graph and the contrasts are very clear, reduction and reproduction don't constitute a problem for visual clarity.

The upper two panels of Fig. 14.11 shows a parallel coordinate plot of the same data, which shows sequence frequencies of two superposed data sets—i.e., two subgroups (males and females). Because the categories are at same vertical position, the principle of easy visual distinction is fulfilled as well. The lower two panels of Fig. 14.11 show an alternative version of this graph, which summarizes the individual sequence data in a slightly different way. The upper two panels show and summarize the sequences with respect to concrete time points, which can be referred to as ‘time similarity’. For example, a person A with a sequence that is composed of 10

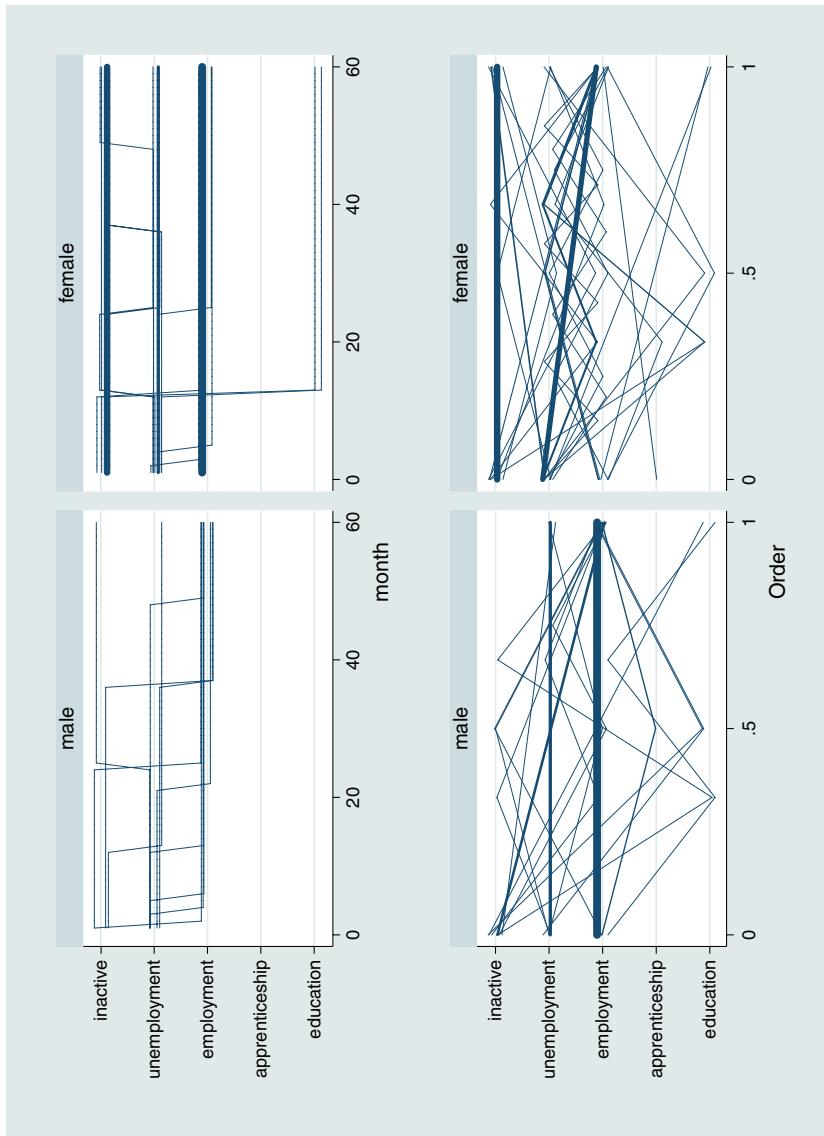


Fig. 14.11 Parallel coordinate plots by gender

months of unemployment followed by 50 months of employment is displayed in a different line than a person B with a sequence that is composed of 20 months of unemployment and 40 months of employment. Both have the same order of statuses, namely unemployment-employment, but because the lengths of the periods differ between person A and person B, they are considered as being different.

In contrast, the parallel coordinate plots in the lower two panels only consider the order of the sequences, while ignoring the absolute length of the episodes. It refers to the so-called ‘same-order similarity’. Here, person A and person B are seen as equal, because the order of statuses is the same. This leads to the result that both sequences are shown in the same line within the parallel coordinate plot, because the similarity refers only to the order—unemployment-employment—which is the same for both persons. Hence, both sequences are summarized in different lines in the upper two panels, while they are summarized in the same line (and constituting its thickness) in the lower two panels.

Regarding the principles of graph construction and clear understanding, there is no difference between the same-order similarity and the time similarity. However, it should be kept in mind that with parallel coordinate plots, there is a substantial loss of information involved because only the most frequent sequences are displayed.

Transition Plots

The transition plot is also helpful for the analysis of sequential information. Transitions are simply changes in the category within the sequence of an observation. They do not genuinely reflect sequential information because they count the number of transition between certain states while disregarding the time dimension and the observation dimension. Figure 14.12 shows a transition plot of the sequence data that also were used for the other figures. The horizontal and the vertical axis show the categories (statuses) of the sequences. Where the reference (dotted) lines do cross, a bubble shows the relative frequency of the transition between the respective statuses from one time point to the other. This graph summarizes all transitions at each time point, but it can be modified so that the transitions for particular time points can be shown. In this graph, the higher the frequency of a transition, the larger the corresponding bubble; hence, the most frequent transition is between employment statuses. The bubbles on the diagonal mean that there is no change or status stability. If one is only interested in changes, then it is helpful not to consider changes between equal statuses—i.e., to not show the bubbles on the diagonal, as it is shown in Fig. 14.13.

The principles are followed in the graphs presented here. Date stand out, and there is no dispensable information in the graph. The graphical elements are visually prominent, overlapping does not exist, and visual clarity is preserved under reduction and (print) reproduction.

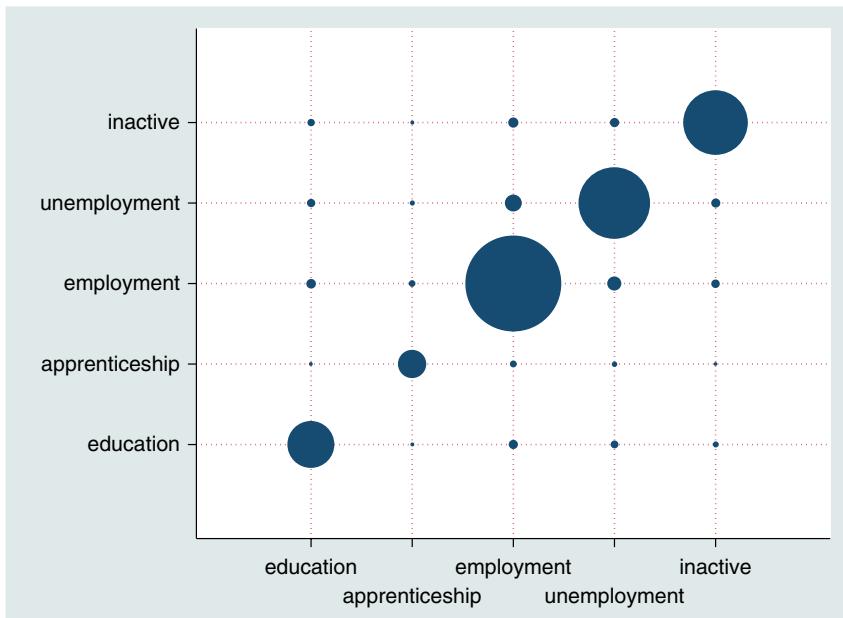


Fig. 14.12 Transition plot with diagonal

Summary

In this chapter, the most useful and applied methods of visualization of sequential information were presented, and their advantages and drawbacks were discussed with respect to principles of graphical display and clear understanding. As categorical time series, sequences contain rich information about individual processes across time, which needs to be presented carefully during the research process of sequence analysis as well as in the end when it comes to result communication. Because of their explorative nature, graphs have a noticeably high relevance even within the research process. Researchers always need to face the trade-off between constructing graphs that are concise in the results they show and that allow for surprising findings regarding hidden structures. The objective of the chapter was to provide sequence analysts insight into graph construction as a means to help them solve this trade-off.

The classical graph for showing results of sequence analyses is still the sequence index plot, predominantly because of its intuitive structure. It fulfills most of the principles of graph construction and clear understanding, but it shows practical problems when it comes to publishing these graphs in printed media. There are some remedies available that can be applied with the statistical software packages able to conduct sequence analysis. Status proportion plots and modal plots are naturally strongly connected to sequence index plots and help to extract important

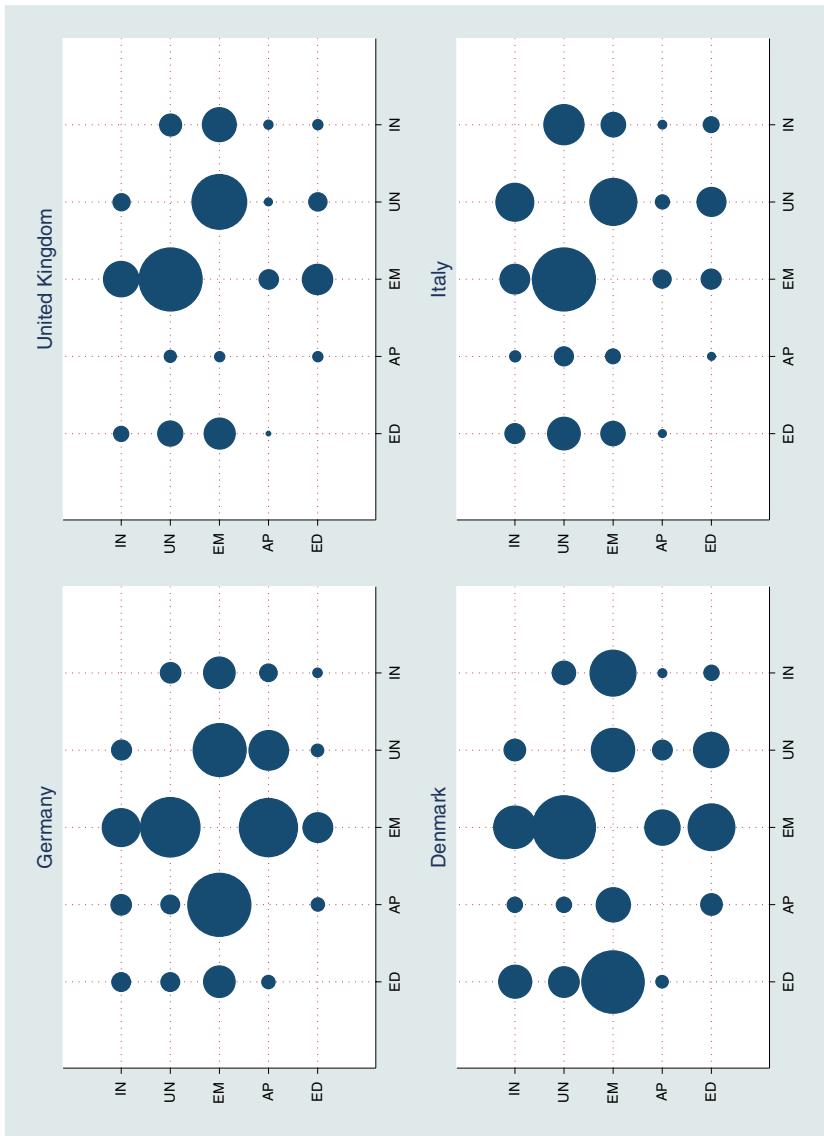


Fig. 14.13 Transition plots by country without diagonal

pieces of information from them. These three graph types constitute a continuum in which the classical trade-off between complexity or information content and visual clarity could be exemplified. The higher the amount of information, the harder it becomes to visualize the results.

Parallel coordinate plots and transition plots are rarely used by sequence analysts, although they provide interesting and valuable information about sequences and transitions. They also follow the principles of graph construction and clear understanding to a larger extent than sequence index plots do. The graphs presented and discussed constitute a rich body of possibilities to successfully conduct an important part of exploratory sequence analysis. Or, as Tukey (1977) explains: They help you “to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.”

Appendix

This list contains a selection of graphing principles developed and assembled by Cleveland (1994). It is limited to those principles that are relevant for the graphs usually applied to sequence analyses.

Elementary principles of graph construction:

- Make the data stand out & avoid superfluity.
- Use visually prominent graphical elements to show the data.
- [...]
- Overlapping plotting symbols must be visually distinguishable.
- Superposed data sets must be readily visually assembled.
- Visual clarity must be preserved under reduction and reproduction.

Principles of clear understanding:

- Put major conclusions into graphical form.
- [...]

General strategy:

- A large amount of quantitative information can be packed into a small region.
- Graphing data should be an iterative, experimental process.
- [...]

References

- Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods*, 16, 129–147.
- Aisenbrey, S., & Fasang, A. (2010). New life for old ideas: The “Second Wave” of sequence analysis bringing the “Course” back into the life course. *Sociological Methods & Research*, 38, 420–462.

- Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27, 17–21.
- Blasius, J., & Greenacre, M. (Eds.). (1998). *Visualization of categorical data*. London: Academic Press.
- Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review*, 23, 409–422.
- Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis using Stata. *Stata Journal*, 6, 435–460.
- Cleveland, W. S. (1994). *The elements of graphing data*. Summit: Hobart Press.
- Elder, J., Glen, H., Kirkpatrick Johnson, M., & Crosnoe, R. (2004). The emergence and development of life course theory. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the life course* (pp. 3–19). New York: Springer.
- Fasang, A. E., & Liao, T. F. (Forthcoming). Visualizing Sequences in the social sciences: Relative frequency sequence plots. *Sociological Methods & Research* doi:10.1177/0049124113506563.
- Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011a). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40, 1–37.
- Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2011b). *Mining sequence data in R with the TraMineR package: A user's guide*. Geneva: University of Geneva.
- George, L. K. (2009). Conceptualizing and measuring trajectories. In G. E. Elder & J. Z. Giele (Eds.), *The craft of life course research* (pp. 163–186). New York/ London: Guldiford Press.
- Good, P. I., & Hardin, J. W. (2003). *Common errors in statistics (and how to avoid them)*. New York: Wiley.
- Heinz, W. R. (2003). Combining methods in life-course research: A mixed blessing? In W. R. Heinz & V. W. Marshall (Eds.), *Social dynamics of the life course* (pp. 73–90). New York: Aldine de Gruyter.
- Piccarreta, R. (2012). Graphical and smoothing techniques for sequence analysis. *Sociological Methods & Research*, 41, 362–380.
- Piccarreta, R., & Lior, O. (2010). Exploring sequences: A graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society (Series A: Statistics in Society)*, 173, 165–184.
- Theus, M. (2006). Statistical Graphics. In A. Unwin, M. Theus, & H. Hofmann (Eds.), *Graphics of large datasets. Visualizing a million* (pp. 31–54). New York: Springer.
- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire: Graphic Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley Pub.Co.
- Wallgren, A., Wallgren, B., Persson, R., Jorner, I., & Haaland, J.-A. (1996). *Graphing statistics & data. Creating better charts*. Newbury Park: Sage.
- Zeileis, A., Hornik, K., & Murrell, P. (2009). Escaping RGBland: selecting colors for statistical graphics. *Computational statistics and data analysis*, 53, 3259–3270.

Chapter 15

Patterns of Contact Attempts in Surveys

Alexandre Pollien and Dominique Joye

Understanding Participation

This study aimed at measuring and understanding survey participation using contact attempt patterns. Survey specialists are becoming increasingly concerned about the growing levels of non-response in surveys, even though efforts to reach people have increased (Groves 2006; Groves and Peytcheva 2008). In addition, the transition from fixed landlines to mobile telephones, coupled with the spreading use of the Internet as a mean of communication, has complicated the work of contacting respondents. Methodologists are considering alternative forms of surveys, such as mixed-mode studies. They are also shifting their attention towards new indicators of quality, following the Total Survey Error paradigm (Groves and Lyberg 2010). The Total Survey Error approach considers all forms of survey error, along with the costs associated with reducing them. Through this method, survey quality is no longer measured solely by the criterion of response rate.

This study aims to complement existing research by proposing to re-establish a sociological approach to survey participation (Pollien and Joye 2011) in order to better understand factors for survey participation. Understanding survey participation should help with understanding the root causes of the non-responses and adjusting samples according to social characteristics.

What About the Quality of Data?

To assess the effects of survey non-response, data about the reference population is required. Many studies have shown that inference of non-response with information restricted to respondents remains problematic (Stoop et al. 2010). Imputation of non-response involves getting as much information about the entire sample frame as possible. Thus, controlling the non-response bias can be achieved by using four

A. Pollien (✉) · D. Joye
FORS, UNIL, Lausanne, Switzerland
e-mail: alexandre.pollien@fors.unil.ch

main sources: sample registry data, indicators related to the geographical location, data collected by the interviewer during the contact procedure, and the non-response survey. The last source covers only part of the sample. We distinguish, in this study, the analyses according to the degree of coverage of the information sources.

Register data covers the whole sample frame; however, the information is limited to administrative fields, including age, gender, civil state, nationality, and address. While these data are interesting, they do not always relate to the variables of interest. Methodologists have also developed analyses based on observations of interviewers during visits (Groves and Couper 1998; Peytchev and Olson 2007). The propensity to cooperate is analysed with environmental variables, such as housing. However, the information provided by these data is limited to observable characteristics, excluding attitudes or opinions.

A non-response survey, an ad hoc study focused on non-respondents, can offer more details on respondents, including education, activities, attitudes, and values. Even if useful, this survey is limited, as it does not provide information about the entire sampled population: it remains a part of the sample that has not participated. If used to compute propensity scores, non-response surveys assume that the persons who did not cooperate in the main survey are not so different from those who participated in the non-response survey. Moreover, this survey follows separate processes, with differences in time and interview modes (face-to-face versus paper and pencil), both of which need to be controlled.

The incompleteness of one method and the limitations of the other stimulate investigations for a new way of analysing survey participation through sequences of contact attempts. Matsuo et al. (2006) have developed a method that focuses on the history of the contact procedure and the role played by the number of contact attempts. The contact process was studied under the event history analysis approach, distinguishing between failed and successful contact. The authors stress the importance of considering the full history of the contact procedure. However, while they focused on counting the elements, they neglected to address the combination and order setting up the sequence of contact. Kreuter and Kohler (2009) considered the trail of paradata. They noted that researchers had thus far neglected to address information pertaining to the history of contact. The meaning of contact attempts change according to the order in which they occur. For example, the definition of non-contact is altered according to whether an appointment has been confirmed in advance.

This study extends this approach, including processual constitution of survey participation.¹ Kreuter and Kohler (2009) restricted their study to the possibility of rectifying the sample with contact data. Our hypothesis states that the sequence of contact consists of a series of interactions with the interviewer concerning participation. This sequence is related in particular to the characteristics of the target

¹ Even if the perspective is quite original, it is not a completely revolutionary idea. In the Fourth Conference of the European Survey Research Association (ESRA), Lausanne, 2011, a number of the presentations discussed the concept. Please see, among others, presentations by Romano and Nalli; Laflamme and Bilocq; and Halbherr et al.

person, e.g., her or his lifestyle. In other words, the way in which the respondent is contacted and agrees to cooperate is a sociologically defined process of involvement in social activity.

We must first address the theories accounting for participation in surveys, through the dimensions of accessibility and cooperation.

The Pursuit of the Respondent: Accessibility and Cooperation

To account for participation, the literature about survey methodology distinguishes between two dimensions: the ease of reaching target persons and the propensity to cooperate (Lynn and Clarke 2002). In face-to-face surveys, the interviewer's first challenge is to reach the target person. Participants may be difficult to locate, and access to physical locations may be restricted by various obstacles, such as entry codes or watchdogs. Even if the address is reached, the contact may still fail due to the absence of the target person. The difficulties of accessibility can be summed up by constraints of housing and problems of coordination between the interviewer and the target person. It is now established that socio-demographic factors characterise the ease of being reached (Keeter et al. 2000). Authors have noticed that the accessibility of respondents is linked to lifestyle, i.e. age, marital status, employment, household size, and housing (Lynn and Clarke 2002). For example, according to Keeter et al. (2000), younger and more highly educated respondents are more difficult to reach.

Once contact is made, the interviewer seeks to obtain the cooperation of the target person. The difference between the two dimensions of non-response can be ambiguous. Access issues are mostly due to social activities that lower the possibility of being contacted at home. However, some access issues can be related to an isolation attitude toward requests for social cooperation. This attitude is reflected both in housing and in behaviour toward others. Refusing to participate can take the ambiguous forms of conflicting appointments, health issues, or language inability. With expressed and motivated refusal, Dunne et al. (1997) proposed to distinguish between general unwillingness and reluctance linked to specific topics. In other words, inclination to cooperate is structured by different contexts of openness within the environment, from general accessibility to propensity to grant a specific request.

Cooperation is related to attitudinal profiles and social integration, namely general or specific interest regarding the request and social participation (Groves et al. 2000; Groves et al. 2004). To understand the inclination to cooperate, the following issues need to be considered at several levels:

At the macro level, surveys are conducted in a society at a given time. This involves considering the perceived legitimacy of surveys (Beatty and Herrmann 2002). Social representations of research institutions can encourage or discourage the propensity to cooperate, as can the topic of the survey. In summary, participation is related to the social representation of utility and trustworthiness of the survey, which is gradually revealed through the notification letter and information given by the interviewer at each step of the contact process.

At the meso level, cooperation depends on how the target person perceived the social interaction of the interview (Nolen and Maynard 2013). The perception of social solicitation can differ according to the timing of contact attempt: night, meal-time, or office hours. The sequence is rolled out through progressive adjustments made by the interviewer, who aims for the best moment (*kairos*) to contact (Pollien and Joye 2011).

At the micro level, participation stems from a contact process that consists of a series of (shorter or longer) interactions between the respondent and the interviewer. Among a whole set of inducements, the saliency of the survey theme and incentives are crucial with regards to cooperation. The series of contact attempts outlines relational sequences of negotiations and influences on target persons.

Leverage-Saliency Theory versus the Sociology of Patterns of Contacts

The manner in which different elements of motivation and inhibition are combined and lead together to cooperation is commonly theorised by using economic decision models. Groves et al. (2000) have proposed to describe the participation in survey as choice made through rewards and burdens. This model postulates that the incentive to cooperate consists of greater or smaller remunerations of various kinds—topic, civic attitude, monetary, and so on—which are of more or less interest to the respondent. On the other side, the disincentive to cooperate is mainly determined by the burden of the interview task and various social inhibitions, insecurities, ignorance, and so on.

According to the leverage theory of Groves and colleagues, there is a point at which the balance between rewards and burdens is determined: the target person weighs the pros and cons. Despite conceptual clarification of the survey, such a model underestimates temporal strategies and procedural decision-making regarding participation: a rational choice model cannot account for changes if there is no change in the arguments, e.g., when participants ‘throw in the towel’. It seems inaccurate to sum up participation through a one-time decision. Some refusals are latent and formulated as adjournment (‘call back later’). In an acquaintanceship that is built upon contact, a social exchange is expected (Dillman 1978). As noted by Bourdieu (1980), the temporality plays a crucial role in the relationship of debt and duty. Time is one of the strategies employed by the interviewer, who tries to put the respondent in a ‘debt’ of participation and make contact at the right moment. The interviewer would prefer an adjournment to a refusal, as the latter is difficult to alter (Shuttles 2008; Joule and Beauvois 2002).

The sequence of contact is part of the process whereby the target person forms an opinion concerning survey participation. Furthermore, economic considerations about cooperation imply the two dimensions of accessibility and cooperation are considered separately (Groves and Couper 1998; Lynn and Clarke 2002; Groves and Peytcheva 2008). A procedural perspective accounts for an interlinking of these dimensions. Accessibility relates to the sociology of everyday life, providing the

context in which negotiation makes sense. The context of the encounter may be more or less public, more or less sensitive to privacy, or more or less comfortable, and it assigns specific roles to the protagonists: intruder, pedestrian, citizen, colleague, and so forth. This perspective moves the analysis from a psychology of decision to a sociology of involvement. Rather than a disembodied preferences analysis or an individual influences model, the sequential analysis explores the sociology of how two actors negotiate a situation to which they attribute a meaning based on social institution, social times, and social roles.

In summary, participation in a survey is the outcome of a selection occurring through a sequence of contact attempts and cooperation requests.² The sequence of contact attempts is a process that results in the involvement of the respondent in a specific social activity.

Our study aims to associate contact sequences with sociological profiles. The hypothesis is that participation in the survey combines accessibility and willingness to cooperate. Our task is to identify the sociological links between characteristics of target persons and the way in which they are contacted. The methodological specifications and strategies of the interviewer meet the characteristics of the respondent to determine the typical sequence of contact. The sequence is a representation of the process of transforming a target person into a surveyed person.

Data and Final Outcomes of the Contact Attempts

The analyses involve contact data from the 2010 European Social Survey (ESS) and the 2011 Measures and Sociological Observation of Attitudes in Switzerland (MOSAiCH).³ Contact procedures are very similar in these surveys and use a sample frame of individuals devised by the Swiss Federal Statistical Office.

Before entering a sequential perspective, let us examine the survey from a final-outcome point of view. Seven categories are distinguished by the American Association for Public Opinion Research (AAPOR).⁴ The aim of the whole procedure is to obtain an interview. The average response rate is almost 52 %, which means the survey fails to include nearly half of the sample population. The category of non-contact (6 %) refers to no successful contact being made at the address. Broken appointment (3 %) refers to the contact failing despite an appointment having been arranged. The general category of inability (9 %) refers mainly to health problems, absence during the fieldwork, or inability to speak a survey language (French,

² The sequential approach stresses each element of fieldwork, including the contact letter, contact, conversions, etc. In this way, the Total Survey Error paradigm acknowledges this procedural perspective by linking all the elements of the survey in order to consider survey quality.

³ ESS (European Social Survey) and MOSAiCH (Measures and Sociological Observation of Attitudes in Switzerland) are two general social surveys conducted face-to-face by FORS. For further information please see: <http://www2.unil.ch/fors/>

⁴ See AAPOR, Standard Definitions, Revised 2011, at <http://www.aapor.org/Home.htm>

German, or Italian). Ineligibility (3 %) refers to a non-existent address, persons who are deceased, or people who have moved into institutions (retirement home, jail) or abroad. Two types of refusal are distinguished according to the persons concerned. The most important cause of non-participation is refusal by the respondent (25 %). Proxy refusal (2 %) refers to refusal by a household member.

The contact procedure involved several steps. After receiving a list of names, interviewers attempted to make face-to-face contact with each target person. They were required to make at least four attempts to contact the target person and accepted up to one refusal. If they failed to obtain an interview, the first phase ended, and the name was reallocated among interviewers to achieve a refusal conversion. Subsequent to another refusal or several non-contact cases (telephone calls were permitted after face-to-face contact had been established) the fieldwork was terminated for that person.

Both the ESS 2010 and the MOSAiCH 2011 were experimentally extended through a non-response survey. A two-page questionnaire was sent to non-contact names and to the people who refused to partake in the survey. In all, 5,259 names were contacted or their address visited. The outcomes after the second phase of contact indicate a gross response rate of 51.7 %. Including the figures from the non-response survey, the total response rate was 72.0 %.

Key Variables

In order to test the relationship between the sequences of contact attempts and characteristics of the sample unit, three sets of variables were defined. They make it possible to measure the composition of the sample. In the last part, they will be used to test whether respondents differ from non-respondents

1. The first set consisted of the final outcomes (see above).
2. The second set included exhaustive information about respondents, making it possible to analyse the characteristics of the entire sample. In addition of the information from the contact procedure, two other sources were identified:
 - a) the register of the Swiss Federal Statistical Office, from which the sample is drawn (nationality, age and civil status, as well as geographic location); and
 - b) the observations made by interviewers during visits. Regarding nationality, three types of citizenship are distinguished: Swiss citizenship; people with citizenship from one of Switzerland's neighbouring countries (Italy, France, Austria, or Germany); and people who are citizens of more distant countries. The scantiness of information provided by the register data allowed for the codification of a rough proxy of family types. Six modalities were differentiated using the age and civil status of the respondent. The age categories were chosen according to life-course considerations. In Switzerland, retirement actually occurs at 65 and 64 years (male/female). Being less than 22 years of age corresponds to a higher probability of living in the parental home. Being 38 years old corresponds to the midpoint of a change in lifestyle, with the

probability of having children at school age and engaging in activities that are no longer those of the youngest age group. As the research is interested in lifestyle and household composition, divorced individuals were grouped together with singles. The 6 categories are: under 22, single from 22 to 38, married from 22 to 38, single from 39 to 64, married from 39 to 64, and both single and married over 64.

The address offers information about the geographic region of the respondents. As per the national statistical office typology, we adopt the classification of urbanity: city centre, agglomeration, and rural.

Furthermore, each address visited, whether contact was successful or not, was also evaluated by the interviewer. The assessment of the address and its environment provides systematic information about the living conditions of the respondent. We selected the evaluation of the economic status (5-point scale from very rich to very modest).

3. The non-response survey provides information about some non-respondents and, therefore, increased the total number of persons on which we have detailed observations. The main benefit of this survey is its ability to probe attitudes. The difficulty lies in providing stable indicators, even more so in a context of changing mode and context of interview. These difficulties can be overcome with control groups. In fact, the participants in non-response survey are nevertheless respondents, even though they have refused to participate in the main survey; we don't have information about the remaining non-respondents and cannot answer the question as whether or not this core of non-respondents is qualitatively different from those who finally participated to the non-response survey.

In terms of content, one question focuses on the respondents' view concerning the utility of the survey (6-point scale agreement); this variable is the one that is the most highly correlated with participation in the survey. The assumption is that some participation factors are linked to the social utility of surveys. The main activity (employed, unemployed, training, retired, at home) is also an interesting variable to test lifestyle hypotheses. A dummy variable that indicates a xenophobic attitude is created through different questions in ESS, such as 'immigrants make Switzerland a worse place to live', and in MOSAiCH, such as 'give more chances to Swiss than to foreigners'.⁵ The openness to otherness seems to apply to both foreigners and various requests, such as survey participation. Another question focuses on the satisfaction with the manner in which democracy works (4-point scale satisfaction). The correlation of the attitude about democracy with survey participation indicates proximity between the democratic process and the process of the survey.

⁵ The comparison between the results of ESS and MOSAiCH confirms that the stability of this index is sufficient to constitute an indicator.

Table 15.1 Categories of contact attempts, ESS 2010 and MOSAiCH 2011, Face-to-face survey, n=5107

	Participant	Abbreviation	Frequency	Percent
Non-contact		NC	6796	38.4
Appointment made with...	Household	AP	1174	6.6
Refusal by...		RP	500	2.8
Appointment made with...		AR	3064	17.3
Refusal by...	Respondent	RR	2702	15.3
Interview of...		IN	2727	15.4
Respondent unavailable		UN	750	4.2
TOTAL			18245	100.0

Handling the Sequences of Contact

The analysis of the sequences of contact attempts elaborated by Kreuter and Kohler (2009) shows that indicators that use sequential information give better predictions of participation than do non-sequential indicators. Yet, even though the sequential indicators are promising in terms of response prediction, Kreuter and Kohler (2009) have not found any sufficient correlations with key variables to adjust for non-response. Their approach differs methodologically from ours in three ways concerning the coding used to handle sequences: Kreuter and Kohler (2009) built sequences composed of four events, namely contact with respondent, contact with someone else, non-contact, and interview; they deleted the last event of each sequence, arguing that this information is redundant with regards to the length of the sequence; and they also remove sequences of one episode, arguing that they include no sequential information. Our approach offers a much richer perspective with regards to different events.

To deal with the complexity and huge variance regarding the sequences of contact attempts, codifications were made to the raw data in order to obtain a balance between fruitfulness and homogeneity of information. Codification of the event is presented in Table 15.1. It does not differ much from the final-outcomes codification. This codification indicates whether the protagonist of the interaction is the selected respondent himself/herself or a member of the household (proxy). Finally, it differentiates the outcome of the contact attempt: interview, refusal, appointment, or other.

Table 15.2 Sequence of contact attempts, ESS 2010 and MOSAiCH 2011

Percent	ESS and MOSAiCH	Frequency
1	AR-IN	565
2	NC-AR-IN	330
3	NC-NC	305
4	IN	287
5	NC-NC-AR-IN	281
6	NC-IN	165
7	UN	151
8	NC-NC-IN	102
9	AR-AR-IN	100
10	NC-NC-RR	82
	Subtotal	2368

	Total	5107
		100.0

Duplicates due to error or the double entry of an interviewer were removed, as well as new contact attempts made in a short time span. In all cases, this lengthens sequences without adding information. We chose the limit of one-day duration because it exceeds the probability of double inscription. Deleting repeated contacts reduced the maximum length from 144 to 44 contact attempts. The second step of reduction is called ‘standardisation of sequences’. Identical consecutive codes were reduced: one consecutive identical code stays as it is, two or three were coded as two elements, and more than three were coded as three elements. Finally, occurrence of any element in a sequence was limited to the maximum number of 4. The result was a set of sequences that is far more homogeneous, with a maximum length of 15 elements. This standardisation keeps the analysis from being saturated by length differences and also facilitates discriminating sequences with respect to their content.

Patterns of Contact Attempts

With this codification, we obtained a total of 5,107 sequences, of which 1,125 are distinct. Despite the reduction of long patterns, the sequences present a huge dispersion. Thirteen different sequences (1%) contribute to 50% of the total; 837 (74.4%) exist in one specimen and contribute to 16.4% of the total. Table 15.2 shows a preview of the 10 most frequent sequences of contact attempts.

The most frequent sequence of contact consists of an appointment made by the selected respondent and an interviewer. Sometimes a non-contact may occur prior to the appointment, or the interview may take place at the first contact. We also noticed situations of repeated non-contacts before a procedure is completed. Sometimes a series of non-contacts finally leads to a contact, which then leads directly to an interview or an appointment. Sometimes non-contact is followed by a refusal. Another typical situation is a series of appointments that barely distinguish between a hidden refusal and difficulties of coordination.

We will analyse below the link between the characteristics of the respondent and the form of the sequence variables. A simple analysis shows that there are associations between them. The sequence in which a first contact attempt succeeds in leading to an appointment, followed by an interview, is over-represented by Swiss citizens from slightly higher socioeconomic groups, while the sequence composed of an appointment made by the proxy, followed by an interview, is most prevalent with young respondents supposedly living at home with their parents. Sequences characterised by non-contact cases that do not result in an interview mostly involve respondents who are fairly young, foreign, and married. These findings show that the forms the sequences take are sociologically significant. To develop a systematic analysis of these relationships, we must first create an object, dissimilarities between pairs of sequences, that provides a numerical model to account for the set of sequences.

Dissimilarities Between Pairs of Sequences

In order to obtain the relevant depiction regarding the set of sequences, we measured the similarity between sequences with the package TraMineR (R library; see Gabadinho et al. 2011). The outcome of the measurement of similarity is a pairwise distance matrix between 5,107 sequences. We tested three methods described by Elzinga (2007): the longest common prefix method (LCP), longest common suffix method (RLCP), and longest common subsequence (LCS) method, as well as optimal matching analysis (OMA) with various parameters (Abbott and Forrest 1986; Abbot and Tsay 2000).

According to Gabadinho et al. (2011), dissimilarity measures can distinguish between measures based on the count of matching attributes and those defined by the cost of transforming one sequence into another. Another interesting distinction is between those that allow and those that do not allow the shifting of matching states inside a sequence, which is called aligning. The simplest distance measurements count the number of successive common positions, starting from the beginning of the sequences (LCP). However, any early deviation implies a distinction. The sequences of contact being a mutual adjustment process with an aim, we tested the inverse measure (RLCP), which looks for the common elements from the end of the sequence. This measure effectively corresponds to the notion that the sequences of contacts are processes directed towards finality, i.e., interview. However, the measure is based on time alignment as well. The LCS metric reduces distances by accounting for position-shifted similarities. Lastly, optimal matching (OM) can be defined as the minimal cost of transforming one sequence into another. Essentially, two types of operations are considered: the substitution of an element by another and the insertion or deletion ('indel') of an element.

Setting a high indel cost relative to substitution costs favours substitutions, while low values favour indel. Substitution costs constitute a matrix that was computed by means of a function implemented in TraMineR. Choosing the dissimilarity measure and setting substitution and indel costs are crucial; favouring insertions and dele-

Table 15.3 Selection of metrics with multi-factor discrepancy analysis

R2 total	RLCP	LCP	LCS	LCS	LCS	LCS	OM	OM	OM
Codification	Full	Full	Full	Full	Standard	Standard	Full	Full	Standard
Normalisation	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes
Indel costs	–	–	–	–	–	–	2	1	1
Cat. of contact attempts	0.323	0.120	0.550	0.509	0.547	0.537	0.346	0.413	0.445
Whole sample var.	0.011	0.012	0.016	0.018	0.016	0.019	0.014	0.016	0.017
Non-response sample var.	0.110	0.023	0.088	0.133	0.106	0.138	0.090	0.125	0.129

tions reduces the importance of time shifts in the comparison, while favouring substitutions gives greater importance to position-wise similarities. While the length difference between a pair of sequences produces, on average, a larger distance, it may become necessary to normalise the distances, as the results do not have to depend on the lengths. Consequently, Elzinga's normalization (Elzinga 2007) is applied.

The distance matrix obtained enables us to use classic data-analysis tools. It permits the running of ANOVA-like analyses (Studer et al. 2010; Gabadinho et al. 2011) to apply multidimensional scaling and geometrical approaches as well as extract representative sequences, such as 'medoids', of a set of sequences. The medoid is the most central object, i.e., the one with the smallest sum of distances in relation to all other objects within the set (Kaufman and Rousseeuw, 2005). To correct for the effect of the contact procedure in the sample quality, we built a weighting based on the ratio of distance to the medoids of the set of sequences leading to interviews. This indicator can be interpreted as scores of propensity to participate in the sequence perspective. The effect of such a weighting has been tested on the key variables and will be presented later.

The Quest for the Metric

A major challenge in sequence analysis is finding the correct model to account for the sequential process. Comparing the results obtained with various settings helps in selecting the most appropriate measure. This testing is presented in Table 15.3. It should be restated that the socio-demographic positions (sample variables) are probably related to accessibility and mainly linked to the length of the sequences, while the attitudinal variables (non-response variables) are most related to cooperation, which does not have much influence on the length of the sequences. Broadly, the multi-factor discrepancy analysis shows that changing the metric does not alter the focus on the phenomenon.

We tested different insertion/deletion (indel) costs. The substitution costs were computed with the 'TRATE' (transition rates between each state) function of TraMineR. Two event codifications were tested: the 'full' and the 'standardised' (deletion of repeated events). We also tested for normalisation distance. We observed that, apart from the LCP metric, the measures yielded similar results. If some differences can be observed according to the method used, the improvement of a

method through a higher R2 coefficient relative to another method applies generally to all variables. We have observed some exceptions. For example, RLCP and LCP are comparable for the whole sample variables but not for non-response variables. Another obvious exception is the methods that are very restrictive on the sequence length (cost higher than 2) or cannot describe the variance of the socio-demographic variable. Shorter (standardised) and normalised sequences focused on extracts (LCS) give better results than does a whole sequence with holistic methods (OMA, in particular with high indel costs). Even though balancing between methods, in particular indel cost and substitution costs in OMA, is balancing between length and content, the quality of the measurement of accessibility (sample's variables) and cooperation (respondents' variables) vary together: there is a broad range where the two dimensions are optimised.

We opted finally for LCS normalised with standardised codification—even if OMA provides good results. RLCP and LCP show the poorest results. This highlights that some typical sub-sequences are likely to be an important determinant in predicting the success of an interview.

Mapping Sequences of Contact Attempts

We applied a categorical principal components analysis (CatPCA) of the 5107 sequences. This analysis presents a systematic view of the relationship between sequences and key variables.⁶

To make a comparison between the analyses, the key variables were considered as supplementary variables and fitted into the solution obtained by the sequences and final result. The analysis produces joint category plots that are single plots of the centroid of each selected variable.

The structure of the sequences may be mapped into a two-dimensional space that allows plotting of the main key variables. We can notice (see Figs. 15.1 and 15.2) that this space is organised into the shape of a triangle. The three poles are defined by the interview, the non-contact, and the refusal. The first dimension, which is given by the axes between interview and refusal, reflects the propensity to cooperate, while the second dimension, between interview and non-contact, assumes the difficulty of reaching the respondent. We can see how the variables are allocated between these dimensions. Non-contact, broken appointment, proxy refusal, and unknown eligibility imply accessibility issues. Unknown eligibility often accounts for removal or disease causing contact difficulties. The proxy refusal occurs more likely when the target person is frequently absent.

The subsequent paragraphs discuss how the key variables are interconnected in relation to these dimensions.

⁶ In fact, from a technical point of view, we begin to decrease the dimensionality using a MDS (multidimensional scaling) solution before applying the CatPCA on the coordinates of the MDS solution.

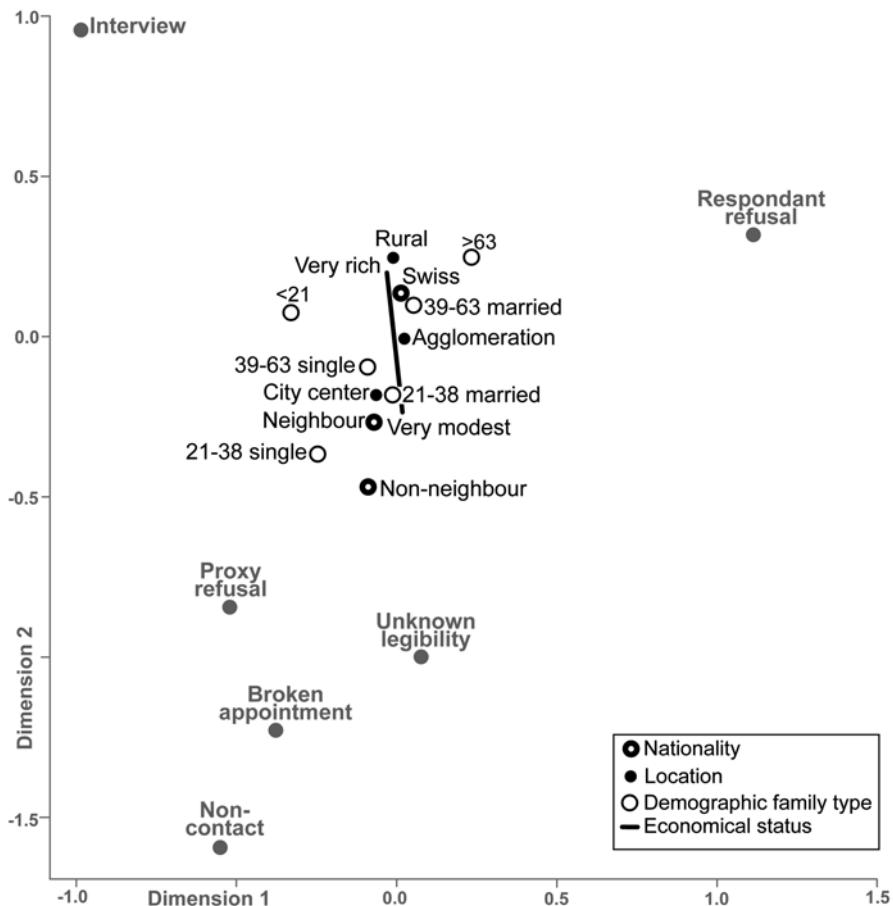


Fig. 15.1 Principal component analysis for whole sample variables

Figure 15.1 presents results for whole sample variables. The results show that most of the socio-demographic characteristics are situated along the axis of accessibility. Contacting appears to be related to the lifestyle of the target person. For example, we notice that young people under the age of 21 are relatively easy to contact. It is likely that they are in school; therefore, their schedules allow them to spend time at home. They rarely live alone, and other family members increased the probability of someone's presence at the residence. The rarity of proxy refusal means that proxies interfere less in the young persons' private lives than do those who are part of a couple, for example. Results highlight that reaching target persons between the ages of 21–38 is challenging, while reaching those between the ages of 39–63 is less challenging. The former more often practice a lifestyle characterised by activities outside the home, especially if they are single. The 39–63 group seems to be more stay-at-home and differs according to marital status, marriage

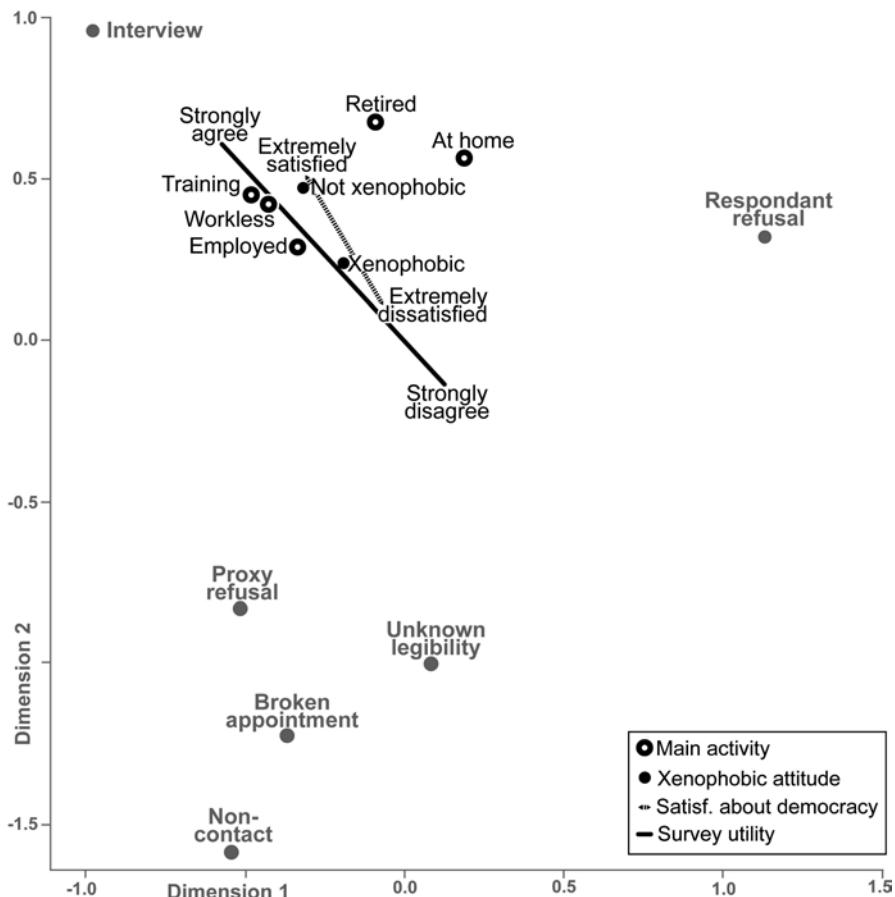


Fig. 15.2 Principal component analysis for respondents' variables

implying more reluctance to participate. In the case of most elderly people (>63 years old), accessibility increases further. As retired people are usually at home, they are likely to be frequently solicited by polling institutes. This population is also more vulnerable to scams, which explains its high level of reluctance to participate. Lower socioeconomic groups are easier to reach than are higher socioeconomic groups, which may be related to their more urban habitat. Marital status has an impact that differs according to age. Foreigners, especially from non-neighbouring countries, are more difficult to reach, which can be attributed to their mobility. Lifestyle, which is represented by socio-demographic variables, appears to determine the accessibility of the respondent. However, the analyses indicate interactions of the dimension of accessibility with the propensity to cooperate.

While variables related to the demographic dimension show which targets are more or less easy to reach, attitudinal variables present a distribution related to the reluctance to participate (Fig. 15.2). We can observe that the people who think

that surveys are useless are, for the most part, reluctant. Persons who are not satisfied with democracy are less likely to cooperate. We can see that both satisfaction with democracy and xenophobia illustrate the same pattern, though with a different intensity and a reversed order. One can assume that these two variables account for a general openness to social environment. Being satisfied by democracy is an opinion linked to the representation of different categories of people in the institutions. This attitude is very close to the opinion about the integration of foreigners. The dimension drawn by the variable on opinions about survey utility appears to be more oriented to the accessibility axis. This meets our assumption of a general availability of oneself that takes the form of both the propensity to accept (or not) an external request and the propensity to have a lifestyle that makes one accessible (or not) to others.

This map distinguishes mechanisms for participation by the way of an analysis of sequences of contact attempts: xenophobia and dissatisfaction with democracy requires the same type of contacting, while disbelief in the usefulness of surveys requires another type of contacting. Does control for type of contacts constitute a way to describe differential participation in the survey?

Coverage and Contact Attempts

The final part of this research attempts to outline an approach to survey coverage using the contact data. This analysis is another way to show the effects of the contact procedure on participation. A score was computed for each respondent based on the LCS distance from the medoid of all sequences of respondents, i.e., the centre of the group of respondents. The score expresses the distance to an ‘average sequence’ of respondents. This distance was used as a weight in order to control the sample for contact procedure. Tables 15.4 and 15.5 present the results for the whole sample variable and for the non-response variables, respectively. The columns present the percentages, respectively, of the categories for the respondent during the first phase of the survey (prior to refusal conversion and a telephone call), respondents during the main survey (after the second phase), enhanced sample by the non-response survey, and—when data are available—the whole sample informed by the register and interviewers’ information. Comparison between columns shows the difference in sample composition. The last column presents the results corrected with weighting. If, for socio-demographic variables, we have the ‘true’ values for variables of attitudes, we must use the estimation of the non-response survey as the most plausible indicator. However, this test is—in this case—probably conclusive, as all the corrections are moving in the direction of more plausible shares. Furthermore, it allows building an interesting theoretical object, which can be likened to a sample with ‘equalised contact procedure’.

With the whole sample variables, we can notice the effect of the correction on the socio-demographic variables. The direction of the correction is consistently good, and the values obtained in the sample of survey respondents are close to the

Table 15.4 Whole sample variables, distribution of key variables

Variables	Modalities	First phase N=2236	Respondents N=2715	Enhanced sample N=3749	Ref: total sample N=5107	Weighted respondents
Nationality	Swiss	84.2	83.6	83.0	78.5	76.0
	Neighbour	7.2	7.6	8.1	9.0	9.8
	Non-neigh- bour	8.7	8.8	8.9	12.5	14.2
	Total	100.0	100.0	100.0	100.0	100.0
Demographic family type	<22	8.2	7.9	7.1	6.4	6.4
	22–38 single	15.0	14.4	15.3	16.0	16.5
	22–38 married	10.0	9.9	10.5	11.1	11.8
	39–64 single	8.9	9.3	9.7	10.2	9.9
	39–64 married	37.4	37.3	36.5	34.8	33.4
	>64	20.5	21.2	21.0	21.4	22.1
	Total	100.0	100.0	100.0	100.0	100.0
Localisation	City centre	24.6	25.9	26.7	29.1	30.1
	Agglomera- tion	43.6	43.6	44.8	44.9	45.2
	Rural	31.8	30.5	28.5	26.0	24.7
	Total	100.0	100.0	100.0	100.0	100.0
Interviewer's assess- ment of economic status	Very rich	3.1	2.9	2.7	2.6	2.6
	Rich	23.1	23.8	22.7	20.9	19.8
	Neither r., nor m.	53.7	53.2	54.5	54.8	54.5
	Quite modest	18.1	18.1	18.1	19.4	20.6
	Very modest	2.0	2.0	2.0	2.4	2.6
	Total	100.0	100.0	100.0	100.0	100.0

values of the register for the complete population. The weighting leads the shares beyond the reference values for nationality and assessment for socioeconomic status. It is apparent that the improvement provided by the weighting affects socio-demographic variables related to accessibility. Table 15.5 confirms the capacity of this weighting based on distance to respondent to also correct for the reluctance of some respondents.

In matters of attitudes, the correction is pushing the share of respondents (half of the population) in a plausible direction, according to the estimations provided by the non-response survey (three quarters of the population). It even accentuates the share of the non-response survey in all variables. The exception affects the Main activity, which can be related to the mitigated results obtained with Demographic family types in the previous analysis. This result is difficult to interpret. We can notice that these two variables correspond less with the 'pure dimension' in terms of

Table 15.5 Respondents' variables, distribution of participation

Variables	Modalities	First phase N=2169	Respondents N=2634	Ref. enhanced sample N=3648	Weighted respondents N=2634
Utility of survey	Strongly agree	12.5	11.6	10.2	9.6
	Agree	64.0	63.3	56.5	53.9
	Neither agree nor disagree	17.8	19.4	23.8	25.6
	Disagree	4.7	4.7	7.3	8.2
	Strongly disagree	0.8	1.0	2.2	2.7
	Total	100	100.0	100.0	100.0
Main activity	Employed	59.6	59.6	60.2	59.7
	Unemployed	2.4	2.3	2.3	2.4
	Training	8.3	7.6	7.1	7.0
	Retired	21.5	22.3	22.0	22.1
	At home	8.2	8.1	8.4	8.7
	Total	100.0	100.0	100.0	100.0
Xenophobic attitude	Xenophobic	21.0	20.7	25.8	27.6
	Not xenophobic	79.0	79.3	74.2	72.4
	Total	100.0	100.0	100.0	100.0
Satisfaction about democracy	Extremely dissatisfied	2.2	2.2	2.8	3.0
	Dissatisfied	9.2	9.6	10.4	11.0
	Satisfied	67.1	67.4	68.0	68.1
	Extremely satisfied	21.5	20.8	18.8	17.9
	Total	100.0	100.0	100.0	100.0

our classification demographic (accessibility) versus attitude (cooperation). Overall, the analyses suggest, nevertheless, a stable contact procedure vis-à-vis the data. The weighting reflects the propensity to participate in the survey, from a global and exhaustive point of view, regarding the average respondent's sequence of contacts.

Towards a Perfect Sampling Process?

Conducted from a sequential perspective, this research offers knowledge regarding the theory of sequences in social activities. Through sociology of involvement in social activity related to survey methods, we suggested a forward-looking perspective in measuring attitudes.

Sequential perspective is rooted in empirical reality. It helps to describe a social process with its ambiguities—a sequence is a transformation. The numerous examples of contacts that first resulted in an initial statement of ineligibility but were restored during a subsequent visit reveal the dynamic nature of the outcomes

of social activities. This perspective shows that the sequences of contact attempts reflect ‘social relationships’ that can also exist for themselves, e.g., through a series of broken appointments or through the interviewer’s strategy of keeping in touch. With respect to this first glance at contacts, the processual nature of sequence of contact attempts may be challenged. Our hypothesis is that even short sequences of few elements are already processes, in the minimal meaning that they could have a greater length. From a technical point of view, short sequences contain little sequential information (transition, order). Therefore, from a sociological point of view, the common length of a social encounter involves few steps. In everyday life, several consecutive unsuccessful contacts appear to be unexpected and frustrating. A commonly expected sequence of social coordination set up is that, after a first contact is made, the cooperation is either rejected or accepted and then can take place either immediately or later through an appointment. The coordination will appear ‘abnormal’ to the actors if the sequence includes additional non-contact, broken appointment, or other vicissitudes (request repeated by interviewer after a refusal; refusal from the interviewee after an appointment). An assumption that can be made is that we are dealing with a change in the forms of social coordination (Thévenot 2006). A sequence of relatively short contact represents a familiar order of engagement. The interviewer uses local and ordinary knowledge about people (at a certain time someone eats, someone sleeps, etc.). As the sequence becomes longer and is influenced by professional principles as dictated by the survey institute, the ways of engaging in social activities changes, i.e., where an ordinary person would have given up, the interviewer must continue the contact process to obtain as many cooperative people as possible. This leads to a model of public justification that could be associated with the industrial order described by Boltanski and Thévenot (1991); the criteria are those of efficiency and objectivity. The main difficulty in the analysis of the sequences of social activities is that their logical principles of constitution vary according to their length.

From the survey-method point of view, this perspective offers a better understanding regarding the chances of obtaining cooperation. This perspective opens the black box of surveys, showing the process of ‘making’ respondents. Applications of these results could lie in the comparison of interviewers’ strategies as well as in international comparisons to show the effects of different ways to deal with survey fieldwork. An important result is that there is a systematic link between the form of the sequences and the final outcome. This means that we can both study and identify the mechanisms by which more efforts in contact process result, or not, in the same population, i.e., diversity in contact procedures broadens the surveyed population.

Nonetheless, it seems difficult to use only information based on these sequences in order to correct a sample, because the sequences of contact are related not only to the respondent but also to the interviewer. Survey is embedded in social interaction (contact processes) and language (interview). Seeking to subtract a survey from its grounds can drive to a deadlock. We prefer, in this sense, to develop the control of social effects and understand what surveying means.

References

- Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16(3), 471–494.
- Abbot, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: review and prospect. *Sociological Methods and Research*, 29(1), 3–33.
- Beatty, P., & Herrmann, D. (2002). To answer or not to answer: Decision processes related to survey item non-response. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. Little (Eds.), *Survey non-response* (pp. 71–85). New York: Wiley.
- Boltanski, L., & Thévenot, L. (1991). *De la justification: Les économies de la grandeur*. Paris: Gallimard.
- Bourdieu, P. (1980). *Le sens pratique*. Paris: Éditions de Minuit.
- Dillman, D. A. (1978). *Mail and telephone surveys: the total design method*. New York: Wiley.
- Dunne, M., Martin, N., Bailey, M., Heath, A., Bucholz, K., Madden, P., et al. (1997). Participation bias in a sexuality survey: psychological and behavioural characteristics of responders and non-responders. *International Journal of Epidemiology*, 26(4), 844–854.
- Elzinga, C. H. (2007). *Sequence analysis: Metric representations of categorical time series*. Amsterdam: Vrije Universiteit. <http://home.fsw.vu.nl/ch.elzinga/MetricsRevision.pdf>. Accessed 16 Sep 2013.
- Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Groves, R. M. (2006). Non-response rates and non-response bias in household surveys. *Public opinion quarterly*, 70(5), 646–675.
- Groves, R. M., & Couper, M. (1998). *Non-response in household interview surveys*. New York: Wiley.
- Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly*, 74(5), 849–879.
- Groves, R. M., & Peycheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly*, 72(2), 167–189.
- Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: description and an illustration. *Public Opinion Quarterly*, 64(3), 299–308.
- Groves, R. M., Presser, S., Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68(1), 2–31.
- Joule, R.-V., & Beauvois, J.-L. (2002). *Petit traité de manipulation à l'usage des honnêtes gens*. Grenoble: Presses Universitaires de Grenoble.
- Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data. An introduction to cluster analysis*. New-York: Wiley-Interscience.
- Keeter, S., Miller, C., Kohut, A., Groves, R., & Presser, S. (2000). Consequences of reducing non-response in a national telephone survey. *Public Opinion Quarterly*, 64(2), 125–148.
- Kreuter, F., & Kohler U. (2009). Analysing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *Journal of Official Statistics*, 25(2), 203–226.
- Lynn, P., & Clarke, P. (2002). Separating refusal bias and non-contact bias: evidence from UK national surveys. *The Statistician*, 51(3), 319–333.
- Matsuo, H., Loosveldt G., & Billiet, J. (2006). The history of the contact procedure and survey cooperation. Louvain-la-Neuve: Quetelet conference. http://www.uclouvain.be/cps/ucl/doc/demo/documents/Matsuo_Loosveldt_Billiet.pdf. Accessed 29 April 2013.
- Nolen, J.A., & Maynard, D.W. (2013). Formulating the request for survey participation in relation to the interactional environment. *Discourse Studies*, 15(2), 205–227.
- Peytchev, A., & Olson, K. (2007). *Using interviewer observations to improve nonresponse adjustments: NES 2004*. Papers presented at the Joint Statistical Meetings, Salt Lake City. <http://www.amstat.org/sections/srms/proceedings/y2007/Files/JSM2007-000695.pdf>. Accessed 16 Sep 2013.

- Pollien A., & Joye, D. (2011). A la poursuite du répondant? Essai de typologie des séquences de contact dans les enquêtes. In D. Joye, C. Pirinoli, D. Spini, & E. Widmer (Eds.), *Parcours de vie et insertions sociales* (pp. 189–212). Zürich: Seismo.
- Shuttles, C.D. (2008). Refusal avoidance training (RAT). In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 702–703). Beverly Hills: Sage.
- Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response. Lessons learned from the European Social Survey*. Chichester: Wiley.
- Studer, M., Ritschard, G., Gabadinho, A., & Müller, N.S. (2010). Discrepancy analysis of complex objects using dissimilarities. In F. Guillet, G. Ritschard, D.A. Zighed, & H. Briand (Eds.), *Advances in Knowledge Discovery and Management* (pp. 3–19). Berlin: Springer.
- Thévenot, L. (2006). *L'action au pluriel: sociologie des régimes d'engagement*. Paris: La Découverte.