Life Course Research and Social Policies 10

Gilbert Ritschard · Matthias Studer
Editors

Sequence Analysis and Related Approaches

Innovative Methods and Applications

Springer Open

Life Course Research and Social Policies

Volume 10

Series editors
Laura Bernardi
Dario Spini
Jean-Michel Bonvin

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.

More information about this series at http://www.springer.com/series/10158

Gilbert Ritschard • Matthias Studer
Editors

Sequence Analysis and Related Approaches

Innovative Methods and Applications

Springer Open

Editors
Gilbert Ritschard
NCCR LIVES and Geneva School
of Social Sciences
University of Geneva
Geneva, Switzerland

Matthias Studer
NCCR LIVES and Geneva School
of Social Sciences
University of Geneva
Geneva, Switzerland

ISSN 2211-7776 ISSN 2211-7784 (electronic)
Life Course Research and Social Policies
ISBN 978-3-319-95419-6 ISBN 978-3-319-95420-2 (eBook)
https://doi.org/10.1007/978-3-319-95420-2

Library of Congress Control Number: 2018957117

© The Editor(s) (if applicable) and The Author(s) 2018, corrected publication 2018. This book is an open access publication
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
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.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume provides innovative methods and applications of sequence analysis (SA) to life course and time use data with a focus on the relationship between SA and other methods for longitudinal data. It originated in the International Conference on Sequence Analysis and Related Methods (LaCOSA II) held in Lausanne, 8–10 June, 2016. The 15 articles that, together with the introduction chapter, constitute the book were selected among 25 propositions received in response to a call for papers addressed to the participants and members of the scientific committee of the Conference. The selection was conducted through peer reviewing with each paper being evaluated by at least three reviewers.

How to Read the Book

The chapters are self-contained and each one can be picked-up independently of the others. They are not ordered by difficulty level but have been organized so as to optimize the topical and methodological consistency in the succession of the chapters. The retained storyline is based on the kind of SA involved and on how SA relates to other methods for longitudinal analysis. There are six parts each with two or three chapters.

In order to guide the reader in selecting the chapters, Table 1 below characterizes their nature by means of a series of attributes. The column "SA type" gives an idea of the kind of knowledge involved.

In addition, the reader will find useful information for her/his choices in the introductory chapter that summarizes the development of SA to date, enshrines the editors' view on the future of SA, explains the storyline of the book, and gives a short overview of what you find in each chapter.

The chapters in Part I by Courgeau and by Eerola that overview different approaches for longitudinal analysis and their relationship with SA are the most general. They probably also are the most difficult papers because of the number of

Table 1 Characteristics of the chapters

Chapter	SA type[a]	Classic SA	Primary applied	Innovative method	Original method combination	Mathematical level
Ritschard, Studer	overview					
Courgeau	overview					
Eerola	*D*, other				x	++
Malin, Wise	*G*, *S*		x			
Lundevaller et al.	*D*, *S*	x	x			
Rossignon et al.	*D*, *S*	x	x	x	x	
Cornwell	*N*, *I*, *G*					+
Hamberger	*N*, *G*		x	x		+
Collas	*D*, *G*	x		x		+
Borgna, Struffolino	*D*, other		x		x	
Helske et al.	*M*, *D*, *G*			x	x	+
Taushanov, Berchtold	*M*		x	x		+
Studer	*D*, *G*	x		x		+
Bison, Scalcon	*D*, *I*	x		x		++
Manzoni, Mooi-Reci	*I*		x	x		+
Ritschard et al.	*I*			x		+

[a]Type of SA: *D* dissimilarity-based, *N* sequence network, *M* Markov-based, *I* individual sequence summary numbers, *S* survival approach, *G* graphical sequence representation

concepts and approaches they refer to. Moreover, the chapter by Eerola is one of most demanding mathematically. Nevertheless, these two chapters are important by demonstrating the range of possibilities and advantages of using SA in conjunction with related approaches. Courgeau's text highlights the added value of SA with regard to other methods used in demography to study the dynamics of trajectories. Eerola's contribution addresses the combination of SA with probabilistic models rarely considered in social sciences, which opens new perspectives for the analysis of life processes.

The first two chapters in Part II, i.e., those by Malin and Wise and by Lundevaller and colleagues, are easily accessible even for newcomers to SA. These chapters are pure applications in which SA is used in combination with event history analysis (EHA) and require just basic knowledge in EHA (i.e., survival analysis). Chapters that use or deal with classic SA, i.e., with the clustering of sequences from pairwise dissimilarities, should also be easily comprehensible especially by readers with basic experience in SA. The other chapters either deal with non-clustering aspects of SA or consider alternative approaches such as sequence networks or Markov-based models. They may require basic knowledge in the concerned fields to grasp all the details.

Acknowledgments

As editors, we would like to thank the authors of the chapters for their insights and contributions to the book. We also warmly acknowledge the members of the review committee and the associated referees—listed on next page—for their involvement in the review process of the chapters. Their in-depth reviewing, criticisms, and constructive remarks significantly contributed to the high quality of the retained papers. Thanks to an anonymous advisor for her/his valuable comments on the overall book.

Many thanks also to Springer and the editors of the Series *Life Course Research and Social Policies* for their confidence in our project.

This book 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 (Grant number: 51NF40-160590). The editors are grateful to the Swiss National Science Foundation for its financial assistance.

Geneva, Switzerland
November 2017

Gilbert Ritschard
Matthias Studer

Review Committee

- Silke Aisenbrey, Yeshiva University, USA
- Jake Anders, UCL Institute of Education, UK
- André Berchtold, University of Lausanne, Switzerland
- Torsten Biemann, University of Mannheim, Germany
- Ivano Bison, University of Trento, Italy
- Philippe Blanchard, The University of Warwick, UK
- Benjamin Cornwell, Cornell University, USA
- Daniel Courgeau, INED, France
- Cees Elzinga, Vrije Universiteit Amsterdam, Netherlands
- Anette Fasang, Humboldt University of Berlin, Germany
- Jacques-Antoine Gauthier, University of Lausanne, Switzerland
- Brendan Halpin, University of Limerick, Ireland
- Satu Helske, University of Oxford, UK
- Jean-Marie Le Goff, University of Lausanne, Switzerland
- Eva Lelièvre, INED, France
- Tim Liao, University of Illinois at Urbana-Champaign, USA
- Léonard Moulin, INED, France
- Madalina Olteanu, SAMM, Université Paris 1, France
- Raffaella Piccarreta, Bocconi University, Italy
- Gary Pollock, Manchester Metropolitan University, UK
- Nicolas Robette, UVSQ, France
- Emanuela Struffolino, WZB Berlin, Germany

Associated Reviewers

- Klaus Hamberger
- Mervi Eerola
- Jouni Helske
- Karina Videgain
- Lydia Malin
- Marta Mier-Y-Teran
- Giampiero D'Alessandro
- Lotta Vikström
- Irma Mooi-Reci
- Erling Lundevaller
- Dan Orsholits
- Olga Ganjour
- Camilla Borgna
- Luize Ratniece
- Matteo Antonini
- Thomas Collas
- Florence Rossignon
- Daniel Lapresa

Contents

Sequence Analysis: Where Are We, Where Are We Going? 1
Gilbert Ritschard and Matthias Studer

Part I About Different Longitudinal Approaches in Longitudinal Analysis

Do Different Approaches in Population Science Lead to Divergent or Convergent Models? 15
Daniel Courgeau

Case Studies of Combining Sequence Analysis and Modelling 35
Mervi Eerola

Part II Sequence Analysis and Event History Analysis

Glass Ceilings, Glass Escalators and Revolving Doors 49
Lydia Malin and Ramsey Wise

Modelling Mortality Using Life Trajectories of Disabled and Non-Disabled Individuals in Nineteenth-Century Sweden 69
Erling Häggström Lundevaller, Lotta Vikström, and Helena Haage

Sequence History Analysis (SHA): Estimating the Effect of Past Trajectories on an Upcoming Event 83
Florence Rossignon, Matthias Studer, Jacques-Antoine Gauthier, and Jean-Marie Le Goff

Part III The Sequence Network Approach

Network Analysis of Sequence Structures 103
Benjamin Cornwell

Relational Sequence Networks as a Tool for Studying Gendered Mobility Patterns 121
Klaus Hamberger

Part IV Unfolding the Process

Multiphase Sequence Analysis 149
Thomas Collas

Unpacking Configurational Dynamics: Sequence Analysis and Qualitative Comparative Analysis as a Mixed-Method Design 167
Camilla Borgna and Emanuela Struffolino

Combining Sequence Analysis and Hidden Markov Models in the Analysis of Complex Life Sequence Data 185
Satu Helske, Jouni Helske, and Mervi Eerola

Part V Advances in Sequence Clustering

Markovian-Based Clustering of Internet Addiction Trajectories 203
Zhivko Taushanov and André Berchtold

Divisive Property-Based and Fuzzy Clustering for Sequence Analysis 223
Matthias Studer

From 07.00 to 22.00: A Dual-Earner Couple's Typical Day in Italy 241
Ivano Bison and Alessandro Scalcon

Part VI Appraising Sequence Quality

Measuring Sequence Quality 261
Anna Manzoni and Irma Mooi-Reci

An Index of Precarity for Measuring Early Employment Insecurity 279
Gilbert Ritschard, Margherita Bussi, and Jacqueline O'Reilly

Correction to: Unpacking Configurational Dynamics: Sequence Analysis and Qualitative Comparative Analysis as a Mixed-Method Design E1

Index 297

Contributors

André Berchtold Institute of Social Sciences and NCCR LIVES, University of Lausanne, Lausanne, Switzerland

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

Camilla Borgna Collegio Carlo Alberto, Turin, Italy

Margherita Bussi University of Louvain, Louvain-la-Neuve, Belgium

Thomas Collas F.R.S.-FNRS – Université de Louvain, Louvain-la-Neuve, Belgium

Benjamin Cornwell Department of Sociology, Cornell University, Ithaca, NY, USA

Daniel Courgeau Institut National d'Etudes Démographiques (INED), Paris, France

Mervi Eerola Centre of Statistics, University of Turku, Turku, Finland

Jacques-Antoine Gauthier NCCR LIVES and University of Lausanne, Lausanne, Switzerland

Helena Haage Umeå University, Umeå, Sweden

Klaus Hamberger Laboratoire d'Anthropologie Sociale, Ecole de Hautes Etudes en Sciences Sociales, Paris, France

Satu Helske Institute for Analytical Sociology, Linköping University, Linköping, Sweden

Department of Sociology, University of Oxford, Oxford, UK

Department of Mathematics and Statistics, University of Jyvaskyla, Jyvaskyla, Finland

Jouni Helske Department of Science and Technology, Linköping University, Linköping, Sweden

Department of Mathematics and Statistics, University of Jyvaskyla, Jyvaskyla, Finland

Jean-Marie Le Goff NCCR LIVES and University of Lausanne, Lausanne, Switzerland

Erling Häggström Lundevaller Umeå University, Umeå, Sweden

Lydia Malin University of Cologne, Cologne, Germany

Anna Manzoni North Carolina State University, Raleigh, NC, USA

Irma Mooi-Reci University of Melbourne, Parkville, Australia

Jacqueline O'Reilly University of Sussex, Brighton, UK

Gilbert Ritschard NCCR LIVES and Geneva School of Social Sciences, University of Geneva, Geneva, Switzerland

Florence Rossignon Swiss Federal Statistical Office, Neuchâtel, Switzerland

Alessandro Scalcon Department of Sociology and Social Research, University of Trento, Trento, Italy

Emanuela Struffolino WZB Berlin Social Science Center – Research Group 'Demography and Inequality' and Humboldt University of Berlin, Berlin, Germany

Matthias Studer NCCR LIVES and Geneva School of Social Sciences, University of Geneva, Geneva, Switzerland

Zhivko Taushanov Institute of Social Sciences and NCCR LIVES, University of Lausanne, Lausanne, Switzerland

Lotta Vikström Umeå University, Umeå, Sweden

Ramsey Wise University of Bremen, Bremen, Germany

The original version of this frontmatter was revised.
A correction to this chapter is available at https://doi.org/10.1007/978-3-319-95420-2_17

Sequence Analysis: Where Are We, Where Are We Going?

Gilbert Ritschard and Matthias Studer

1 Sequence Analysis: Optimal Matching and Much More

Sequence Analysis (SA) has gained increasing importance in the field of social sciences since the pioneering contributions of Andrew Abbott (e.g., Abbott 1983; Abbott and Forrest 1986) and has become a popular tool with the release of powerful dedicated software (Brzinsky-Fay et al. 2006; Gabadinho et al. 2011; Halpin 2014). The increasing availability of longitudinal data sources such as panel and retrospective surveys also contributed to the rise of interest in SA. In recent times, SA has garnered a central role in life course studies to appraise, for example, occupational careers, cohabitation pathways, or health trajectories. In addition, it is effectively used in domains such as time use, spatio-temporal development of economic activities, and historical evolution of political institutions. In fact, sequences are a convenient way of coding individual narratives into a form suitable for quantitative analysis.

Briefly, SA primarily provides a comprehensible overall picture of sets of individual categorical sequences—the retained coding of the narratives—and involves using this overall picture for objectives such as discovering the characteristics of a set of sequences, identifying possible atypical or deviant individual trajectories, and comparing trajectory patterns among groups such as sexes, birth cohorts, or regions.

Abbott and Tsay (2000) describe the typical SA as a three-step program: (1) code the narratives as sequences, (2) compute pairwise dissimilarities between sequences, and (3) analyse the sequences based on their dissimilarities. The coding stage involves the selection of a suitable alphabet for the states or events to be considered, and the definition of the timing scheme to be used to time align the sequences.

G. Ritschard (✉) · M. Studer
NCCR LIVES and Geneva School of Social Sciences, University of Geneva, Geneva, Switzerland
e-mail: gilbert.ritschard@unige.ch; matthias.studer@unige.ch

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_1

The computation of the dissimilarities implies the choice of a suitable dissimilarity measure. The analysis itself typically involves building a typology of the trajectories by applying a clustering algorithm that takes the dissimilarities as input, even though Abbott and Tsay (2000) also consider multidimensional scaling.

The optimal matching (OM) distance borrowed from signal processing (Levenshtein 1966; Hamming 1950) and biology (e.g., Needleman and Wunsch 1970; Sankoff and Kruskal 1983) and popularized in the social science by Abbott and Forrest (1986) was typically used to compare the sequences. OM was so intimately connected with SA that the expression 'optimal matching analysis' was—and still is—often used as a synonym for SA, and this even when no or non-optimal matching distances are used.

In fact, during the second wave of SA development (see Aisenbrey and Fasang 2010) most methodological developments focused on the measurement of the dissimilarities between sequences and more specifically on OM. Several new variants for defining the OM-costs, but also new non-OM-based distance measures were proposed to address criticisms raised in the literature. This scattered development was recently summarized by Studer and Ritschard (2016) in their comparative review of dissimilarity measures.

More recently, the use of SA has seen developments in several other directions. The visualisation of sequences started veritably with the release of software dedicated to SA (Brzinsky-Fay et al. 2006; Gabadinho et al. 2011). The multiple possibilities to render a set of sequences with easily interpretable colourful plots undoubtedly boosted interest in SA. The most popular of these plots are index plots (Scherer 2001) that render the individual sequences and their diversity, and chronograms, which display the evolution of the cross-sectional distribution at successive time points. The cluttered aspect of the former and the over simplification of the latter—which completely overshadows the diversity of the sequences—naturally called for lightened forms of index plots. The search and plot of representative sequences by Gabadinho and Ritschard (2013) and the relative frequency sequence plot of Fasang and Liao (2014) offer solutions to this challenge. The decorated parallel coordinate plot (Bürgin and Ritschard 2014) focuses on sequencing within trajectories and is useful for identifying the most typical sequencing patterns while rendering the diversity of the entire set of observed trajectories at the same time.

Multichannel sequences, i.e., the joint analysis of narratives of different domains such as linked lives, or occupational and cohabitation trajectories, also received increasing attention. Here, one difficulty is the exploding size of the alphabet that results from the combination of different dimensions. For the specific case of OM, some authors (e.g., Pollock 2007) proposed tricks for defining the substitution costs between different combinations of states from costs set for each individual dimension. This dramatically reduces the number of parameters that need to be specified. The measurement of the strength of association between channels is probably more interesting and promising to study the relationship or interaction between dimensions such as familial and professional pathways. The contributions of Piccarreta and Elzinga (2013) and Piccarreta (2017) are path-breaking in that respect. The graphical rendering of multichannel sequences also requires special attention and effective solutions are, for example, provided by Helske and Helske (2017) and their `seqHMM` package.

There has been some work on exploiting the pairwise dissimilarities between sequences differently from clustering and multidimensional scaling. Studer et al. (2011) showed how to conduct an ANOVA-like analysis and to grow regression trees for sequence data. Gabadinho and Ritschard (2013) used the dissimilarities to find representative sequences such as the most central sequences or sequences with the densest neighbourhood.

With regards to clustering, there are attempts to get rid of explicit dissimilarity measures by resorting to latent class approaches (Vermunt et al. 2008; Barban and Billari 2012), or more or less similarly hidden Markov models (e.g., Helske and Helske 2017; Bolano et al. 2016), to clustering the sequences. Markov-based approaches may also contribute to understanding the dynamics that drive the unfolding of the sequences. However, the difficulty to synthesize the outcome of Markov-based transition models—especially when more realistic models with order greater than one are considered—negatively affected the extension of their usage. The graphical rendering of hidden Markov models (HMM) by Helske and Helske (2017) (see also Helske et al. 2018, in this bundle) as well as the use and rendering of probabilistic suffix trees (Gabadinho and Ritschard 2016) for sequence analysis should facilitate the access to such probabilistic approaches and shed light on how the current situation is linked to the history of previous situations.

More or less independently of classical SA, graph and network approaches (e.g., Butts and Pixley 2004; Bison 2014; Cornwell and Watkins 2015) have proven to provide useful summaries of individual sequences as well as synthetic views at the population level. See also Cornwell (2018) and Hamberger (2018) in this bundle.

Alongside the discovery of characteristics of the set of sequences, such as the diversity among sequences and typologies of trajectories, there are works concerned with summary numbers of individual sequences. More specifically, here, the aim was to complement simple indexes such as the sequence length, the number of different states visited, and the number of transitions, with measures of internal diversity and complexity of an individual sequence. Contributions in that direction have been made, for example, by Brzinsky-Fay (2007), Elzinga and Liefbroer (2007), Elzinga (2010), and Gabadinho et al. (2010).

2 Towards Stronger Interaction with Related Approaches

The short survey included above covering what has been done in SA is certainly incomplete. In particular, we did not mention two important issues that already received some attention in the literature: the handling of censored sequences and more generally of missing values, and the possibility to study the relationship of sequences with time-varying covariates. These issues have no definitive solution thus far and deserve further research. While different schemes for imputing missing values have been proposed (e.g., Halpin 2015; Gabadinho and Ritschard 2016), there remains the question of the maximal proportion of missing values that are appropriate to impute in a sequence. Moreover, the real impact of such imputations

on the SA outcome remains to be investigated. Regarding time-varying covariates, Studer et al. (2018a,b) proposed procedures combining SA with event history analysis in order to study the influence of such time-dependent covariates on a trajectory in a semi-holistic perspective. A possible completely holistic solution lies in the multichannel approach and the joint analysis (Piccarreta 2017) of the dependent channel with those defined by the history of values of each time-varying covariate. Here again, further investigation seems necessary. Therefore, there is still room for further development in classical SA. However, we think that the future of SA is intimately linked with the development of its interaction with other—more inferential and/or probabilistic—methods for longitudinal data.

Despite few attempts to introduce inferential methods in SA with ANOVA-like analysis and Markov-based modelling of sequences, SA essentially remains exploratory and needs to be complemented with other modelling tools, especially when it comes to testing hypotheses or studying the dynamics that drive trajectories. The powerfulness of SA as an exploratory tool has been largely demonstrated by many substantive studies that use SA tools; moreover, most of these studies run SA in conjunction with other approaches, typically involving the use of the obtained typology of sequences either as an explanatory variable or as the response variable in a regression analysis. In these studies, the SA outcome serves as input for the regression stage, but we could imagine using regression outcome, e.g., the most contrasting profiles with respect to the regression response variable, to guide the SA analysis.

One advantage often discussed for SA is its holistic perspective (see e.g., Billari 2005) meaning that SA sheds light on the entire trajectory rather than, for example, on specific transitions in the trajectory. With this holistic perspective, sequences are considered as static objects, which is not suited for studying the process that generates the sequences. For investigating sequence dynamics, we require alternative methods such as probabilistic models of the occurrences of successive transitions in the sequence. As already mentioned above, an issue with such transition models is the difficulty to present their outcome synthetically. Here, SA could help rendering the outcome of the modelling phase.

As illustrated, by the abovementioned two examples, an intimate combination of SA and related methods seems necessary to achieve a better understanding of life course data. We describe below how the chapters of this book lead in that direction.

3 Directions for the Future: The Chapters of this Book

In Part I, two chapters address the relationship between SA and other methods for analysing longitudinal data. In the first chapter, Daniel Courgeau (2018) examines four major approaches for longitudinal analysis in population science, namely, approaches based on sequences, duration between events, multiple levels, and networks. He first identifies the proper characteristics of each approach in terms of the mathematical tools involved and the considered statistical unit (e.g., event,

individual, group), among others. He then depicts a general robust program that could lead to the convergence of these different models. The next chapter by Mervi Eerola (2018) discusses three original ways of combining SA—in fact, the clustering of sequences—and probabilistic modelling. This is done through a short presentation of three case studies of life course analysis. With case study 1, Mervi Eerola considers the combined use of SA and prediction probabilities obtained with a marked point process—a kind of multistate model. In case study 2, SA is used to identify pathways to adulthood and is used in conjunction with a structural equation factor model of social and achievement strategies, and a model for transitional pathways accounting for the strategies. In case study 3, SA is used to identify the most vulnerable individuals among Finns between 18 and 25 years of age and Mervi Eerola addresses the use of either a latent transition model or a HMM for analysing their risk pattern, e.g., risk to be outside work force, to be living on social benefits, and to have the lowest educational attainment. These three case studies are good illustrations of the many possibilities of combining SA with modelling approaches.

Part II is devoted to the combination of SA and event history analysis or equivalently, survival analysis. The strength of the connection between SA and the survival models increases in each of the successive chapters. In the first chapter, Malin and Wise (2018) propose a study of gender differences in career advancement across occupations in West Germany. In this study, sequence visualization is used to provide an overall view of the data at hand, and the focus is then placed on the study of the time to a leadership position and time to leaving a leadership position by means of Kaplan-Meier (KM) survival curves and Cox regressions. The connection between sequence and survival analysis remains loose and no explicit SA outcome is used in the survival analysis. The second chapter by Lundevaller et al. (2018) studies the mortality of disabled and non-disabled individuals in nineteenth century Sweden. A classical clustering of life trajectories is realized—separately for women and men—and the obtained types are used as covariates in the survival analyses carried out with stratified KM and Cox regressions. Finally, the chapter by Rossignon et al. (2018) introduces an innovative method—called Sequence History Analysis (SHA)—where SA and event history analysis are more tightly entwined. The method consists of an event history analysis that accounts for the past trajectory at each time point. More specifically, SA is used to determine the type of past trajectory at each time point, which makes the past trajectory type a time-varying covariate. This method is applied to study how the risk of leaving home depends on the past co-residence trajectories in Switzerland.

Part III includes two papers concerned with the network approach in SA. In the first chapter, Benjamin Cornwell (2018) starts by explaining, with some detail, how a sequence can be represented as a network with states as nodes and time-adjacency between states as links. He then shows how a series of network concepts—such as network density, centralization, and homophily—prove useful for characterizing the structure of individual sequences and to compare multiple sequence structures with each other. The approach is illustrated with an analysis of daily activities using data from the American Time Use Survey. The next chapter by Klaus Hamberger (2018) introduces relational networks and demonstrates their scope in a study of mobility

patterns in Togo. Relational networks are built from networks of kinship and mobility relations. The nodes of the relational network are the classes of (mobility) events obtained by classifying the events according to the type of relation—e.g., kinship, employer, friend—between the individuals involved. The arcs indicate the immediate succession of the events. Then, the author proposes two complementary ways of using the personal networks. First, he aggregates the individual networks into one network for women and one for men with node sizes and arc widths proportionate to the observed counts, which allows the visual identification of the gender differences in the mobility itineraries. Second, he orders the individual networks along a spanning tree. Here, again, we see that women and men occupy different areas of the tree and thus reveal gender differences.

Part IV is composed of three chapters that attempt to gain knowledge about the process behind the observed trajectories. The chapter by Thomas Collas (2018) suggests that life trajectories decompose into phases and follow a different logic—possibly characterized with a different alphabet—in each phase. He shows how such multiphase trajectories can be formalized and rendered, and proposes a dissimilarity measure that can account for this decomposition into phases. The method is illustrated with an application to competition trajectories of French pastry cooks with an explicit distinction between the junior and senior phases. The next chapter by Borgna and Struffolino (2018) combines SA with qualitative comparative analysis (QCA), which is a method related to the mining of association rules, to find out factor configurations that are 'logically sufficient' to be in employment or education at crucial time points in divided Germany. Discrepancy analysis, more specifically the analysis of the evolution of the discrepancy among sequences along the time frame, is used to identify the crucial turning points in the divergence between sequences. QCA is then applied at the identified turning points to find out the relevant 'sufficient' factor combinations. The third chapter by Helske et al. (2018) also considers that trajectories belonging to a same group share similar phases. Here, however, the phases are not predefined, but are associated to the hidden states of a HMM and thus probabilistically determined. In fact, this chapter presents a general framework for the combined use of SA and HMM to analyse multichannel sequences. SA is used to cluster the sequences and then the HMM is used to identify similar phases within each group. In addition, the authors propose two original compressed representations of a group of (multichannel) sequences, including a graph of the structure of the hidden states and the transitions between them, and plots of the most probable individual pathways predicted by the HMMs for each group. The method is illustrated with data from the German National Educational Panel Survey (NEPS).

Part V includes three chapters that present advances in the original task of SA, namely the clustering of sequences. The chapter by Taushanov and Berchtold (2018) proposes clustering sequences of continuous data by means of a Markov-based mixture model—the hidden mixture transition distribution (HMTD) model—and applying their method to Swiss data obtained from the internet addiction test (IAT). The clustering is achieved by setting the transition matrix of the hidden states as the identity matrix, which makes their model a mixture of Gaussian distributions. One

advantage of the proposed approach is the possibility of accounting for covariates at the clustering level. The authors compare the results provided by their method with those obtained by means of a growth mixture model (GMM). Note that this chapter is the only one in that volume that deals with continuous data. The next chapter by Matthias Studer (2018) investigates two original dissimilarity-based ways of clustering sequences: a divisive property-based method and fuzzy clustering. The former orders the splits of a discrepancy-based regression tree that provides classification rules defined in terms of covariates, and considers the partition that results from the splits up to a given, optimally chosen rank. Splits are ordered according to the overall share of reduction of discrepancy that each of them produces. In fuzzy clustering, each sequence may belong to more than one cluster and cluster membership may have different degrees. The method is especially useful when clusters are not well separated. Here, the author addresses a series of issues such as the graphical representation of fuzzy clusters and how to measure the effect of covariates on the individual strengths of membership. The methods are illustrated with the school-to-work transition data from McVicar and Anyadike-Danes (2002). The last chapter by Bison and Scalcon (2018) focuses on the measure of the dissimilarity between sequences. The authors decompose each sequence into basic binary sequences that indicate for each element of the alphabet whether it is active at successive time points. Then, they associate to each binary sequence two index numbers, the first one indicating the proportion of time spent in the concerned state and the second one synthesizing when the concerned state is actually observed. The dissimilarity between a pair of binary sequences is obtained as the Euclidean distance between the couples of index numbers, and the dissimilarity between the original—possibly multichannel—sequences as the sum of the distances between the underlying binary sequences. The method is applied to cluster data describing the time-use during a typical day of Italian dual earners.

Finally, the book concludes with Part VI with two papers concerned with summary numbers of individual sequences. Both papers aim to measure the quality of sequences, i.e., to get an index that would allow the ranking of, for example, occupational sequences from the most negative—insecure, undesirable—to the most positive ones. The solutions proposed are, however, very different. In the first chapter, Manzoni and Mooi-Reci (2018) assume that each state of the alphabet can be classified as a success or a failure. Then, the proposed quality index is defined so as to increase with the proportion and recency of the successes. The index is applied to a study of the quality of employment career after a first spell of unemployment using data from HILDA, an Australian household survey. In the second chapter by Ritschard et al. (2018), the quality index—named precarity index—is defined based on the quality of the transitions rather than the states themselves. Assuming that the states of the alphabet can be (partially) ordered, the authors define the index as a complexity index corrected by a factor that depends on the proportion of upward and downward transitions. The scope of the index is illustrated with the school-to-work transition data from McVicar and Anyadike-Danes (2002) by showing the strong impact of the quality of the initial trajectory on the situation three years later.

4 Conclusion

To conclude, let us highlight how this volume traces the expected trend for the future of SA. First, we observe a shift in the methodological concern. The measure of dissimilarities between sequences that was the major central aspect of SA until recently, is the concern of only two papers (Collas 2018; Bison and Scalcon 2018) out of the fifteen.

The trend seems to be more oriented toward alternative approaches for SA and the combined use of dissimilarity-based SA with other related methods. Five papers address alternatives to the classical 'compute dissimilarities—partition the set of sequences' approach. These alternatives include feature-based and fuzzy clustering (Studer 2018) and non-dissimilarity-based methods such as those based on network representations of sequences (Cornwell 2018; Hamberger 2018), and Markov-based models (Helske et al. 2018; Taushanov and Berchtold 2018). Alongside the two general papers (Courgeau 2018; Eerola 2018) in Part I, five papers demonstrate the benefit of combining SA with other methods to grasp the dynamics that drive the trajectories. SA is combined with survival models (Malin and Wise 2018; Lundevaller et al. 2018; Rossignon et al. 2018), with QCA (Borgna and Struffolino 2018), and with hidden Markov models (Helske et al. 2018).

Finally, there seems to be an increasing interest in individual sequence summaries. Such summary numbers are central in Part VI (Manzoni and Mooi-Reci 2018; Ritschard et al. 2018), but also play an important role in two other chapters (Cornwell 2018; Bison and Scalcon 2018).

Acknowledgements The authors acknowledge 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 (grant number: 51NF40-160590).

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., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History, 16*, 471–494.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research, 29*(1), 3–33. (With discussion, pp 34–76).

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 and Research, 38*(3), 430–462.

Barban, N., & Billari, F. C. (2012). Classifying life course trajectories: A comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 61*(5), 765–784.

Billari, F. C. (2005). Life course analysis: Two (complementary) cultures? Some reflections with examples from the analysis of transition to adulthood. In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, & E. Widmer (Eds.), *Towards an interdisciplinary perspective on the life course* (Advances in life course research, Vol. 10, pp. 267–288). Amsterdam: Elsevier.

Bison, I. (2014). Sequence as network: An attempt to apply network analysis to sequence analysis. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 231–248). Cham: Springer.

Bison, I., & Scalcon, A. (2018). From 07.00 to 22.00: A dual-earner couple's typical day in Italy. Old questions and new evidence from social sequence analysis. In Ritschard and Studer (2018) (this volume).

Bolano, D., Berchtold, A., & Ritschard, G. (2016). A discussion on hidden Markov models for life course data. In *Proceedings of the International Conference on Sequence Analysis and Related Methods*, Lausanne, 8–10 June 2016.

Borgna, C., & Struffolino, E. (2018). Unpacking configurational dynamics: Sequence analysis and qualitative comparative analysis as a mixed-method design. In Ritschard and Studer (2018) (this volume).

Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review, 23*(4), 409–422.

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with Stata. *The Stata Journal, 6*(4), 435–460.

Bürgin, R., & Ritschard, G. (2014). A decorated parallel coordinate plot for categorical longitudinal data. *The American Statistician, 68*(2), 98–103.

Butts, C. T., & Pixley, J. E. (2004). A structural approach to the representation of life history data. *The Journal of Mathematical Sociology, 28*(2), 81–124.

Collas, T. (2018). Multiphase sequence analysis. In Ritschard and Studer (2018) (this volume).

Cornwell, B. (2018). Network analysis of sequence structures. In Ritschard and Studer (2018) (this volume).

Cornwell, B., & Watkins, K. (2015). Sequence-network analysis: A new framework for studying action in groups. In S. R. Thye & E. J. Lawler (Eds.), *Advances in group processes* (Vol. 32, pp. 31–63). Bingley: Emerald Group Publishing Limited.

Courgeau, D. (2018). Do different approaches in population science lead to divergent or convergent models? In Ritschard and Studer (2018) (this volume).

Eerola, M. (2018). Case studies of combining sequence analysis and modelling. In Ritschard and Studer (2018) (this volume).

Elzinga, C. H. (2010). Complexity of categorical time series. *Sociological Methods & Research, 38*(3), 463–481.

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, 23*, 225–250.

Fasang, A. E., & Liao, T. F. (2014). Visualizing sequences in the social sciences: Relative frequency sequence plots. *Sociological Methods & Research, 43*(4), 643–676.

Gabadinho, A., & Ritschard, G. (2013). Searching for typical life trajectories applied to childbirth histories. In R. Levy & E. Widmer (Eds.), *Gendered life courses – Between individualization and standardization. A European approach applied to Switzerland* (pp. 287–312). Vienna: LIT-Verlag.

Gabadinho, A., & Ritschard, G. (2016). Analysing state sequences with probabilistic suffix trees: The PST R library. *Journal of Statistical Software, 72*(3), 1–39.

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des nouvelles technologies de l'information RNTI, E-19*, 61–66.

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. (2014). SADI: Sequence analysis tools for Stata. Department of Sociology Working Paper Series WP2014-03, University of Limerick.

Halpin, B. (2015). MICT: Multiple imputation for categorical time-series. Department of Sociology Working Paper Series WP2015-02, University of Limerick, Ireland.

Hamberger, K. (2018). Relational sequence networks as a tool for studying gendered mobility patterns. In Ritschard and Studer (2018) (this volume).

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal, 26*(2), 147–160.

Helske, S., & Helske, J. (2017). Mixture hidden Markov models for sequence data: The seqHMM package in R. Vignette of the seqHMM package, CRAN.

Helske, S., Helske, J., & Eerola, M. (2018). Combining sequence analysis and hidden Markov models in the analysis of complex life sequence data. In Ritschard and Studer (2018) (this volume).

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707–710.

Lundevaller, E. H., Vikström, L., & Haage, H. (2018). Modelling mortality using life trajectories of disabled and non-disabled individuals in 19th-century Sweden. In Ritschard and Studer (2018) (this volume).

Malin, L., & Wise, R. (2018). Glass ceilings, glass escalators and revolving doors: Comparing gendered occupational trajectories and the upward mobility of men and women in West Germany. In Ritschard and Studer (2018) (this volume).

Manzoni, A., & Mooi-Reci, I. (2018). Measuring sequence quality. In Ritschard and Studer (2018) (this volume).

McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A, 165*(2), 317–334.

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. (2017). Joint sequence analysis: Association and clustering. *Sociological Methods & Research, 46*(2), 252–287.

Piccarreta, R., & Elzinga, C. H. (2013). Mining for association between life course domains. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (Quantitative methodology, pp. 190–220). New York: Routledge.

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 A, 170*(1), 167–183.

Ritschard, G., & Studer, M. (Eds.) (2018). *Sequence analysis and related approaches: Innovative methods and applications*. (Life course research and social policies). Cham: Springer.

Ritschard, G., Bussi, M., & O'Reilly, J. (2018). An index of precarity for measuring early employment insecurity. In Ritschard and Studer (2018) (this volume).

Rossignon, F., Studer, M., Gauthier, J.-A., & Le Goff, J.-M. (2018). Sequence history analysis (SHA): Estimating the effect of past trajectories on an upcoming event. In Ritschard and Studer (2018) (this volume).

Sankoff, D., & Kruskal, J. B. (Eds.) (1983). *Time warps, string edits, and macro-molecules: The theory and practice of sequence comparison*. Reading: Addison-Wesley.

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European Sociological Review, 17*(2), 119–144.

Studer, M. (2018). Divisive property-based and fuzzy clustering for sequence analysis. In Ritschard and Studer (2018) (this volume).

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society, Series A, 179*(2), 481–511.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research, 40*(3), 471–510.

Studer, M., Liefbroer, A. C., & Mooyaart, J. E. (2018a). Understanding trends in family formation trajectories: An application of competing trajectories analysis (CTA). *Advances in Life Course Research, 36*, 1–12.

Studer, M., Struffolino, E., & Fasang, A. E. (2018b). Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. *Sociological Methodology*. (First Published Online).

Taushanov, Z., & Berchtold, A. (2018). Markovian-based clustering of internet addiction trajectories. In Ritschard and Studer (2018) (this volume).

Vermunt, J., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 373–385). Burlington, MA: Elsevier.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part I
About Different Longitudinal Approaches in Longitudinal Analysis

Do Different Approaches in Population Science Lead to Divergent or Convergent Models?

Daniel Courgeau

1 Introduction

Since its introduction by Graunt in 1662, the scientific study of population, initially called *political arithmetick*, has become possible not only for demography but also for epidemiology, political economy, and other fields. For more than 200 years, researchers adopted a *cross-sectional approach* in which social facts in a given period exist independently of the individuals who experience them, and can be explained by various characteristics of the society of that period. After the end of World War II, researchers examined social facts from a new angle, introducing individuals' life experience. This *longitudinal approach* holds that the occurrence of a given event, during the lifetime of a birth cohort, can be studied in a population that maintains all its characteristics as long as the phenomenon persists. However, this condition was too restrictive, triggering the development of new approaches that we shall discuss in greater detail here, with an emphasis on the scope for convergence or divergence.

From the comparison of these new approaches, we shall try to identify the conditions that would allow a synthesis of these approaches by means of a Baconian inductive analysis. This induction method consists in discovering the principles of a social process by experiment and observation. It is therefore based on the fact that, without these principles, the observed properties would be different. It will enable us to draw a conclusion.

D. Courgeau (✉)
Institut National d'Etudes Démographiques (INED), Paris, France
e-mail: daniel.courgeau@wanadoo.fr

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_2

2 Different Approaches

We shall describe and focus our discussion on four major approaches based on: duration between events, sequences, multiple levels, and networks; we set aside agent-based models (i.e., based on agents' decisions) which are of a totally different type. We will try to show in what these approaches are different and in what they may converge.

2.1 An Approach Based on Duration Models

The first approach emerged in the social sciences in the early 1980s, more than 30 years after the introduction of longitudinal analysis. However, it was already used by statisticians, such as Ville in 1939 and Doob in 1953, in association with the concept of martingale. It was Cox who, in 1972, recommended the simultaneous use of life tables and regression methods, and Aalen in 1975 who proposed the use of counting process theory for the simultaneous analysis of several events that an individual may experience over time.

The principle of this approach is that "throughout his or her life, an individual follows a complex life course, which depends at any moment on the past history and the information acquired previously" (Courgeau and Lelièvre 1997, p. 5).

This "event-history" approach rests on robust mathematical and statistical foundations that make it possible to determine risk factors and to process censored observations. It has been described in many statistical works since 1980 (Kalbfleisch and Prentice 1980; Cox and Oakes 1984; Aalen et al. 2008). It can be used to analyze changes of state, however diverse, and to demonstrate the role of many individual characteristics that can change over time or during transitions. The application of these methods in demography yielded significant advances with respect to longitudinal analysis (see, for example, Courgeau and Lelièvre 1992). Many other social sciences adopted it including epidemiology, biostatistics, sociology, econometrics, actuarial sciences, and medicine.

The event-history approach eliminates the need for the highly restrictive hypotheses of longitudinal analysis while maintaining the individual point of view. Individuals can be followed for relatively long spells of their lives by means of retrospective or prospective surveys that track a large number of events. For example, the French 1981 "Triple event history" survey ("3B survey") allowed a simultaneous analysis of family-related, occupational, and migration-related events occurring from their birth to the survey date, for cohorts born between 1911 and 1936 (Courgeau 1999). As the approach makes it possible to process censored observations, the persons still economically active at the time of the survey (four-fifths of the sample) could be studied over the duration of their working careers in the same way as the persons already retired (one-fifth of the sample).

The approach relies mainly on semi-parametric methods that, while maintaining a nonparametric vision of time between events, use parameters to describe the effect of personal characteristics.

However, the event-history approach raises a set of problems to which we now turn. All the problems mentioned here are more structural than technical.

The first problem concerns unobserved heterogeneity. How does this heterogeneity affect the parameters of observed characteristics? An important result obtained by Bretagnolle and Huber-Carol (1988), but often overlooked by the models' users, provides an answer to the question. These authors show that when omitted characteristics are independent of observed characteristics, the omission does not impact the signs of the estimated parameters, but only reduces their absolute values. Therefore, if the effect of a characteristic is fully significant, the introduction of the initially unobserved characteristic will increase this effect alone. Conversely, a characteristic without significant effect may have one when the unobserved characteristics are introduced (Courgeau and Lelièvre 1992). One must be aware of this risk.

When the observed and omitted characteristics are interdependent, the situation is more complex. It may be tempting to introduce this heterogeneity in the form of a distribution of a known type, which Vaupel et al. (1979) have called *frailty*. When information on this distribution is available, its introduction is entirely legitimate. The problem is that, in most cases, we know nothing about this distribution, and it is often chosen for no other valid reason than convenience. In such circumstances, some distributions may change the sign of an estimated parameter, whereas a model without frailty avoids this drawback (Trussell and Richards 1985).

We therefore totally agree with Aalen et al. (2008, p. 425), who, in their extensive studies on stochastic processes, have tried to identify individual frailty:

> As long as there is no specific information about the underlying process and observations are only made once for each individual, there is little hope of identifying what kind of process is actually driving the development.

Indeed, for the analysis of non-repetitive events, there is only one model without unobserved heterogeneity, but an infinity of models with unobserved heterogeneity (Trussell 1992). Their estimates differ, but none has a solid justification. By contrast, if we are analyzing repetitive events—such as successive births or migrations—we have the option of estimating multilevel models that allow the introduction of unobserved heterogeneity, which reflects the multiple events experienced by every individual. We shall present these multilevel models later.

The second problem concerns the concept of probability used. Most of the statisticians who developed the method chose an *objective probabilistic approach*. Could an *epistemic approach* to probability enable us to lift some of the constraints involved in objective probability? Space precludes a full description of these different approaches, but we can focus on the constraints linked to *statistical inference* (Courgeau 2012).

The purpose of statistical inference is to optimize the use of incomplete information in order to arrive at the best decision. Statistical inference will therefore

consist in providing the best possible analysis of a past phenomenon and the best possible prediction of a similar phenomenon to come. The first point is important for sciences such as demography or epidemiology, which must analyze human behavior. The second point is crucial for disciplines such as medicine or public health, which aim to produce the best possible forecast of the outcome of a treatment course or a decision on the best policy to implement for achieving a specific effect. Statistical inference leads, among other things, to testing various hypotheses about the phenomena studied.

Objectivist methods seek to verify whether a given factor does or does not affect the phenomenon studied. This brings us to the notion of statistical test, which involves treating the sample under analysis as one possible selection from an infinity of samples that we could extract from a population also assumed to be infinite. When we assign a confidence interval of, say, 95% to a parameter estimated on this sample, we wish to conclude that the probability of the unknown parameter lying in the interval is 0.95. In fact, however, the objectivist tells us that this conclusion is wrong. All we can state is that if we draw an infinity of new samples, then the new estimated parameters will lie in that interval 95% of the time. As Jeffreys wrote (1939, p. 377):

> The most serious drawback of these definitions, however, is the deliberate omission to give any meaning to the probability of a hypothesis. All they can do is to set up a hypothesis and give arbitrary rules for rejecting it in certain circumstances.

That is exactly what happens with objectivist statistical tests. Similarly, the use of frequentist methods—here, for prediction—will consist in taking the parameters estimated, for example, by means of maximum likelihood and introducing them into the distribution function of the new observation. But this does not allow us to factor in the uncertainty of the parameter estimations, and it will cause us to underestimate the variance of the predicted distribution.

This is why Jeffreys showed that, if we adopt an epistemic approach, a 95% confidence interval will mean precisely an interval in which the statistician can conclude that the unknown parameter will lie with a probability of 0.95. Moreover, this approach offers a more satisfactory resolution of the prediction problem than we could obtain with objective probability. All we need to do is calculate the "posterior predictive distribution" of a future observation from the initially observed data, which are known. What we obtain is not a single value, as with the objectivist solution, but a distribution.

These advantages of an epistemic method have led a number of authors to propose it for event-history analysis (Ibrahim et al. 2001).

The final problem we would like to discuss is the risk of atomistic fallacy involved in the event-history approach (Robinson 1950). If we can take all individual characteristics into account to explain a behavior, we may overlook the context in which the behavior occurs. If, instead, we use a cross-sectional approach, we only introduce the characteristics of society to explain social facts. This aggregate approach is, by contrast, vulnerable to the risk of ecological fallacy. We can easily show that the relationships between two characteristics measured on individuals

or on proportions applied to different aggregates are generally far from identical (Courgeau 2007). In the third subsection of this section, we shall see the solution for overcoming these divergences.

2.2 An Event Sequences Approach

Sequence analysis was first introduced in computer science (Levenshtein 1966), then in molecular biology to study DNA and RNA sequences (Levitt 1969). It was imported into the social sciences by the sociologist (Abbott 1983, 1984) to study the social processes occurring in sequences over a long period.

The main principle of this approach is that social organization derives "from the regular and predictable pattern of temporal, spatial, hierarchical and other ordered phenomena that result". (Cornwell 2015, p. 24–25). So that the key assumption is that there is such a pattern and that it is socially meaningful. Its purpose appears to be quite different from the one of the previous models, and is now based in the order rather than in the duration of phenomena.

In the social sciences, however, this approach rests on less robust mathematical and statistical foundations than event-history analysis. In its most common setting, its goal is to describe complete sequences in terms of types reflecting socially significant trajectories of subjects (individuals or more general entities such as stimuli in psychology or artifacts in archeology). The approach comprises two stages. First, it tries to calculate a distance between sequences with the aid of certain operations (insertions or deletions called "indels" or substitutions) with a given cost for each operation. The most widely used metric is called Optimal Matching (OM), but we shall discuss other methods for calculating distances below. In the second stage, cluster analysis is used to detect types of sequences, grouping subjects into mutually exclusive categories. Many natural, biologic and social sciences use these methods, including computer science, biology, sociology, demography, psychology, anthropology, political science, and linguistics.

With sequence analysis, we can move from a Durkheimian search for causes (1895), to an emphasis on contexts, connections, and events—a shift about which (Abbott 1995, p. 93) says "a quiet revolution is underway in social science." The surveys tracking individuals for long spells of their lives resemble event-history surveys, with a focus on observing complete processes without censorship. Their objectives, by contrast, are very different: while event-history analysis seeks the causes of phenomena, sequence analysis explores the paths followed without looking for the reasons for the underlying processes that generate the paths (Robette and Bry 2012). Individual characteristics are thus of little value in this approach, apart from the sequence of events and certain characteristics preceding the sequence analyzed. For instance, the 2001 survey by Lelièvre on "Event histories and contact circle," covering a sample of cohorts born between 1930 and 1950, applied a sequence analysis of occupational trajectories of mothers and daughters in order to compare them (Robette et al. 2012). However, there exist some attempts to look at how individual factors may explain the observed heterogeneity between sequences (Studer et al. 2011).

This approach is largely based on non-parametric methods that make no assumptions about the underlying process over a lifetime. Its goal is to explore and describe the course of events as a whole, without worrying about the risk of knowing the events or their determinants. There have also been recent attempts at a Bayesian extension of the sequence approach with the aid of Hidden Markov Models (Bolano 2014; Helske et al. 2018) or variable length Markov models (Gabadinho and Ritschard 2016).

Sequence analysis, in turn, raises a new set of problems that differ from those of event-history analysis. The first problem is mainly a technical one, the others being more theoretical.

The first problem concerns the metric used in social sciences. The OM metric was imported from information theory and molecular biology, where it is fully justified by the basic assumptions. In the social sciences, however, the structure of sequences is far more complex and the metric used is less self-evident. As Wu (2000, p. 46) notes:

> Part of my skepticism stems, in part, from my inability to see how the operations defining distances between trajectories (replacements and indels) correspond, even roughly, to something recognizably social.

For example, if we interpret substitutions as transitions, assigning the same cost for a substitution from employment to non-employment and for the substitution from non-employment to employment seems scarcely plausible. Bison (2009) clearly shows, using simulations, that different substitution costs yield inconsistent results. As a result, we can find regularities even when none exist (Bison 2014).

To solve these problems, a number of generalizations of the OM method have been proposed, such as variable substitution costs, different distance measures, spell-adjusted measures, non-alignment techniques, and monothetic divisive algorithms. However, with the increase in the number of distances and costs, their comparability becomes increasingly difficult. We do have some comparisons between different metrics, but the only studies comparing a large number of metrics using a set of artificial sequences are those of Robette and Bry (2012) and Studer and Ritschard (2016). In the former study, the authors did not try to find the best metric but "rather to unravel the specific patterns to which each alternative is actually more sensitive" (p. 2). Although they found differences between the results obtained with different metrics, "the main patterns they conceal will be uncovered by most of the metrics" (p. 14). However, these differences exist, and Bison's inconsistent results leave the comparability problem largely unsolved.

The second problem lies in the use of cluster analysis to detect classes of sequences. This classification method was used long before sequence analysis, for example by a psychologist (Tryon 1939) for manual calculation. The advent of computers spurred the development of many methods to detect significant groups, but also raised numerous complications, which we shall summarize here.

One of the most important criteria for a good classification is the number of groups that should exist in a given study. Unfortunately, when the classification criterion is plotted against the number of groups, in most cases, there is no "sharp

step" that we can use to determine the ideal number of classes. The choice becomes highly subjective (Everitt 1979). The assessment of the validity and stability of the clusters found using different approaches is equally problematic. Regrettably, there are few validity tests for these approaches, and even fewer tests of their social significance. A recent comment (Byrne and Uprichard 2012, p. 11) concludes: 'Although written in the late 1970s, actually many of the "unresolvable problems" raised in Everitt's article are still problems today'.

The emphasis on context, connections, and events leads sequence analysis to abandon regression methods and to view the search for causes as obsolete. This raises a third problem: "whether the clusters obtained under this method might be an artifact or something else social" (Wu 2000, p. 51). While unobserved heterogeneity was a significant problem in event-history analysis, here even observed heterogeneity creates difficulties. As sequence analysis tries to capture trajectories as a whole, the only characteristics that can be introduced are the ones measured before the start of the trajectory. Introducing characteristics measured later, or time-dependent characteristics, will raise a host of conceptual problems that are hard to solve. In section three below, however, we shall see that new attempts to combine event-history and sequence analysis may offer a partial solution to these problems.

A fourth problem is linked to the fact that sequence analysis—unlike event-history analysis—cannot fully handle censored observations (Wu 2000, p. 53; Studer et al. 2018). Such a limitation entails the exclusion of incomplete trajectories and confines us to a study of the past. For example, as French retirement age was 65 years at the time of the 3B survey, a sequence analysis of occupational careers would have to be confined to persons born between 1911 and 1916, i.e., one-quarter of the sample.

Sequence analysis is intended to allow a description of trajectories in terms of classes, meant to reflect types of social behavior adopted by groups of individuals. The fifth problem raised is that the meaning of these behaviors is not as clear as one might imagine. First, as an individual is assigned to only one type, the resulting classification is very narrow, whereas we know that an individual may in fact be assigned to a large number of groups such as family, business firms, organizations, and contact circles. These groups are real entities, whereas the classes obtained with sequence analysis are open to question. Consequently, what are the grounds for believing in the existence of these types? Abbott and Tsay (2000, p.27) argue that sequence methods "would find this particular regularity because people in particular friendship networks would turn up in groupings of similar fertility careers". Their argument, however, assumes that data on these individuals' friendship networks are available simultaneously with data on their fertility history. To the best of my knowledge, however, there are no examples showing the congruence of cluster results with friendship networks, but only examples of impact of trajectories on personal network (Aeby et al. 2017).

More recently, several authors have similarly argued that network analysis may be a valuable tool for solving these problems. Bison (2014), for example, suggests converting individual sequences into network graphs. While this method makes it possible "to bring out career patterns that have never previously been observed" (p. 246), it has major limitations. The most important one, advanced by Bison (id),

> that creates the greatest methodological and philosophical problems is the annulment of individual sequences. [...] 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.

If we want to stay in the purely descriptive field of sequence analysis, this characteristic is a genuine hindrance. More recently, Cornwell (2015) goes further and devotes an entire chapter to "Network methods for sequence analysis" (p. 155–209). While some methods used in network analysis may be useful in sequence analysis, it is important to grasp the difference between the goals of the two approaches. The main goal of sequence analysis, as noted earlier, is to understand a life history as a whole and to identify its regularities and structures. Network analysis, as we shall see in the fourth subsection of this section, is focused on understanding the relationships between entities (individuals, or more general levels of "collective agency") and to see how changes at each level drive changes at other levels. We suggest a solution to this problem in the final synthesis of the third section.

2.3 A Level Based Approach

While the two preceding analyses operated at a given aggregation level, we shall now introduce the effects of multiple levels on human behavior. These methods derive from the hierarchical models used in biometrics and population genetics since the late 1950s (Henderson et al. 1959). They were then applied in the social sciences—in sociology by Mason et al. (1983) and in education science in 1986 by Goldstein (2003).

The simplest solution is to incorporate into the same model the individual's characteristics and those of the groups to which (s)he belongs. These "contextual" models differ from cross-sectional models, which explained aggregate behavior by equally aggregate characteristics. We can thus eliminate the risk of ecological fallacy, for the aggregate characteristic will measure a different construct from its equivalent at individual level. It now acts not as a substitute, but as a characteristic of the sub-population that will influence the behavior of one of its members. Simultaneously, we remove the atomistic fallacy, as we take into consideration the context in which the individual lives.

However, contextual models impose highly restrictive conditions on the formulation of the log-odds (logarithm of relative risks) as a function of characteristics. In particular, the models assume that individual members of a group behave independently of one another. In practice, the risk incurred by a member of a given group depends on the risks encountered by the group's other members. Overlooking this intra-group dependence biases the estimates of the variances of contextual effects, generating excessively narrow confidence intervals. Moreover, for individuals in different groups, the log-odds cannot vary freely but are subject to tight constraints (Loriaux 1989; Courgeau 2007).

Multilevel models offer a solution to this double problem. By incorporating different aggregation levels into a single model, they generalize the usual regression models. The basic assumption is that the groups' residuals are normally distributed. The analysis can thus focus exclusively on their variances and covariances, but may introduce individual or group characteristics at different levels.

Multilevel analysis no longer focuses on the group, as in the aggregate approach, or on the individual, as in the event-history approach. Instead, it incorporates the individual into a broader set of levels. It thus resolves the antagonism between holism and methodological individualism (Franck 1995, p. 79):

> Once we have admitted the metaphysical or metadisciplinary concept of hierarchy, it no longer makes sense to choose between holism and atomism, and—as regards the social sciences—between holism and individualism.

We can finally say that this approach regards "a person's behavior as dependent on his or her past history, viewed in its full complexity, but it will be necessary to add that this behavior can also depend on external constraints on the individual, whether he or she is aware of them or not" (Courgeau 2007, p. 79). It can be seen as complementary of event-history analysis but is less linked to sequence analysis.

This approach, however, requires new types of surveys to define and capture the various levels to examine (Courgeau 2007). It has been used in biometrics, population genetics, education science, demography, epidemiology, economics, ecology, and other disciplines. Its methods are basically semi-parametric but can take non-parametric forms, as in factor analysis models.

Although some of these models use the frequentist paradigm, they generally adopt the Bayesian paradigm in order to deal effectively with nested or clustered data (Draper 2008). However, as Greenland (2000) notes, the multilevel approach makes it possible to unify the two paradigms, leading to an empirical Bayes estimate.

But again new problems arise: the three first ones are technical while the last one is mainly structural.

As group characteristics, the multilevel approach often uses mean values, variances or even covariances of group members' characteristics. The first problem is that we must go beyond this approach, for we need a fuller definition of the aims and rules prevailing in a group in order to explain a collective action. What are the mechanisms of social influence that permit the emergence of a collectively owned social capital in different contexts—a capital that "is more than the sum of the various kinds of relationship that we entertain"? (Adler and Kwon 2002, p. 36).

The second problem is that "independence among the individuals derives solely from common group membership." (Wang et al. 2013, p. 125). In fact, the groups are generally more complex. For example a family, generally treated as a simple group, is composed of parents and children, who can play very different and even conflicting roles. This dissymmetry of roles partly undermines the value of the family for multilevel analysis, in which we are looking for what unites group members rather than what divides them. As a result, we take into account the interactions between group members and their changes over time in order to fully incorporate their social structure. In the next subsection, we discuss how a multilevel network approach makes it possible to avoid this problem.

The third problem stems from the difficulty of defining valid groups. It leads to the use of geographic or administrative groupings that often have little impact on their inhabitants' behavior. However, by observing existing networks through more detailed surveys, such as those included in the Stanford Large Network Data Set Collection, we should be able to avoid using these unsatisfactory groupings.

The fourth problem is that, while multilevel analysis enables us to incorporate a growing number of levels that constitute a society, it continues to focus on only one of these levels—an event, an individual or a group. As a result, this "approach assumes that links between groups are non-existent." (Wang et al. 2013, p. 1). On the contrary, it is important to take the analysis further by trying to identify the interactions that necessarily exist between the various levels. In Franck's words (1995, p. 79): "the point now is to determine how the different stages or levels connect, from top to bottom and from bottom to top." We shall now see how the analysis of social networks allows us to solve this problem.

2.4 A Network Based Approach

While earlier examples exist, research on social networks effectively began with the work of the sociologists Moreno and Jennings, particularly with a paper (1938) in which they used the term "network theory" and proposed statistics of social configurations. Until the 1970s, however, while research teams in various social sciences worked on network analysis, no cumulative theory resulted (Freeman 2004). Social networks did not begin to be regarded as a full-fledged research field until the 1970s and 1980s. The development of structural models introduced by White et al. (1976) and Freeman (1989) made it possible to examine the interdependent relationships between actors and the similar relationships between actors' positions in the different social networks.

The principle of this approach is to "identify different levels of agency, but also intermediary levels and social forms (such as systems of social niches and systems of heterogeneous dimensions of status), and relational infrastructures that help members in constructing new organizations at higher levels of agency and in managing intertwined dilemmas of collective actions" (Lazega and Snijders 2016, p. 360). It permits to answer to the two last structural problems of sequence analysis, in introducing networks, and the last one of multilevel analysis, in introducing the interactions which exist between the various levels.

This approach rests on robust mathematical foundations. These, however, differ substantially from the previous ones, as the assumption that observations of individuals are independent no longer holds: network analysis argues that units do not act independently but influence each other. The use of graph theory and matrix analysis is important in this field (Wasserman and Faust 1994). Many disciplines—and not only the social sciences—have adopted this approach. They include information science, computer science, management, communication, engineering, economics, psychology, political science, public health, medicine, physics, sociology, geography, and demography.

More recently, we have seen the development of a multilevel network analysis that has provided the link with multilevel analysis. While network theory generally analyzes one given level, the newer approach examines not only the networks that exist at different levels but also the links between levels. It has led to major extensions of existing models of social structures, with networks as dependent variables. One class of models tries to "reveal the interdependencies among the micro-, macro-, and meso-level networks," (Wang et al. 2013, p. 97), the meso-level being defined here as between nodes of two adjacent models". They generalize graph models for multiple networks. A second category of models accommodate "multiple partially exchangeable networks for parameter estimation, as well as pools information for multiple networks to assess treatment and covariate effects" (Sweet et al. 2013, p. 298). Often called hierarchical network models, they are a generalization of the multilevel models described in the previous section. A third type of model "partition[s] the units at all levels into groups by taking all available information into account and determining the ties among these groups." (Žiberna 2014, p. 50). It is a generalization of classical blockmodeling developed for relationships between individuals.

Like the multilevel approach, many of these models use Bayesian estimators—which offer many algorithmic advantages, particularly for non-nested data structures—and Markov Chain Monte Carlo (MCMC) algorithms. They use the frequentist paradigm as well as the epistemic paradigm, producing more general estimators of the empirical Bayes estimator type (Greenland 2000).

This approach requires surveys capable of capturing different levels simultaneously. For example, a survey on relationship networks captured the family, occupational relationships, friendly relationships, and memberships in various organizations and groups for individuals living in a rural area (Courgeau 1972). A network analysis of this survey (Forsé 1981) used a complete diagram of acquaintance networks to construct "sociability" groups distinguished by social and demographic characteristics. Other examples of more restricted networks include biomedical research networks and a secluded monastery (White et al. 1976), as well as larger networks such as those found in the Stanford Large Network Data Set Collection, which comprises social networks, citation networks, collaborative networks, Internet networks, and so on (Leskovec et al. 2009).

What new problems will this approach now encounter? They are now mainly technical ones.

An initial problem is the difficulty of capturing the ties between individuals or in collecting available data on the subject. To begin with, the ties will never be exhaustive, and the many reasons for their limitation complicate their study. Very often, such surveys can capture only a limited number of ties, and the number may vary substantially between surveys. There is also an ambiguity about how to designate ties: the term "best friends" may have a different meaning from "friends most frequently met" or "most trustworthy person." While a survey may ask a respondent for information on different kinds of relationship networks such as family, friends or work colleagues, an existing data collection, such as people linked on Facebook, will not allow this distinction. Some respondents may even report more connections with popular, attractive or powerful persons than they actually maintain.

A second problem is that network clusters are generally created by the researcher rather than pre-existing. The method used to create them requires many decisions that are hard to take in a truly scientific way. As Žiberna (2014, p. 50) noted:

> In conceptual terms, the main disadvantages are that there are no clear guidelines concerning what are the appropriate restrictions for ties between levels and what are appropriate weights for different parts of multi-relational networks, that is for level specific one-mode networks and for the two-mode networks.

While this statement relates more specifically to Žiberna's blockmodeling approach, it also applies to the more general multilevel network approach. In both cases, the researcher must decide whether to include or exclude people from a given network, merge or divide network clusters, and so on. Such decisions are needed to allow statistical analysis later on.

A third problem is the difficulty of introducing individual or network characteristics in the study. This can be done only by using the hierarchical network model. But, even in this case, few data sets give measures of the effects of characteristics or measures of network structure (Sweet et al. 2013). These characteristics may be individual, network-specific, tie-specific, or a combination of the three.

A fourth problem concerns the introduction of time in these studies. Here as well, very few surveys enable us to observe changes in networks over time. Some multi-wave surveys capture network structure at different times. Lazega et al. (2011) used a three-wave survey to show that an organization's structure remains the same regardless of its membership's turnover. However, we need more detailed surveys on the changes in networks over a long, continuous period in order to study the changes that may occur, up to and including the end of a network.

We can conclude this examination of the problems and challenges of multilevel network analysis with the following quotation (Lazega and Snijders 2016, p. 260):

> Among the most difficult [challenges], we find combining network dynamics and multilevel analysis by providing statistical approaches to how changes at each level of collective agency drive the evolution of changes at other levels of collective agency. In all these domains, much remains to be done.

Arguably, these problems should be seen more as a challenge for future research than as insuperable difficulties.

3 Toward a Synthesis

After describing and assessing four approaches—with different goals—to understanding human behavior, let us now see if we can give a more synthetic view of them. We begin by examining two basic concepts without which no social science would be possible.

The first concept is the creation of an abstract fictitious individual, whom we can call a statistical individual as distinct from an observed individual. While for Aristotle (around 350 BC, Book I, Part 2, 1356b) "individual cases are so infinitely

various that no systematic knowledge of them is possible," Graunt (1662) was the first to introduce the possibility of a population science by setting aside the observed individual—too complex for study—and using statistics on a small number of characteristics, yielding a statistical individual. In Courgeau's words (2012, p. 197):

> Under this scenario, two observed individuals, with identical characteristics, will certainly have different chances of experiencing a given event, for they will have an infinity of other characteristics that can influence the outcome. By contrast, two statistical individuals, seen as units of a repeated random draw, subjected to the same sampling conditions and possessing the same characteristics, will have the same probability of experiencing the event.

The statistical individual having been thus defined, the key assumption that allows the use of probability theory here is that of exchangeability (de Finetti 1937), which we can formulate simply as follows: n trials will be said to be exchangeable if the joint probability distribution is invariant for all permutations of the n units. Social scientists routinely use exchangeability for the residuals obtained, taking into account the various characteristics included in their analysis. In so doing, they distinguish the statistical individual from the observed individual.

The second concept is the statistical network, as distinct from observed networks. It was introduced more recently, by Coleman (1958). While observed networks may be as diverse as the different kinds of ties existing between individuals—consistently with Aristotle's comment on individuals—statistical networks are obtained from an analysis of ties between individuals along with and the choice of criteria to circumscribe the ties. Here as well, the basic assumption that allows the use of probability theory is that of the exchangeability of networks and the individuals that compose them, taking into account the characteristics introduced at each level.

It is interesting to compare these two concepts with the contexts proposed by Billari (2015) to explain population change, namely, the micro- and macro-level contexts. In fact, Billari clearly recognizes the abstract concept of statistical individual—the same concept proposed here—as the basis of the micro-level context. For the macro-level context, however, he only proposes to examine how "population patterns re-emerge from action and interaction of individuals" (p. S13), without fully recognizing the abstract concept underlying the interactions: the statistical network, which makes it possible to flesh out this macro-analysis. As an example, we have already seen how multilevel analysis reconciles the macro- and micro-level results.

Once these two main concepts are defined, we can see that the study of time between events and the study of event sequences are directly connected to the same concept of statistical individual. Despite the earlier-noted difference in their approaches to this individual, we can regard them as two complementary ways to study the individual. Furthermore, some recent studies combine the advantages of the two approaches by modeling "interaction between macro-institutional configurations and individual life-course trajectories" (Studer et al. 2018). The definition of event-history analysis, already given on the first subsection of the second section of this chapter easily extends to sequence analysis. The itinerary is followed event after event in the first approach and with more complex sequences of events in the second approach.

Similarly, we can see that the contextual, multilevel, and multilevel network approaches are simultaneously connected to the same concept of statistical network. They also seem complementary. We can say that contextual and multilevel analysis focuses on attributes of both individuals and levels, whereas network multilevel analysis focuses on relationships combining the different levels. The paradigm offered for the contextual and multilevel approach is "to explain dependent variables by models containing multiple sources of random variation and including explanatory variables defined as aggregate or other higher-order units" (Lazega and Snijders 2016, p. 3). They can easily extend this paradigm to the network-based approach, with the additional specification that it "means analyzing separately, then jointly, several models of collective agency" (p. 4).

Interestingly, multilevel approaches may be seen as complementing event-history analysis by introducing the effects of membership of different levels on individual behavior. Similarly, multilevel network analysis may be seen as complementary to sequence analysis. This proximity may explain why Cornwell (2015) tries to introduce network analysis methods in sequence analysis. However, sequence methods rely mostly on a grouping of statistical individuals determined by personal criteria, while network methods introduce statistical networks from the outset.

The problems encountered when using one of the four approaches above are easily solved by simultaneously examining the statistical individual and the statistical network by means of a more general biographical multilevel network analysis. As noted earlier, such an approach avoids the risk of atomistic or ecological fallacy by using a synthesis of holism and methodological individualism. It also avoids having to choose between Bayesian and frequentist probability through the use of a more general compromise on confidence distributions (Schweder and Hjort 2016), paving the way for a more satisfactory statistical inference. By introducing networks that yield a better understanding of human behavior, it offers solutions to several problems posed by unobserved heterogeneity. It is also likely that a number of problems involved in sequence analysis—such as the choice of metric, cluster analysis, and the question of whether the groups formed actually exist—can be solved by undertaking more complex surveys on social networks. These would enable us to replace theoretical clusters with real networks of individuals linked together by existing social forces. Similarly, the main problems raised by multilevel analysis could be solved more easily by multilevel network analysis, such as the use of a Multilevel Social Influence (MSI) model (Agneessens and Koskinen 2016) to explain the emergence of social capital, and the use of Exponential Random Graph Models (ERGMs) to show that within-level network structures depend on network structures at other levels (Wang et al. 2016).

Lastly, we believe that the problems posed more recently by multilevel network analysis should be seen as a challenge for future research rather than as insuperable difficulties. For example, such an analysis will reach its full potential when truly longitudinal observations of the multiple levels analyzed become available, providing a combination of individual and network event histories. Collecting data and providing valid statistical approaches to solve this problem will be necessary and appear to be a challenge for such a biographical multilevel network analysis.

4 Conclusion

If we define a scientific approach solely by its methods, we inevitably adopt a partial view of the core of the approach. We must now set up a more robust research program for demography and, more generally, the social sciences—a program that converges with the now well—established program of the physical and biological sciences. The source for this program can be traced back to Bacon in 1620 (Bacon et al. 2000, XIX, p. 36):

> There are, and can be, only two ways to investigate and discover truth. The one leaps from senses and particulars to the most general axioms, and from these principles and their settled truth, determines and discovers intermediate axioms; this is the current way. The other elicits axioms from sense and particulars, rising in a gradual and unbroken ascent to arrive at last at the most general axioms; this is the true way, but it has not been tried.

Bacon calls the second approach induction, not in the meaning later given to the term by Hume and his empiricist tradition—i.e., the generalization of observations—but in the sense of the search for the structure of observed phenomena. That is how Galileo, Newton, Graunt, Einstein, Darwin, and others developed their approach to the study of phenomena—whether physical, biological or social.

It is important for the social sciences to begin by observing and measuring social facts, for this measurement, far from being of secondary importance, makes it possible to assess the "potentialities" of a social fact (Courgeau 2013). Next, instead of relying on often arbitrary hypotheses, the modeling of observed phenomena should follow the method recommended by Bacon by analyzing the interactions between the networks created by people and seeking their structure (Franck 2002; Courgeau et al. 2017).

While we can argue that individuals each have an unlimited and unknowable number of characteristics with their own freedom of choice, social science can show that they are born in a given society with its rules and laws, which restrain their freedom, and that they are subject to biological laws, which are the same for all humans. This is what allows the existence of a social science that takes into account a limited number of characters and is based on a set of concepts without which these characters would be inconceivable or impossible (Franck 2002).

We should like to conclude by emphasizing the following point. We have more often viewed the social sciences as a whole to which certain approaches applied and not others. We must now consider that it is not by erasing the boundaries between disciplines that we can improve our knowledge (Franck 1999). The boundaries are real, for each discipline endeavors to analyze different properties of human societies. However, we believe it is possible to construct a new formal object that can explain certain properties of human societies—an object that transcends existing disciplines and allows their synthesis.

Acknowledgements I thank the three anonymous referees for their detailed and thoughtful comments on an earlier draft of the manuscript, and Jonathan Mandelbaum for its English translation.

References

Aalen, O. (1975). *Statistical inference for a family of counting processes*. Ph.D. thesis, Institute of Mathematical Statistics, Copenhagen.

Aalen, O. O., Borgan, Ø., & Gjessing, H. K. (2008). *Survival and event history analysis: A process point of view*. New York: Springer.

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 16*(4), 129–147.

Abbott, A. (1984). Event sequence and event duration: Colligation and measurement. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 17*(4), 192–204.

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21*(1), 93–113.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology. *Sociological Methods & Research, 29*(1), 3–33.

Adler, P. S., & Kwon, S.-W. (2002). Social capital: Prospects for a new concept. *The Academy of Management Review, 27*(1), 17–40.

Aeby, G., Gauthier, J.-A., Gouveia, R., Ramos, V., Wall, K., & Cesnuityte, V. (2017). The impact of coresidence trajectories on personal networks during transition to adulthood: A comparative perspective. In V. Cesnuityte, D. Lück, & E. Widmer (Eds.), *Family continuity and change: Contemporary European perspectives* (pp. 211–242). London: Palgrave Macmillan.

Agneessens, F., & Koskinen, J. (2016). Modeling individual outcomes using a multilevel social influence (MSI) model: Individual versus team effects of trust on job satisfaction in an organisational context. In E. Lazega & T. A. Snijders (Eds.), *Multilevel network analysis for the social sciences: Theory, methods and applications*, (pp. 81–105). Cham: Springer.

Aristotle (350 BC). *Rhetoric*. Translated by W. Rhys Roberts (http://classics.mit.edu/Aristotle/rhetoric.html).

Bacon, F., Jardine, L., & Silverthorne, M. (2000). *The new organon* (Cambridge texts in the history of philosophy). Cambridge: Cambridge University Press.

Billari, F. C. (2015). Integrating macro- and micro-level approaches in the explanation of population change. *Population Studies, 69*(sup1), S11–S20.

Bison, I. (2009). OM matters: The interaction effects between indel and substitution costs. *Methodological Innovations Online, 4*(2), 53–67.

Bison, I. (2014). Sequence as network: An attempt to apply network analysis to sequence analysis. In P. Blanchard, F. Bülmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 231–248). Heidelberg: Springer.

Bolano, D. (2014). Hidden Markov models: An approach to sequence analysis in population studies. In *Annual Meeting of the Population Association of America, Boston*, 1–3 May 2014.

Bretagnolle, & Huber-Carol (1988). Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics, 15*(2), 125–138.

Byrne, D., & Uprichard, E. (2012). Introduction. In D. Byrne & E. Uprichard (Eds.), *Cluster analysis* (Vol. 2, pp. vii–xii). London: Sage Publication Ltd.

Coleman, J. (1958). Relational analysis: The study of social organizations with survey methods. *Human Organization, 17*(4), 28–36.

Cornwell, B. (2015). *Social sequence analysis: Methods and applications*. Cambridge: Cambridge University Press.

Courgeau, D. (1972). Les réseaux de relations entre personnes. etude d'un milieu rural. *Population, 27*(4–5), 641–683.

Courgeau, D. (1999). L'enquête "triple biographie: Familiale, professionnelle et migratoire". In G. de réflexion sur l'approche biographique (Ed.), *Biographies d'enquêtes* (pp. 59–74). Paris: INED.

Courgeau, D. (2007). *Multilevel synthesis: From the group to the individual*. Dordrecht: Springer.

Courgeau, D. (2012). *Probability and social science: Methodological relationships between the two approaches*. Dordrecht: Springer.

Courgeau, D. (2013). La mesure dans les sciences de la population. *Cahiers philosophiques, 135*(4), 51–74.

Courgeau, D., & Lelièvre, E. (1992). *Event history analysis in demography*. Oxford: Clarendon Press.

Courgeau, D., & Lelièvre, E. (1997). Changing paradigm in demography. *Population, 9*, 1–10.

Courgeau, D., Bijak, J., Franck, R., & Silverman, E. (2017). Model-based demography: Towards a research agenda. In A. Grow & J. Van Bavel (Eds.), *Agent-based modelling in population studies: Concepts, methods, and applications* (pp. 29–51). Cham: Springer.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187–220.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. London: Chapman and Hall.

de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Pointcarré, 7*(1), 1–68.

Doob, J. L. (1953). *Stochastic processes*. New York: Wiley.

Draper, D. (2008). Bayesian multilevel analysis and MCMC. In J. Deleeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 77–140). New York: Springer.

Durkheim, É. (1895). *Les règles de la méthode sociologique*. Paris: Alcan.

Everitt, B. S. (1979). Unresolved problems in cluster analysis. *Biometrics, 35*(1), 169–181.

Forsé, M. (1981). Les réseaux de sociabilité dans un village. *Population, 36*(6), 1141–1162.

Franck, R. (1995). Mosaïques, machines, organismes et sociétés. examen métadisciplinaire du réductionnisme. *Revue Philosophique de Louvain, 93*(1–2), 67–81.

Franck, R. (1999). La pluralité des disciplines, l'unité du savoir et les connaissances ordinaires. *Sociologie et sociétés, 31*(1), 129–142.

Franck, R. (Ed.) (2002). *The explanatory power of models: Bridging the gap between empirical and theoretical research in the social sciences*. Boston: Kluwer Academic.

Freeman, L. C. (1989). Social networks and the structure of experiment. In L. C. Freeman, D. R. White, & A. K. Romney (Eds.), *Research methods in social network analysis* (pp. 11–40). Fairfax: George Mason University Press.

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, BC: BookSurge.

Gabadinho, A., & Ritschard, G. (2016). Analysing state sequences with probabilistic suffix trees: The PST R library. *Journal of Statistical Software, 72*(3), 1–39.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.

Graunt, J. (1662). *Natural and political observations mentioned in a following index and made upon the bills of mortality*. London: Tho. Roycroft.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology, 29*(1), 158–167.

Helske, S., Helske, J., & Eerola, M. (2018). Combining sequence analysis and hidden Markov models in the analysis of complex life sequence data. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications* (Life course research and social policies). Berlin: Springer (this volume).

Henderson, C. R., Kempthorne, O., Searle, S. R., & von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics, 15*(2), 192–218.

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon Press.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Lazega, E., Sapulete, S., & Mounier, L. (2011). Structural stability regardless of membership turnover? The added value of blockmodelling in the analysis of network evolution. *Quality & Quantity, 45*(1), 129–144.

Lazega, E., & Snijders, T. A. B. (Eds.) (2016). *Multilevel network analysis for the social sciences: Theory, methods and applications*. Heidelberg: Springer.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2009). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics, 6*(1), 29–123.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*, 707–710.

Levitt, M. (1969). Detailed molecular model for transfer ribonucleic acid. *Nature, 224*(5221), 759–763.

Loriaux, M. (1989). L'analyse contextuelle: Renouveau théorique ou impasse méthodologique? In J. Duchêne, G. Wunsch, & E. Vilquin (Eds.), *Explanation in the social sciences. The search for causes in demography* (Chaire Quetelet, Vol. 1987, pp. 333–368). Louvain-la-Neuve: Ciaco.

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual analysis through the multilevel linear model. *Sociological Methodology, 14*, 72–103.

Moreno, J. L., & Jennings, H. H. (1938). Statistics of social configurations. *Sociometry, 1*(3/4), 342–374.

Robette, N., & Bry, X. (2012). Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 116*(1), 5–24.

Robette, N., Lelièvre, E., & Bry, X. (2012). La transmission des trajectoires d'activité: Telles mères, telles filles? In C. Bonvalet & E. Lelièvre (Eds.), *De la famille à l'entourage* (pp. 395–418). INED.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357.

Schweder, T., & Hjort, N. L. (2016). *Confidence, likelihood, probability: Statistical inference with confidence distributions*. Cambridge: Cambridge University Press.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society, Series A, 179*(2), 481–511.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research, 40*(3), 471–510.

Studer, M., Struffolino, E., & Fasang, A. E. (2018). Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. *Sociological Methodology* (First Published Online).

Sweet, T. M., Thomas, A. C., & Junker, B. W. (2013). Hierarchical network models for education research. *Journal of Educational and Behavioral Statistics, 38*(3), 295–318.

Trussell, J. (1992). Introduction. In J. Trussell & R. Hankinson (Eds.), *International Studies in Demography* (pp. 1–7). Oxford: Clarendon Press.

Trussell, J., & Richards, T. (1985). Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure. In N. B. Tuma (Ed.), *Social and behavioral science series*(Vol. 15, pp. 242–276). San Francisco, CA: Jossey-Bass.

Tryon, R. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Ann Arbor: Edwards brother.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography, 16*(3), 439–454.

Ville, J. (1939). *Etude critique de la notion de collectif*. Paris: Gauthier-Villars.

Wang, P., Robins, G., & Matous, P. (2016). Multilevel network analysis using ERGM and its extension. In E. Lazega & T. A. Snijders (Eds.), *Multilevel network analysis for the social sciences* (pp. 125–143). Cham: Springer.

Wang, P., Robins, G., Pattison, P., & Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks, 35*(1), 96–115.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

White, H. C., Boorman, S. A., & Breiger, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology, 81*(4), 730–780.
Wu, L. L. (2000). Some comments on "Sequence analysis and optimal matching methods in sociology: Review and prospect". *Sociological Methods & Research, 29*(1), 41–64.
Žiberna, A. (2014). Blockmodeling of multilevel networks. *Social Networks, 39*, 46–61.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Case Studies of Combining Sequence Analysis and Modelling

Mervi Eerola

1 Introduction

The recent two decades have shown that sequence analysis is a valuable tool for life course analysis. While the significance of the past and future is of fundamental importance in event-history analysis, sequence analysis, which in its basic form is ignorant to this distinction, seems to highlight the diversity of life patterns in a way that cannot be achieved with traditional statistical modelling. Several improvements and contributions either to the dissimilarity measures or to the cost matrix have been suggested to the original version. However, the question still remains whether the colourful figures of individual index plots or state distribution plots merely pose interesting hypotheses and further questions rather than provide analytic answers to the causes of life course differences. Therefore, current consensus in this research area seems to emphasize combining the benefits of both approaches. In this methodological paper, I shall present three case studies of life course analysis in which the clustering of the sequences has been combined, or contrasted, with modelling. Two of the studies are already published and one is an ongoing project. The results and complete versions can be found in the References. In the Discussion, the experiences of using both methods are compared in more detail and their role in life course analysis is evaluated.

M. Eerola (✉)
Centre of Statistics, University of Turku, Turku, Finland
e-mail: mervi.eerola@utu.fi

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_3

2 Case Study 1: Prediction of Excess Depressive Symptoms and Life Events

This study investigated how the timing and pattern of certain life events, here partnership formation and steady employment, affect the prediction of parenthood, especially remaining childless, and whether this is associated with excess depressive symptoms in middle age. The participants of the Finnish Jyväskylä Longitudinal Study of Personality and Social Development (JYLS), born in 1959, were from 12 randomly selected second-grade classes in Jyväskylä, Central Finland. They were followed from age 8 to 50. The original sample consisted of 173 girls and 196 boys. A life history calendar (LHC) was used to collect information about partnership status, children, studies, and work, as well as other important life events. The occurrence, timing, and duration of the transitions were recorded annually from age 15 to age 50 during interviews in which 275 participants gave reports based on memory and visual aids provided by the LHC-sheet. Since both partnership formation and career events can have variable patterns in time, and be interpreted as 'states' also, we were interested in investigating what information probabilistic multistate models on one hand, and sequence analysis, on the other hand, can provide about the study question.

2.1 Multistate Models

We considered the life course events in an observation interval $\mathscr{T}$ as a *marked point process* (T, X) specifying the sequence of events by a pair of random variables, the occurrence time T and a mark X identifying the event (e.g. Arjas 1989). Other presentations of multistate models can be found, for example, in Andersen and Keiding (2002).

The discrete time event-specific hazard of event x is the conditional probability

$$p_x(t) = P(\Delta N_x(t) = 1 \mid \mathscr{F}_{t-1}^N)$$

of a jump of type x in time interval $t \in \mathscr{T}$ in the counting process $N_x(t) = \sum_{t \geq 1} 1(T \leq t, X = x)$, given its internal history $\mathscr{F}_{t-1}^N$ generated by the points and marks until time $t - 1$. We denote the extended history by $\mathscr{H}_t = \mathscr{F}_t^N \vee \mathbf{Z}_0$, where $\mathbf{Z}_0$ are covariates fixed at time $t = 0$ already. The crude hazard that any event happens in the interval t is the sum over event-specific hazards $p(t) = \sum_x p_x(t)$.

While the hazard of an event gives a very short-term prediction of the life course, *prediction probabilities* associated with the marked point process give a long-term prediction of some random event for the *whole observed* interval in a life course (e.g. Eerola 1994; Putter et al. 2007; Eerola and Helske 2016). They are functions of event-specific hazards but provide more comparable results with sequence analysis than simple hazard analysis.

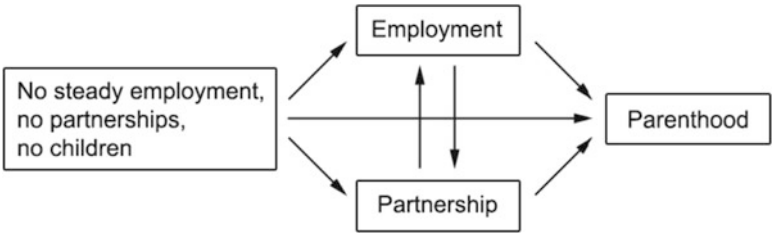

Fig. 1 A schematic model of multistate model for the JYLS data. (Source: Eerola and Helske 2016, reprinted by permission of SAGE publications)

In a multistate model the state-space can increase rapidly, so we only considered the first occurrences of partnership formation (P), child births (C), and steady employment (W) for each of which we specified event-specific hazards (that is, for $X = W$, P or C). For example, the hazard of entering steady employment when the other events have not yet occurred, is in the general form

$$p_W(t) = P(T_W = t \mid T_W \geq t, T_P \geq t, T_C \geq t)$$

where T_W is the time (age) of first steady employment, and the other event time variables are defined accordingly. As a statistical model for the discrete time hazard of event x, a piecewise constant logistic model with time-dependent indicator variables for earlier events

$$p_x(t) = (1 + \exp(-\boldsymbol{\beta}' Z(t)))^{-1}$$

was used. The covariate vector $Z(t)$ comprises indicator variables for the events in Fig. 1, as well as piecewise constant indicators for time (age). For example, the covariate $Z_W(t) = 1$ if steady employment was reached at age t and 0 before that.

The prediction probabilities of remaining childless are sums of the probabilities of all paths of remaining childless within the prediction interval when all possible timings of partnership formation and steady employment are considered. The most complicated path results when nothing has yet happened at the prediction time t, the other paths being special cases of it. In particular, when initial partnership (P) and entering working life (W) have already occurred by the prediction time t, the prediction is simply the survival probability (for time points $0 < v \lesseqgtr w < t < u$)

$$P(T_C > u \mid T_W = v, T_P = w, T_C \geq t) = \prod_{s=t+1}^{u} (1 - p_{C|WP}(s \mid v, w)).$$

Fixing the prediction time t, the prediction interval from $t + 1$ to u (the last observation time), or the history, results in different visual representations of the predictions. For example, fixing prediction interval and history, and identifying

t with the occurrence time of a life event, compares *factual* and *counterfactual* predictions of remaining childless, depending on whether the event x in fact occurred at t, or not. Finally, the prediction of excess depressive symptoms if the person remained childless until age 42, given the history of partnership formation and entry to stable employment, is the joint conditional probability (for $15 < t \leq 42$)

$$P(T_C > 42, D(42) > d^* \mid \mathscr{H}_t) = P(D(42) > d^* \mid T_C > 42)\, P(T_C > 42 \mid \mathscr{H}_t)$$

in which $D(42)$ is the score of depressive symptoms at age 42 and d^* is the median score in the study population.

2.2 Sequence Analysis

As a comparison, we performed multidimensional sequence analysis. Pairwise comparison of the original sequences using the Hamming distance resulted in eight clusters. They differed mostly in terms of timing of partnership and parenthood, and to a lesser extent in terms of the length of education. To associate these results with depression in middle age, we used the individual cluster membership indicator as a covariate in a logistic regression predicting higher than median depression score d^*, as before. This covariate was used as a 'proxy' variable for parenthood, partnership and employment history. For a generic individual, the model was

$$\text{logit}(P(D(42) > d^* \mid c)) = \alpha + \beta Z(c)$$

with $Z(c) = 1$, if the individual was a member of cluster c. Only the most deviant cluster ("singles or late family") had significantly higher odds of excess depressive symptoms than the other clusters. This supports our results with the prediction probabilities but is less informative in terms of separating individual effects or timing effect in general. More results of the study can be found in Eerola and Helske (2016).

3 Case Study 2: Antecedents and Consequences of Transitional Pathways to Adulthood

This study from developmental psychology linked two types of longitudinal data: the sequences of young people's transitions to adulthood and longitudinal data of psychological resources. The study investigated the extent to which university students' depressive symptoms, and the strategies they deploy in achievement and social situations (Nurmi et al. 1995) at the beginning of university studies

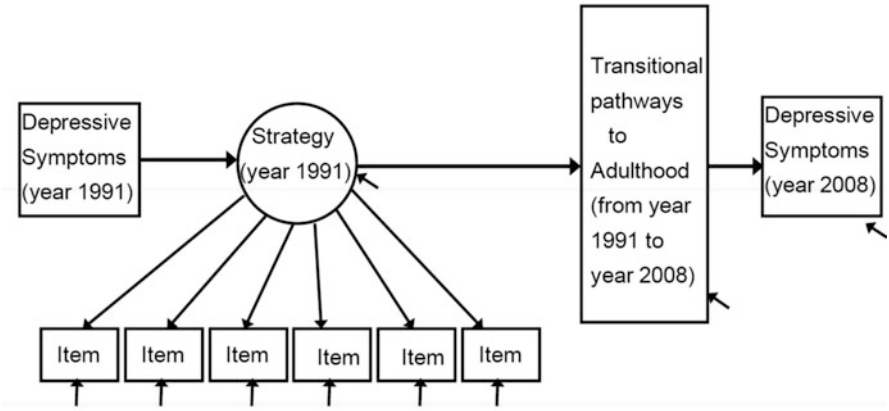

Fig. 2 A schematic model of the structural equation model for the HELS data. (Source: Salmela-Aro et al. 2014, reprinted by permission of Springer)

would predict their transitional pathways, and the extent to which the pathways contribute to depressive symptoms later in adulthood. The study was part of the Helsinki Longitudinal Student Study (HELS study). The participants were 182 undergraduates who started their studies at the University of Helsinki in 1991 and were born in or around 1970. A life history calendar was completed in 2008, retrospectively reporting on key life events during the years 1991–2008 (residence, partnership, parenthood and career).

In a previous study (Salmela-Aro et al. 2011), six clusters (transitional pathways) were identified by sequence analysis. Figure 2 shows a schematic model of the study. A strategy-specific *structural equation model* combined the submodels for achievement and social strategies with depressive symptoms at the start of follow-up, with the model for transitional pathways, and finally with the model for depressive symptoms at the end of follow-up.

3.1 Model for Strategies Accounting for Depressive Symptoms

We defined hierarchical factor models for social and achievement strategies (social optimism, social withdrawal, achievement optimism or task-irrelevant behaviour) as

$$y_{ijs} = \lambda_i \eta_{js} + \epsilon_{ijs}, s = 1, \ldots, 4, j = 1, \ldots, 182$$
$$\eta_{sj} = \eta_{0s} + \gamma_s x_j^{pre} + \zeta_{sj}$$

where y_{ijs} is the ith item measuring a particular strategy s and η_{sj} is the factor corresponding to that strategy for individual j, λ_i is the factor loading of item i, and ϵ_{ijs} is the unique factor of item i for strategy s. The hierarchical or

multilevel structure of the factor model (Muthén 1994; Goldstein 2011) assumes that depressive symptoms at the start of the follow-up may affect the strategies, and this is accomplished in the model by allowing each strategy factor to depend on the individual's depression score. The parameter η_{0s} is the mean level of the factor s, γ_s is the regression coefficient of depressive symptoms score x_j^{pre} at the start of studies in 1991, and ζ_{sj} refers to the individual-specific deviation from the mean factor level η_{0s}. In this model, only the items of the strategy measures and the depression scores are observable. The factors η_{sj} and the error variables ϵ_{ijs} for items and ζ_{sj} for factors are assumed zero-mean normal random variables.

3.2 Model for Transitional Pathways Accounting for Strategies

The predictive value of social and achievement strategies for the probability of following a particular pathway was studied by binary or multinomial logistic regression models with the membership indicator of a particular transitional pathway as the dependent variable. Achievement and social strategies were included as separate predictors. Since the most distinguishing factor between the six pathways was that some life events did not occur at all, or that their timing was exceptionally late, we estimated a joint model for the pathways that we called postponed (singles with slow career who never lived in a partnership during the follow-up, and slow starters, whose transitions were postponed in general). They were combined to a *postponed group* (n = 56). The remaining four pathways (fast starters, fast partnership and late parenthood, career and family, and career and unsteady partnerships) were combined to a *non-postponed group* (n = 126).

The model for the log-odds of belonging to the postponed vs. non-postponed pathway which accounts for social and achievement strategies was of the form

$$\theta_{sj}^{post} = \text{logit}(P(y_j^{post} = 1 \mid \eta_{sj})) = \alpha_s + \beta_s \eta_{sj}$$

Here y_j^{post} is the membership indicator of postponed pathway, β_s the regression coefficient of factor η_{sj} of strategy s for individual j.

3.3 Model for Depressive Symptoms When Accounting for Pathways

The model for the expected level of depressive symptoms in 2008, which accounts for pathway and indirectly also the initial level of depression and the strategies, is for each strategy

$$\mu_{sj} = \mu_0 + \delta_{1s} \exp(\theta_{sj}^{post}).$$

The parameter μ_0 is the mean level of depressive symptoms in 2008, $exp(\theta_{sj}^{post})$ the individual- and strategy-specific odds of following the postponed pathway, and δ_{1s} are the direct effects of pathway, containing also the *indirect* effects of initial-level depressive symptoms and strategies. This model was contrasted with the model of *direct* effect of the initial level of depressive symptoms

$$\mu_j = \mu_0 + \delta_2 x_j^{pre}$$

on the level of depression in 2008, where the parameter δ_2 is the direct effect of the initial level of depression symptoms in 1991, without considering the effect of strategies or pathways. This 18-year follow-up showed that depressive symptoms at the beginning of studies were associated with pessimistic and avoidant strategies in both achievement and social situations, which further predicted postponed pathways later on. The transitional pathways also contributed subsequent changes in depressive symptoms. More results of this study can be found in Salmela-Aro et al. (2014).

4 Case Study 3: Pathways to Social Exclusion

The so called NEET problem (Not in education, not in employment or training) has in many countries initiated special government policy acts to prevent young people, especially young men, from ending up in social exclusion. By social exclusion is usually meant a combination of problems such as unemployment, unfinished education, low incomes, alcohol problems, crime, bad health and unstable family conditions. These problems are linked and mutually reinforcing, and can create a vicious cycle in a person's life course. Cross-sectional studies are not helpful when trying to understand the dynamics of this process.

In this ongoing study, we are in particular interested in originating events or factors of risk accumulation and potential turning points in a young person's trajectory. We use the Finnish National Birth Cohort 1987 (around 60,000 individuals) data from the years 2005–2012 when the members were 18 to 25 years old. The cohort can be combined with all official registers, from which we restrict to data on unemployment, education and use of social benefits, episodes in mental health care and reimbursement of medicine expenses for mental illness, inpatient days due to intoxicant abuse and notifications in crime register. As usual with register data, it is important to analyse carefully which outcomes are results of the social benefit system itself to prevent from meaningless modelling.

4.1 Sequence Analysis

Sequence analysis is here used to find the most vulnerable individuals for the follow-up. Two clusters out of 12 (together around 10% or 6000 individuals) having the most fragmentary trajectories in terms of the main activity classification ("Employed", "Unemployed", "Studying", "Other") are chosen for further analysis. Several approaches can be suggested to analyse underlying lifetime periods characterised by the accumulation of risk factors.

4.2 Risk Pattern Analysis

Let $y_a = (y_{1a}, \ldots, y_{6a})'$ be individual's observed *risk pattern* at age a where $y_1, \ldots, y_6$ are indicators of the measured risk factors (outside of work force, lowest educational attainment, living on social benefits, mental health care or medication, intoxicant abuse and criminal record, respectively). This amounts to $M = 2^6$ possible binary risk patterns in each follow-up year.

We assume that η_a is a latent state with values $s \in S$ representing underlying situational characteristics of a young person at age a. As usual in hierarchical modelling, we assume that the observed indicators $\{y_{ia}\}$ are conditionally independent given η_a at each a. This is a *latent transition model* (e.g. Collins and Lanza 2010) of observed risk patterns given the dynamics of the underlying latent states.

Denote the conditional probability of risk factor i at age a by $P(Y_{ia} = 1 \mid \eta_a = s) = p_a(i \mid s)$ and the transition probability to latent state s at age a by $P(\eta_a = s \mid \eta_{a-1} = r) = q_a(s \mid r), s, r \in S$, given that the previous latent state at age $a - 1$ was r. For a generic individual, the (marginal) probability of risk patterns in the follow-up is then

$$
\begin{aligned}
P(Y = y) &= \sum_{s_a} \prod_{a=18}^{25} \{P(\eta_a = s \mid \eta_{a-1} = r) P(Y_a = m \mid \eta_a = s)\} \\
&= \sum_{s_a} \pi_s \prod_{a=19}^{25} q_a(s \mid r) \prod_{a=18}^{25} [\prod_{i=1}^{6} p_a(i \mid s)^{y_{ia}} (1 - p_a(i \mid s))^{1-y_{ia}}] \\
&= \sum_{s_a} \pi_s \prod_{a=19}^{25} q_a(s \mid r) \prod_{a=18}^{25} [\prod_{m=1}^{M} p_a(m \mid s)^{1(y_a=m)}]
\end{aligned}
$$

when summing over all possible latent states at ages $a = 18, \ldots, 25$. π_s is the initial probability of latent state s at age $a = 18$ and $1(y_a = m) = 1$ if the observed risk pattern at age a is m.

If we, in turn, assume that $\eta_a = \eta$, where η is an inherent tendency or 'trait', predisposing to marginalisation, which can be partially observed in terms of the accumulating risk factors, we would consider it as a fixed continuous latent variable. If the probability of the observed risk factors change by age, this is a dynamic *latent trait model* (e.g. Lord and Novick 1968).

A *hidden Markov model* (e.g. Rabiner 1989) has a similar probability structure but the observed states y would then be the main activity groups "Employed", "Unemployed", "Studying", "Other". To include the risk factors, we can either enlarge the state space by combining the statuses of the risk factors with the main activity groups resulting in states such as "Other/LowEdu/MHealth/Drugs/Crime" which would resemble multidimensional sequence analysis. A more natural way is to define the transition rates or transition probabilities between the four main activity groups with time-dependent covariates as in Case 1.

4.3 Predictions of Positive Trajectories

Since there already exists several studies on the prevalence of risk factors for NEET, yet another approach is to estimate predictions of *no* marginalisation, that is, predictions of integration into the labour market or in educational trajectories by age 26 when *avoiding* a particular risk factor along the developmental pathway while experiencing others. As in Case study 1, such "What if" -analyses compare two probabilities: that of an individual, initially at high risk, but who avoids a particular risk factor (mental health problems, criminal records, living on social benefits, lowest educational level) *at least until age a*, with the probability of not avoiding it, given other risk factors. Since we are interested in the effects of risk factors on the *positive outcome* (integration into labour market or education), it is the difference of these probabilities that allows us to evaluate the effect of timing on the positive trajectory.

5 Discussion

This paper has illustrated three case studies which combine sequence analysis and probabilistic modelling in life course analysis. In the first, prediction probabilities for individual's entire observed life trajectory were estimated to find out how the timing of certain life events affects the prediction of an outcome. Since all of the life events could repeat in time, sequence analysis provided a much more detailed picture of the life patterns while multistate models restricted to the first events only. Nevertheless, for the analytic and 'causal-like' questions posed in the study, sequence analysis turned out to be less helpful.

In the second case study, multivariate psychological measures before and after the pathway analysis were combined into a larger structural equation model. Embedding the clustering results from sequence analysis in it allowed for including the multidimensional information about the trajectories in a way that could hardly be achieved with a few covariates. However, this information is often weak because individuals may in fact have characteristics of several overlapping clusters. Since clustering is based on the matrix of pairwise distances, and not on the individual sequences any more, explanatory models based on membership indicators can be rather unspecific. The cluster characteristics are not then representative to all its members, and sensitivity analysis with, for example, MDS plots can be useful. Lundevaller et al. (2018) used the combined SA states directly as covariates in Cox regression models but this approach would require a larger dataset than they had. Rossignon et al. (2018) propose to add the individual trajectory as a time-dependent covariate which resembles our logistic risk models with time-dependent indicators for the events of the multistate model in Case 1.

The third case study uses sequence analysis to find the most plausible cases for the follow-up from a large register data while leaving the rest of the cohort as a reference, if needed. Initial clustering with SA allows again multidimensional and time-dependent criteria to extract out a subgroup for further analysis. In Helske et al. (2018) clustering with SA was used to get initial values for the latent states of a hidden Markov model.

Preserving diversity in life-histories in the preliminary stage usually means that we need dimension reduction in later stage. In this paper, we have used latent variable (hierarchical) modelling in various ways for this purpose. Latent variables may have different interpretations: individual-specific tendency to respond (latent response models), deviation from group-specific mean behaviour (mixed models), frailty (excess risk for an event in survival models), or an underlying hypothetical construct or trait which can be observed as a pattern of multiple items (latent trait models, IRT models etc.). In latent transition models or hidden Markov models the latent structure is dynamic and, especially for multichannel problems, the interpretation of the latent states becomes sometimes quite difficult.

Sequence analysis is undoubtedly most effective in grasping the 'big picture' of the state dynamics in population-level studies. It provides an easily understandable visual representation (proportions of states by time) of multidimensional longitudinal data with minimal simplification of the original data. The figures lead to questions as to why these observed differences between population groups exist. This often requires individual-level information, which unfortunately, apart from the membership indicator, is lost in clustering. More specific causal inquiries, such as "How would the trajectory of an individual of certain type most likely be, had he/she faced (or avoided) a particular life event, which he/she didn't, given that

everything else had been the same?" are only possible in probabilistic modelling. In a more general sense, however, combining SA with statistical modelling allows quantitative comparative analysis of observed differences in terms of explanatory covariates, and evaluation of their significance.

Acknowledgements The HELS study led by Katariina Salmela-Aro and Jari-Erik Nurmi and the JYLS study initiated by Lea Pulkkinen and led by Katja Kokko are acknowledged for the permission to use the data in the Case studies. The referees are acknowledged for their valuable comments.

References

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research, 11*(2), 91.

Arjas, E. (1989). Survival models and martingale dynamics. *Scandinavian Journal of Statistics*, 177–225.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Probability and statistics). New York: Wiley.

Eerola, M. (1994). *Probabilistic causality in longitudinal studies* (Lecture notes in statistics, Vol. 92). New York: Springer.

Eerola, M., & Helske, S. (2016). Statistical analysis of life history calendar data. *Statistical Methods in Medical Research, 25*(2), 571–597.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). Hoboken: Wiley.

Helske, S., Helske, J., & Eerola, M. (2018). Analysing complex life sequence data with hidden Markov modelling. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Lundevaller, E., Vikström, L., & Haage, H. (2018). Modelling mortality using life trajectories of disabled and non-disabled individuals in 19th-century Sweden. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*(3), 376–398.

Nurmi, J.-E., Salmela-Aro, K., & Haavisto, T. (1995). The strategy and attribution questionnaire: Psychometric properties. *European Journal of Psychological Assessment, 11*, 108–121.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in medicine, 26*(11), 2389–2430.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Rossignon, F., Studer, M., Gauthier, J.-A., & Goff, J.-M. L. (2018). Sequence history analysis (SHA): Estimating the effect of past trajectories on an upcoming event. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

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., Kiuru, N., Nurmi, J.-E., & Eerola, M. (2014). Antecedents and consequences of transitional pathways to adulthood among university students: 18-year longitudinal study. *Journal of Adult Development, 21*(1), 48–58.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part II
Sequence Analysis and Event History Analysis

Glass Ceilings, Glass Escalators and Revolving Doors

Comparing Gendered Occupational Trajectories and the Upward Mobility of Men and Women in West Germany

Lydia Malin and Ramsey Wise

1 Introduction

Several studies have demonstrated a female disadvantage with regards to upward occupational mobility due to structural barriers commonly referred to as "glass ceilings" (Maume 1999a; Reskin 1993; Cotter et al. 2001). These barriers are often attributed to prejudice based on gender stereotypes of social roles (e.g. Eagly 2003; Eagly and Karau 2002) as well as discrimination and stigmatization, particularly of mothers (Aisenbrey et al. 2009; Benard and Correll 2010; Budig et al. 2012; England 2005; Gangl and Ziefle 2009). In addition to studies of a female disadvantage in male-typical occupations, Williams (1992) demonstrated men to experience more career opportunities for promotion compared to women in female-typical occupations (i.e. "the glass escalator" effect).

We provide a holistic description of how gender effects on upward occupational mobility vary by gender-typical occupations.[1] To this end, much of the empirical research concerning gender differences in career advancement has focused on either

[1]The gender-typicality of occupations is defined as follows: occupations with more than 70% of female employees are defined as female-typical, occupations with 30 up to 70% women as mixed, and occupations with less than 30% of female employees subject to social insurance contributions as male-typical or henceforth referred to as female, mixed and male.

L. Malin (✉)
University of Cologne, Cologne, Germany
e-mail: malin@iwkoeln.de

R. Wise
University of Bremen, Bremen, Germany
e-mail: rwise@bigsss.uni-bremen.de

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_4

the American (e.g. Maume 1999a; Budig 2002) or Scandinavian context (e.g. Hultin 2003). For Germany, there are many studies on gendered occupational careers (for the motherhood penalty in downward occupational mobility see e.g. Aisenbrey et al. 2009, for gender-pay gap see e.g. Brückner 2004, for gender inequalities in occupational prestige see e.g. Härkönen et al. 2016, or Manzoni et al. 2014), on the importance of partner resources for occupational promotion (Bröckel et al. 2015) or the gender pay gap in managerial positions (Busch and Holst 2009; Holst 2006).

Only one study controls for gender-typical occupational differences. In analyzing the gender gap in attaining a first management position, Ochsenfeld (2012) uses field of study as measurement of gender-typicality of occupation. However, this study only considers access into leadership, but does not consider a potential revolving door mechanism, whereby access to leadership position may not guarantee remaining in this position. Similarly, Dämmrich and Blossfeld (2017) recently investigated a female disadvantage in holding a supervisory position from a country comparative perspective. For Germany they found that women working in male occupations do not significantly differ from men in holding supervisory positions. Although they accounted for horizontal gender segregation, we contribute to the literature by taking into account two dimensions of a potential male advantage in upward occupational mobility: (1) accessibility and (2) the likelihood to stay in or to leave a leadership position.

Subsequently, we investigate to what extent glass ceiling and glass escalator effects are indeed evident in West Germany. More precisely, we ask: (1) Do men demonstrate an advantage regarding access to and staying in leadership? (2) To what extent does occupational segregation explain gender differences in upward mobility? (3) Do gender effects vary across occupations?

To answer these research questions, the West German case is of special interest. Despite recent changes in work family policies in Germany that are work-family oriented, our observation time is better reflected in the long-standing tradition of the male breadwinner and female caregiver household division of labor that has reinforced gender norms over time (Trappe et al. 2015). Moreover, this dynamic has been further strengthened by strong horizontal sex segregation, whereby women typically belong to different occupations than men (Jacob et al. 2013).

Previous research has linked this selection process of men and women into gender-typical jobs to explain gender differences in vertical sex segregation (Dämmrich and Blossfeld 2017; Charles 2003). Others have investigated whether men and women are more advantaged in gender-typical or gender-atypical occupations. Some have found evidence of a "glass ceiling" effect for women in male occupations (Reskin and Roos 1990), but a "glass escalator" effect for men in female occupations (Williams 1992; Maume 1999b; Cotter et al. 2001). To this end, we aim to demonstrate to what extent these gender differences are attributed to horizontal sex segregation in West Germany.

Section 2 presents our hypotheses, which are derived from theory and empirical evidence. In Sect. 3, we discuss our sample using data from the German National Educational Panel Study (NEPS). As this provides monthly employment histories, we use discrete-time hazard models to estimate the influence of gender, gender-

typical occupations and the interaction of both on the probability to enter and to stay in leadership positions. The results of the analyses are provided in Sect. 4. Section 5 discusses results and avenues for further research.

2 Theoretical Considerations and Hypotheses

The expected male advantage regarding upward occupational mobility denotes two dimensions: (1) a higher probability to enter leadership positions and (2) a higher probability to stay in leadership positions. This assumption receives support from several theoretical approaches discussed in more detail in the following sections. As we aim to also disentangle the main and interaction effects between gender and occupational gender composition, we have organized theoretical considerations by (1) gender effects, (2) gender compositional effects and (3) how gender effects vary across female, mixed and male occupations.

2.1 Gender and Upward Occupational Mobility

Despite some improvements in female educational attainment and labor market participation in younger cohorts, women often fail to attain leadership positions, which are dominated by men (e.g. Eagly 2003). Several theories have been put forward to explain the well-documented male advantage in upward mobility. For example, the "glass ceiling" effect refers to structural barriers that women face when rising up the career ladder. Consequently, the male advantage is stronger at the top of the status hierarchy than at lower levels (Cotter et al. 2001).

Albeit a highly complex phenomenon, many sociologists have emphasized how gender norms contribute to prejudice against women with regards to obtaining promotions during the career (e.g. Ridgeway 2001; Ridgeway and Correll 2004; Williams 1992). For example, "role congruity theory" argues that women hold fewer leadership positions because these positions are typically associated with characteristics attributed to men (Eagly and Karau 2002). The perceived incongruity between traditional female role characteristics and leadership roles thus stigmatizes women as less appropriate for leadership. Eagly and Karau (2002) further observed that women exhibiting male characteristics are also stigmatized and devalued in comparison to their male counterparts (England et al. 1994; Ridgeway 2001), despite being more congruent with leadership characteristics. Therefore, the male advantage is not only observed when entering leadership, but also over the occupational trajectories of men and women. Based on these theoretical considerations, we hypothesize:

- H1a: *Men are more likely to enter a leadership position compared to women, irrespective of the gender composition of the respective occupation held.*
- H1b: *Men are less likely to drop out of leadership positions compared to women, irrespective of the gender composition of the respective occupation held.*

2.2 Gender Composition and Upward Occupational Mobility

In additional to gender effects, there are several studies that attribute male advantages in the labor market to occupational sex segregation (e.g. Charles 2003; Ko et al. 2015 for the US; Hultin 2003 for Sweden; Busch 2013 for Germany). However, most of these studies do not take into account the role of labor market segmentation or provide theoretical arguments for differences in the institutional set-up structuring upward mobility in female and male occupations. Because men and women often (self-)select employment in gender-typical occupations, we argue that much of the gender effect can be explained by the different work arrangements of these occupations. For this reason, we are interested in how gender composition influence leadership opportunities regardless of gender.

As an important aspect of mobility research, labor market sociologists have long debated the relationship between labor market segmentation and opportunities for promotion (Edwards 1979; Sengenberger 1987). The growth of large firms is argued to have contributed to labor market segmentation, as hierarchical career ladders were created as a means to secure employee commitment, control the workplace and to reduce sunk costs caused by worker turnover (Farkas and England 1988; Sørensen and Kalleberg 1981). These characteristics, however, largely describe the career trajectories in male occupations.

In contrast, female occupations tend to be primarily aligned either with low-skilled, service sector or semi- and high-skilled, professional occupations. The first type of female occupations exhibits the "revolving doors analogy" comprising low-wage, dead-end jobs that do not provide opportunities for career advancement (Jacobs 1989; Charles and Grusky 2004; Williams 2013). The second type of female occupations is more closely associated with occupational-specific professions (e.g. teaching professions or health professions).

As upward mobility opportunities are highly differentiated across occupations, we expect that female occupations offer fewer opportunities for promotion than male occupations irrespective of the employees' gender. Subsequently, much of the so-called gender effect may actually reflect the selection of women into female occupations that do not offer many opportunities for promotion. Therefore, we hypothesize:

- H2a: *Men and women are more likely to hold a leadership position in male-typical occupations and less likely in female-typical ones compared to mixed occupations.*

- H2b: *Men and women are less likely to drop out of leadership positions in male-typical occupations and more likely in female-typical ones compared to mixed occupations, irrespective of gender.*

2.3 Gender Composition and Upward Occupational Mobility, by Gender

In addition to the direct effects of gender and occupational sex segregation, other researchers have argued that the effect of occupational sex segregation may also vary by gender (Dämmrich and Blossfeld 2017; Maume 1999b; Reskin and Roos 1990; Reskin 1993; Cotter et al. 2001). To this end, we lastly inquire whether the male advantage is stronger in male or female occupations. In the following paragraphs, we review several theories that offer polarized viewpoints that we have adopted here as competing hypotheses.

The implicit effect that gender has on the job-matching processes has been extensively demonstrated in relation to statistical discrimination and others means of social closure, i.e. the process by which a group attempts to maintain their position by preventing others from entering (Reskin 1988; Acker 1990; Baron and Newman 1990; Cockburn 1991; Maume 1999a). Women entering male occupations, they not only enter a job queue as job search and job matching theories suggest, but they also enter a "gender queue" whereby employers rank women beneath men due to gender stereotypical belief (Jacobs 1989; Reskin and Roos 1990). For this reason, women are often more disadvantaged when competing for jobs and promotions so that they often are eventually driven out of male occupations due to discrimination or the lack of opportunities (Reskin and Roos 1990).

Kanter's theory of "tokenism" similarly argues that all tokens or minorities are disadvantaged due to heightened visibility, prejudice and gender segregating processes that contribute to social exclusion (Kanter 1977). In line with this theory, men and women are more likely to hold a leadership position in gender-typical occupations than in atypical ones. Respectively, a third hypothesis tested here is:

- H3a: *The likelihood to enter a leadership position is higher through gender-typical occupations than gender-atypical ones.*

In line with the revolving doors analogy (Jacobs 1989), individuals in gender-typical occupations are less likely to drop out of these occupations. Thus, we further hypothesize that men and women spend more time in leadership in gender-typical occupations:

- H3b: *The likelihood to drop out of leadership positions is lower in gender-typical occupations rather than in gender-atypical ones.*

In contrast to Kanter's theory of tokenism, however, role congruity theory argues that men have a greater advantage in upward occupational mobility in female

occupations because they are "only" competing with women whose gender roles are less closely aligned to leadership role characteristics. Similarly, Williams (1992) also argues that men demonstrate a greater advantage in female occupations due to gender stereotyping prejudice in favor of men for leadership positions. Empirical support for this argumentation is given by Dämmrich and Blossfeld (2017). In a country comparative study they investigate women's disadvantage in holding supervisory positions based on the ISCO classification of occupations. Coined as the "glass escalator" effect, this perspective presents competing hypotheses to H3a and H3b:

- H4a: *The male advantage in entering a leadership position is highest in female occupations rather than male ones.*
- H4b: *The male advantage regarding a lower drop out of leadership position is highest in female occupations rather than male ones.*

3 Data and Methods

3.1 Data and Sample

To compare gender and gender compositional effects on upward occupational mobility, we use information on monthly employment biographies from the NEPS, starting cohort 6, (see Blossfeld et al. 2011). This longitudinal dataset contains retrospectively collected employment biographies of individuals born between 1944 and 1986. We use the first four waves available as scientific use file (SUF), carried out from 2009 to 2013. Furthermore, a previous wave of the adult survey was conducted from 2007 to 2008 by the Institute for Employment Research (IAB) under the title, Working and Learning in a Changing World" (ALWA).

We follow individuals from their first significant job for a period of 15 years (180 months). Hence, recent changes in work-family policies are not covered by our data. The first significant job is defined as the first job between the age of 15 and 35 that lasted at least 6 months, which has been similarly used in several previous studies (e.g. Lindemann and Kogan 2013; Smyth 2005). Jobs in preparation for a career, such as internship, traineeship, preparatory service and jobs as student worker are not included. We also excluded respondents who never had a first significant job or have missing information for additional sample-defining characteristics, such as gender or birth date. Furthermore, we excluded individuals born after 1975, as there are too few individuals in the latter birth cohort that adhere to our selection criterion of 180 months of observation following their first significant job. After data preparation and cleaning, our sample consists of 6,402 individual employment

biographies of which 2,926 are female (45.7%) and 3,476 are male (54.3%). We cover the birth cohorts from 1944 to 1955 (32.8%), from 1956 to 1965 (41.1%) and from 1966 to 1975 (26.1%).[2]

3.2 Variables

Our primary variables of interest are: (1) upward occupational mobility, (2) the gender of respondent and (3) gender composition of the occupation held at each point in time. In the following we show how these concepts are operationalized.

3.2.1 Upward Occupational Mobility

With regards to upward occupational mobility, we are chiefly interested in whether men are more likely, to enter and to stay in leadership compared to women. A leadership position is defined as supervisors and executives, coded with "9" as digit four of the KldB2010—the German job classification—coding, and coded with 3 or 4 as digit five of KldB2010 (educational requirement level). Following the German statistical office, we also code 71104 (Managing directors and executive board members-highly complex tasks), 71214 (Legislators-highly complex tasks) und 71224 (Senior officials of special interest organizations-highly complex tasks) as leadership positions (Eisenmenger et al. 2014). Regarding this definition 253 of the 1286 occupations are defined as leadership position. However, not all occupations are represented in our sample.

The first outcome "entering leadership position" is defined as first month of an employment in a leadership position after entering the labor market; the second outcome "staying in" versus "leaving" is defined as any state which is not a leadership position after holding one. This is irrespective to job change, i.e. if the individual continues in another leadership position at a different job, the time spent is viewed as leadership continuous. Furthermore, we are not able to control for if the drop out is voluntary or involuntary.

3.2.2 Gender and Gender-Type of Occupation

While the interviewers report the respondents' gender, the occupation is surveyed by the open question: "Let's start with the first job you had since <DATE>. Please tell me what occupation this was!" The additionally merged gender composition of occupations based on the German Mikrozensus is provided by the German Labor

[2]Further descriptive statistics are available here: https://www.researchgate.net/publication/320036349_Appendix_only_online.

Agency. As the gender composition of occupations is subject to changes over time, we use the mean share of female employees between the years 2001 and 2011.[3] The occupations were then categorized as female occupations with more than 70% of female employees, mixed occupations with 30 up to 70% women, and male occupations with less than 30% of female employees subject to social insurance contributions.[4]

3.3 Methods

For a first glance we use sequence visualization to describe occupational biographies of men and women. Therefore, we distinguished between nine mutually exclusive states that are based on employment activity, the gender composition of a job held and whether or not the position is in a managerial capacity. These include: (1) manager in female occupation, (2) employee in female occupation (3) manager in mixed occupation, (4) employee in mixed occupation (5) manager in male occupation, (6) employee in male occupation, (7) parental leave, (8) unemployment and (9) education and training. Additionally, we had to include a tenth state for gaps.

In a second step we look at Kaplan-Meier survival functions. To compare the survivor functions between our groups of interest, we are calculating risk sets for each of the 180 month of observation "for being in a leadership" position or "not being in a leadership" position. Firstly, we compare leadership positions held by men and women; secondly, we compare the duration of men and women in leadership position by gender-type of occupation. For both calculations, we use four test statistics: Log-rank, Wilcoxon, Tarone-Ware and Peto-Peto test as recommended by Blossfeld et al. (2012).

In a third step, we use discrete-time event history models, which documents whether, and if so, when events occur (Andreß et al. 2013). We use separate analyses for our two outcome variables of interest: (1) we model the accessibility of leadership for the whole sample and (2) we estimate the probability to leave the leadership position for those who at least once in observation time hold a managerial position. Subsequently, our dependent variables are conditional transition probabilities that individual i will experience the respective event at time t, given that the individual hadn't such a transition already in the past. We do not allow for repeating events. Thus, we only consider the first managerial position observed and assume that the dependent variables are dichotomous: "0" for the origin state and "1" for the destination state. Observations after the first occurrence of the event of interest are no longer part of the analysis.

[3]The mean is based on the data from 30th of June as record date for each year to prevent bias of seasonal variation. At that time, the labor market is sturdiest due to stable weather conditions.

[4]With a stricter cutting point of, for example, 80%, there are too few occupations female-typical and with a lower cutting point, such as 60%, occupations that have a nearly balanced gender ratio are also defined as female. However, lower and higher cutting points are used for robustness checks.

As the starting point of our observation period, we identify the first significant job as the point of entry. For the second analysis, we begin with entry into the first leadership position. Although our data are not left censored, we do not know or do not take into account if any of the observed individuals move up into or leave a leadership position after observation time. Thus, we may have right censored data. For this reason we chose this period length of 15 years apart from labor market entry to have a balanced panel data set.

The event history model is estimated using logistic regression, including a time variable as independent covariate, and time-constant as well as time-varying control variables (for more detailed discussion see e.g. Andreß et al. 2013). Furthermore, we use robust standard errors to take into account that month are nested within individuals.

4 Results

To first examine the relationship between the probability to enter a leadership position with gender and gender-typical occupations, Sect. 4.1 presents the visualization of occupational biography sequences; in Sect. 4.2 we report results for Kaplan-Meier Survivor Functions by gender and gender-type of occupation, as well as regression results of event history analysis for entering a leadership position; and finally Sect. 4.3 shows the results for the probability to drop out of leadership for the subsample of those who hold such a position.

4.1 Leadership Position by Gender and Gender-Typical Occupation

Based on the KldB2010 measure, only 6.2% of individuals in the sample hold a leadership position. With regards to gender differences, men appear to have a comparative advantage over women: Only 3.4% of women hold a leadership position, compared to 9% of men.

Table 1 shows the duration (Share of Month) in leadership for the subsample of those at least once in leadership. Here again, all gender differences are highly significant and as expected. Women seem to have an advantage in female occupations, while men demonstrate a comparable advantage in mixed and male occupations. Compared to women, men are only more likely to stay in a leadership position in mixed or male- occupations compared to female ones. From these descriptive results, men are more likely to stay in leadership positions in gender-typical occupations. A male advantage in female occupations is not observable.

In Fig. 1, we illustrate the distribution of occupational states by gender for a period of 180 months, following the first significant job. Both men and women are likely to start their employment biography in a gender-typical occupation (62% for women and 56% for men). Thereby, at least at labor market entry, men are less gender-typical than women. In addition, the difference between distributions of

Table 1 Duration in leadership, by gender and occupational gender-type

	Women	Men	Total	Chi^2 (Pr)
Female-typical leadership occupation (n = 7,686)	58.1	49.2	56.6	0.000
Mixed leadership occupation (n = 26,288)	62.3	89.2	78.4	0.000
Male-typical leadership occupation (n = 54,708)	62.6	87.7	81.6	0.000
All occupations (n = 88,695)	61.5	87.3	78.5	0.000

Data: NEPS SUF, SC6 D-5.1.0

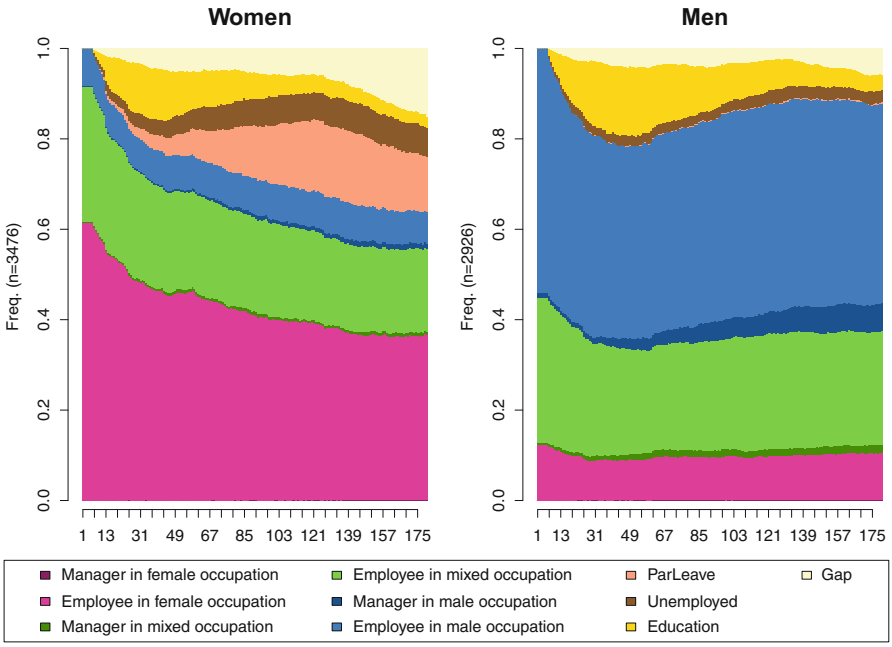

Fig. 1 Sequence distribution plot of occupational states, by gender. Time (horizontal axis) is the number of months following the first significant job. Data: NEPS SUF, SC6 D-5.1.0

occupational states at 1 and 180 months is much more varied for women than it is for men. For example, the share of women working in a female occupation has decreased from 60 to less than 40% by the end of observation period.

For men, the share employed in a male occupation is nearly the same in month 180 after labor market entry. It should also be noted, however, that roughly 35% of women have dropped out of the labor market by month 180, presumably accounting for much of the decline of women in female occupations. It is also observable that the gender differences in holding a leadership position, is smallest at the beginning of observation time. As time goes by, more men than women enter leadership positions, especially in male occupations. Regardless of the occupation, however, the highest proportion of leadership positions is observable at the end of observation time, for both men and women.

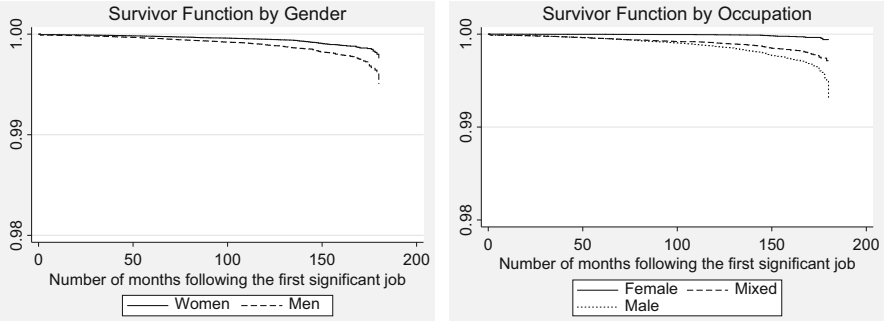

Fig. 2 Access to leadership, by gender and gender-type of occupation. Data: NEPS SUF, SC6 D-5.1.0

4.2 Access to Leadership Positions

4.2.1 Kaplan-Meier Survivor Function

In this section, we present results from product-limit estimations by gender and gender-type of occupation. This technique has the advantage to be a time-driven estimation technique, meaning that we can demonstrate how differences develop over observation time. The survival curves demonstrated in Fig. 2 reflect the effect of gender—or respective gender-type of occupation—on the probability to "survive" without entering a leadership position. Thus, a "failure" means upward occupational mobility. Subsequently, we applied several test statistics to test whether differences between the groups are significant.

It is obvious that differences are increasing over time, even if they remain relatively small. However, we can observe the expected pattern: men seem to have an advantage to "not survive" without upward mobility. All four applied test statistics confirm that gender differences are highly significant ($p < 0.001$).

Similarly, the survivor functions by gender-type of occupation meet our expectations: Individuals in female occupations are at lower risk to take up a leadership position then individuals in mixed and especially male occupations. All four test statistics again are highly significant ($p < 0.001$). However, none of those survivor functions take into account a possible interaction of gender and gender-composition. Furthermore, it is not controlled for further heterogeneity between the groups. Therefore, we show results of event history models in the following.

4.2.2 Regression Results

To disentangle the relevance of gender, gender-type of occupation and their interaction for the upward occupational mobility of men and women, we estimate hierarchical discrete-time event history models with robust standard errors (Table 2). In the first model we only include gender as explanatory variable beside all control

Table 2 Logistic EHA for access to leadership positions

	M1	M2	M3
State number per ID	0.021***	0.020***	0.020***
Men	0.792***	−0.096	0.294 #
Gender-type of occupation (Ref. mixed)			
Sextype female		−2.213***	−2.329***
Sextype male		0.728***	1.230***
Interaction of gender and gender-type of occupation (Ref. mixed)			
Men*female occupation			0.793 #
Men*male occupation			−0.768***
Time constant control variables			
Cohort (Ref. 1944–1955)			
1956–1965	−0.022	−0.122	−0.113
1966–1975	0.085	−0.012	−0.008
Born in Germany	−0.115	0.029	0.026
Age at LM Entry	0.033 *	0.040 **	0.038 **
Educational Degree at LM Entry (Ref. without vocational degree)			
With VET	0.238#	0.229	0.265#
With higher educational degree	0.484*	0.381	0.394
Time varying control variables			
Marital status (Ref. single)			
Married	−0.226#	−0.222#	−0.218#
Divorced	0.038	−0.003	0.007
Number of children	−0.165	−0.109	−0.125
Number of month employed	−0.028***	−0.028***	−0.028***
Number of month in parental leave	−0.007	−0.003	−0.003
Number of month in unemployment	−0.014*	−0.012#	−0.012
Number of month in further education	−0.006	−0.004	−0.005
Employed as public official	−1.281 ***	−0.849*	−0.937**
Selfemployed	0.118	−0.372	−0.487
Constant	−8.465***	−8.135***	−8.304***
N	836474	836474	836474
Pseudo R-squared	0.03	0.07	0.07
AIC	9253.72	8945.95	8927.97
BIC	9474.82	9190.33	9195.62

Data: NEPS SUF, SC6 D-5.1.0
#$p < 0.1$; *$p < 0.05$; **$p < 0.001$; ***$p < 0.001$

variables; in model 2 we add the gender-type of occupation and model 3 contain both plus their interaction. In this way, Likelihood-Ratio tests can be used additionally to the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to assess the model fit.

The overall model fit measure AIC indicates that model 3 including also the interaction effects is the best model. The BIC measure is slightly lower for M2, but the AIC measurement is more straightforward than the BIC, therefore, the recommended choice if there are contradictory outcomes. Furthermore, model 3 is

also recommended looking at the LR-Test results: M1-M2 ($p < 0.001$) and M2-M3 ($p < 0.001$).

In line with the results of the Kaplan-Meier survivor curves, the time variable shows a general increase of the conditional transition probability that individual i will experience upward occupational mobility at time t, given that the individual hadn't such a transition already in the past. Also the main effects of gender and gender-type of occupation are significant and confirm the results from Kaplan-Meier estimation. When controlling for gender-typicality of occupation (M2), the male gender has a non-significant, negative effect on the conditional transition probability. However, this can be explained through the missing interaction effect, as the gender effect returns to be significantly positive after the interaction is included. Thus, it can be concluded that the consideration of only one aspect—or both main effects–does not lead to proper estimations.

Our results show that women have a significant disadvantage to enter leadership positions compared to men in all occupations. Thus, we find support for H1a: Men are more likely to enter leadership positions compared to women, irrespective of the gender composition of the respective occupation held. Furthermore, we can support H2a: Men and women are more likely to enter a leadership position in male occupations than in mixed ones and less likely to enter a leadership position in female occupations.

Additionally, we find supporting evidence for H3a from the interaction effect of gender and gender-type of occupation. The effect for men is the sum of the main coefficient of "sextype male" (1.230) plus the interaction for men in male occupations (−0.768), which results in a significant positive effect (0.462). Thus, men are more likely to enter a leadership position through male and mixed rather than female occupations, even if the advantage in male occupations is smaller for men than for women. For women, H3a has to be rejected because they have the highest likelihood to enter leadership in male occupations. This finding is in line with previous research in Germany that found women to be less disadvantaged in male occupations (Dämmrich and Blossfeld 2017). The absence of a strong male advantage in male occupations may in part be explained by unobserved personality traits of women who take up male occupations (e.g. lower risk aversion, career-orientation etc.). However, the disadvantage to enter leadership in female occupations is less pronounced for men, while the advantage in male occupations is smaller for men than for women. Thus, we as well do find support for a male advantage—in form of a smaller disadvantage—compared to women in gender-atypical occupations, which supports H4a.

4.3 Leaving Leadership Positions

In the following, we restrict our observations to those individuals who already entered a leadership position. We are now interested in a possible male advantage of staying in a leadership position. Like in the previous section, we first report results from survivor analyses and second from event history regression.

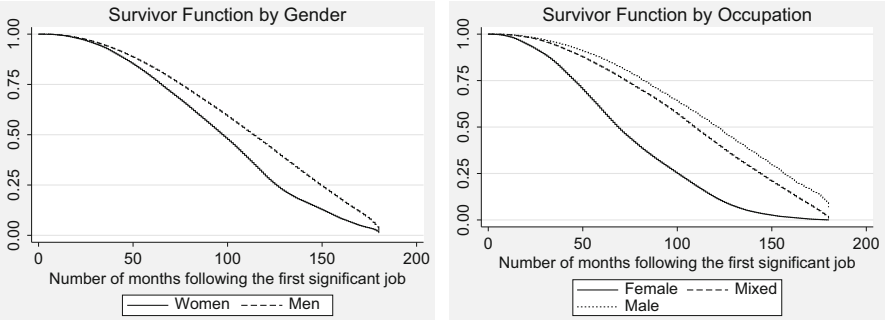

Fig. 3 Dropping out of leadership, by gender and gender-type of occupation. Data: NEPS SUF, SC6 D-5.1.0

4.3.1 Kaplan-Meier Survivor Function

The survival curves in Fig. 3 reflect the effect of gender and gender-type of occupation on the probability to "survive" within a leadership position. Subsequently, a "failure" means the dropout of leadership and stands for the revolving doors. With regards to the "survival" in a leadership position, a comparable male advantage is not observable. Following, the applied test statistics do not confirm gender differences, except the Wilcoxon: Log-rank ($p = 0.697$), Wilcoxon ($p = 0.008$), Tarone-Ware ($p = 0.124$), Peto-Peto ($p = 0.312$). However, a gender-typicality effect is indeed evident and in line with our expectations. The probability to "survive" within a leadership position is steeply decreasing over the observation period, lowest in female occupations and highest in male ones. All four test statistics are highly significant ($p < 0.001$).

While the descriptive results above indicate a stronger gender effect with nearly no differences between mixed and male-typical leadership positions, the Kaplan-Meier estimations are inconsistent. Therefore, it is important to have a closer look at the multivariate analysis for a final assessment of results.

4.3.2 Regression Results

The event of interest for the following event history analysis is "dropping out of leadership" and refers to the revolving door analogy. The month of entry in the first leadership position is the new starting point of analysis. As in the first analysis, a "failure" or drop out of leadership aligns with sample attrition as an individual that already left leadership is no longer "at risk" of dropping out of leadership.

As in the previous section, results (Table 3) are presented including the model fit measures. Most obvious, the model seems to be more appropriate to estimate the conditional probability of "surviving" within a leadership position. Nearly all coefficients are highly significant. The same is true for the LR-Tests which suggest that the full model including the interaction effects is the most appropriate one (M1-M2: $p < 0.001$ and M2-M3: $p < 0.001$). AIC and BIC confirm this suggestion.

Table 3 Logistic EHA for dropping out of a leadership position

	L1	L2	L3
State number per ID	0.022***	0.023***	0.023***
Men	0.115	0.529#	0.531
Gender-type of occupation (Ref. mixed)			
Sextype female		2.381***	2.952***
Sextype male		−0.573*	−0.849#
Interaction of gender and gender-type of occupation (Ref. mixed)			
Men*female occupation			−1.258#
Men*male occupation			0.33
Time constant control variables			
Cohort (Ref. 1944–1955)			
1956–1965	0.014	0.007	0.001
1966–1975	0.168	0.181	0.17
Born in Germany	0.031	0.040	0.008
Age at LM entry	−0.087*	−0.063	−0.067#
Educational degree at LM entry (Ref. without vocational degree)			
With VET	−0.084	−0.044	−0.043
With higher educational degree	0.320	0.323	0.363
Time varying control variables			
Marital status (Ref. single)			
Married	−0.192	−0.344	−0.339
Divorced	0.800**	0.733*	0.757*
Number of children	0.611**	0.546**	0.571**
Number of month employed	−0.006*	−0.007*	−0.007**
Number of month in parental leave	−0.003	−0.009	−0.011
Number of month in unemployment	0.051*	0.042#	0.042#
Number of month in further education	0.000	0.002	0.001
Employed as public official	0.095	−0.262	−0.185
Selfemployed	0.896	0.643	0.489
Constant	−0.091	−0.832	−0.702
N	50892	50892	50892
Pseudo R-squared	0.13	0.21	0.21
AIC	53654.54	48808.49	48546.69
BIC	53813.61	48985.24	48741.12

Data: NEPS SUF, SC6 D-5.1.0
#$p < 0.1$; *$p < 0.05$; **$p < 0.001$; ***$p < 0.001$

The time effect is increasing again in this model, as it was already visible in the Kaplan-Meier curves. In comparison to the results for leadership access, there is no general male advantage for not leaving leadership position. Deviating, the gender-effect indicates a higher transition probability out of leadership for men, but only in M2, without control for the interaction. Thus, we are unable to confirm H1b that men have a general advantage for remaining in a leadership position.

However, the effect of gender composition is in line with our expectations in H2b. Women and men do have a higher probability to drop out of leadership in female compared to mixed occupations. The lower probability to drop out of leadership in male occupations compared to mixed ones is only significant for women so that men do not have an advantage in gender-typical occupations. Thus H3b has to be rejected. However, the disadvantage of a higher drop out risk in female occupations is lower for men. Thus, we find empirical support for H4b, as a male advantage is again visible in form of a lower disadvantage in female occupations.

5 Discussion

We have argued that both gender and occupational gender composition have an independent effect on the likelihood to enter and to stay in leadership. To draw support for this claim, we have presented several theoretical perspectives that offer potential explanation for these effects, including gender role congruity theory, labor market segmentation theory, devaluation theory, tokenism theory and social closure theory. Using these arguments, we contribute to the literature by specifically theorizing as to why gender-typical occupations present different opportunity structures for entering leadership positions and how these may vary by gender.

Using sequence visualization, Kaplan-Meier survivor analysis and event history regression, we examined conditional transition probabilities of men and women into and out of leadership positions. Thereby the aim of this paper was to disentangle effects of gender, gender composition of occupations and their interaction.

For access to leadership positions, most of our hypotheses are supported (see Table 4). Men do have a comparable advantage in entering leadership positions. Their likelihood to enter leadership is highest in male occupations. However, their comparable advantage over women is highest in female occupations but in form of a smaller disadvantage. Our analyses support the presence of a male advantage with regards to upward occupational mobility, even when controlling for occupational gender composition (H1a). Additionally, we were able to show the importance of the gender-typicality of occupations. We presented discrete-time event history results for each of the outcome variables. In line with our theoretical expectations, we find that compared to mixed occupations, female-typical occupations have a negative effect on access to leadership, while male-typical occupations have a positive effect (H2a).

A particular surprising result, however, is the interaction effect between the two. In the theoretical section, we presented two competing hypotheses. We hypothesized that leadership access would be higher in gender-typical occupations rather than atypical ones (H3a). We further tested for a glass escalator effect, whereby the male advantage was hypothesized to be stronger in female occupations rather than male ones. We found that the likelihood for leadership access is highest in gender-typical occupations, but this is only the case for men. Moreover, we found support for H4a, although the male-advantage is only evident in form of a smaller disadvantage

Table 4 Hypotheses and findings

	Access to leadership (a)	Finding	Staying in leadership (b)	Finding
H1	Men are more likely to enter a leadership position compared to women, irrespective of the gender composition of the respective occupation held.	Yes	Men are more likely to stay in a leadership position compared to women, irrespective of the gender composition of the respective occupation held.	No
H2	Men and women are more likely to hold a leadership position in male-typical occupations than in female-typical ones.	Yes	Men and women in male-typical occupations are more likely to stay in leadership positions, irrespective of gender.	Yes
H3	The male advantage in entering a leadership position through gender-typical occupations is greater than gender-atypical ones.	Yes	The male advantage in staying in a leadership position is higher in gender-typical occupations rather than in gender-atypical occupations.	Yes
H4	The male advantage in entering a leadership position is highest in gender-atypical occupations rather than gender-typical ones.	Yes, in form of smaller disad-vantage	The male advantage in staying in a leadership position is highest in gender-atypical occupations rather than gender-typical ones.	Yes, in form of smaller disad-vantage

compared to women. Thus, we do not find any evidence of a glass escalator for men but an advantage compared to women in female occupations.

The second dimension of the comparable male advantage in upward occupational mobility refers to the revolving doors analogy, meaning wherein that individuals—especially women—who manage to enter a leadership position are forced out again. The findings from this analysis are perhaps the most surprising as the male advantage is not statistically significant (H1b). However, the expected gender composition effect is indeed evident: Men as well as women have the highest dropout risk in female occupations (H2b) even if this disadvantage in female occupations is again less pronounced for men, so that H3b as well as H4b receives

support from our results and are not thus competing as expected. There is no male advantage compared to women to stay in leadership position in male occupations.

The general conclusion from our results is that it is not appropriate to analyze gender differences in upward occupational mobility without taking into account the gender composition of occupations and especially the interaction effects. Furthermore, investigating only the access to leadership provides only limited insight into the male-advantage regarding leadership. While many studies have focused on gender differences in leadership access, we found only significant gender differences in terms of staying in leadership for female occupations.

Nevertheless, our results suggest that the (self-)selection into gender-typical occupations largely fosters a male advantage regarding access and lead to a gender difference in dropout risks out of leadership. The lower dropout risk for women in mixed and especially male occupations is likely to reflect a specific selection of women into leadership and into male occupations regarding other factors, such as personality or career-orientation. Unfortunately, we were not able to control for or analyze a probable mechanism of (self-)selection with our data. Therefore, further research is needed to assess results properly. The use of experiments is particularly promising to provide important insights into these mechanisms underlying gender differences and gender discrimination (Correll et al. 2007, see also).

Acknowledgements This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Adults, doi:10.5157/NEPS:SC6:5.1.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

References

Acker, J. (1990). Hierarchies, jobs, bodies: A theory of gendered organizations. *Gender & Society, 4*(2), 139–158.

Aisenbrey, S., Evertsson, M., & Grunow, D. (2009). Is there a career penalty for mothers' time out? a comparison of Germany, Sweden and the United States. *Social Forces, 88*(2), 573–605.

Andreß, H.-J., Golsch, K., & Schmidt, A. W. (2013). *Applied panel data analysis for economic and social surveys*. Berlin/Heidelberg: Springer.

Baron, J. N., & Newman, A. E. (1990). For what its worth: Organizations, occupations and the value of work done by women and non-whites. *American Sociological Review, 55*(2), 155–175.

Benard, S., & Correll, S. J. (2010). Normative discrimination and the motherhood penalty. *Gender & Society, 24*(5), 616–646.

Blossfeld, H.-P., Golsch, K., & Rohwer, G. (2012). *Event history analysis with Stata*. Mahwah: Lawrence Erlbaum.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a lifelong process – The German National Educational Panel Study (NEPS)* (Vol. 14). Sonderheft: Zeitschrift für Erziehungswissenschaft.

Bröckel, M., Busch-Heizmann, A., & Golsch, K. (2015). Headwind or tailwind: Do partners' resources support or restrict promotion to a leadership position in Germany? *European Sociological Review, 31*(5), 533–545.

Brückner, H. (2004). *Gender inequality in the life course: Social change and stability in West Germany 1975–1995*. New York: Aldine de Gruyter.

Budig, M. J. (2002). Male advantage and the gender composition of jobs: Who rides the glass escalator? *Social Problems, 49*(2), 258–277.

Budig, M. J., Misra, J., & Boeckmann, I. (2012). The motherhood penalty in cross-national perspective: The importance of work-family policies and cultural attitudes. *Social Politics, 19*(2), 163–193.

Busch, A. (2013). *Die berufliche Geschlechtersegregation in Deutschland: Ursachen, Reproduktion, Folgen*. Wiesbaden: Springer VS.

Busch, A., & Holst, E. (2009). Glass ceiling effect and earnings: The gender pay gap in managerial positions in Germany. DIW Discussion Papers 905, Berlin.

Charles, M. (2003). Deciphering sex segregation: Vertical and horizontal inequalities in ten national labor markets. *Acta Sociologica, 46*(4), 267–287.

Charles, M., & Grusky, D. B. (2004). *Occupational Ghettos: The worldwide segregation of women and man*. Stanford, CA: Stanford University Press.

Cockburn, C. (1991). *In the way of women: Men's resistance to sex equality in organizations* (Cornell international industrial and labor relations report, Vol. 18). Ithaca: ILR Press.

Correll, S. J., Thébaud, S., & Benard, S. (2007). An introduction to the social psychology of gender. In S. J. Correll (Ed.), *Social psychology of gender* (Advances in group processes, Vol. 24, pp. 1–18). Amsterdam: Emerald.

Cotter, D. A., Hermsen, J. M., Ovadia, S., & Vanneman, R. (2001). The glass ceiling effect. *Social Forces, 80*(2), 655–681.

Dämmrich, J., & Blossfeld, H.-P. (2017). Women's disadvantage in holding supervisory positions. variations among European countries and the role of horizontal gender segregation. *Acta Sociologica, 60*(3), 262–282.

Eagly, A. H. (2003). Few women at the top: How role incongruity produces prejudice and the glass ceiling. In D. van Knippenberg & M. A. Hogg (Eds.), *Leadership and power: Identity processes in groups and organizations* (pp. 79–93). London: Sage Publications.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573–598.

Edwards, R. (1979). *Contested terrain: The transformation of the workplace in the twentieth century*. New York: Basic Books.

Eisenmenger, M., Loos, C., & Sedmihradsky, D. (2014). Erwerbstätigkeit in Deutschland – Ergebnisse des Zensus 2011. *WISTA Wirtschaft und Statistik, 9*, 544–560.

England, P. (2005). Gender inequality in labor markets: The role of motherhood and segregation. *Social Politics, 12*(2), 264–288.

England, P., Herbert, M. S., Kilbourne, B. S., Reid, L. L., & Megdal, L. M. (1994). The gendered valuation of occupations and skills: Earnings in 1980 census occupations. *Social Forces, 73*(1), 65–100.

Farkas, G., & England, P. (Eds.). (1988). *Industries, firms, and jobs: Sociological and economic approaches*. New York: Plenum Press.

Gangl, M., & Ziefle, A. (2009). Motherhood, labor force behavior, and women's careers: An empirical assessment of the wage penalty for motherhood in Britain, Germany, and the United States. *Demography, 46*(2), 341–369.

Härkönen, J., Manzoni, A., & Bihagen, E. (2016). Gender inequalities in occupational prestige across the working life: An analysis of the careers of West Germans and Swedes born from the 1920s to the 1970s. *Advances in Life Course Research, 29*, 41–51.

Holst, E. (2006). Women in managerial positions in Europe: Focus on Germany. *Management revue, 17*(2), 122–142.

Hultin, M. (2003). Some take the glass escalator, some hit the glass ceiling: Career consequences of occupational sex segregation. *Work and Occupations, 30*(1), 30–61.

Jacob, M., Kleinert, C., & Kühhirt, M. (2013). Trends in gender disparities at the transition from school to work: Labour market entries of young men and women between 1984 and 2005 in West Germany. *Journal of Vocational Education & Training, 65*(1), 48–65.

Jacobs, J. A. (1989). *Revolving doors: Sex segregation and women's careers*. Stanford, CA: Stanford University Press.
Kanter, R. M. (1977). *Men and women of the corporation*. New York: Basic Books.
Ko, I., Kotrba, L., & Roebuck, A. (2015). Leaders as males? The role of industry gender composition. *Sex Roles, 72*(7–8), 294–307.
Lindemann, K., & Kogan, I. (2013). The role of language resources in labour market entry: Comparing Estonia and Ukraine. *Journal of Ethnic and Migration Studies, 39*(1), 105–123.
Manzoni, A., Harkonen, J., & Mayer, K. U. (2014). Moving on? a growth-curve analysis of occupational attainment and career progression patterns in West Germany. *Social Forces, 92*(4), 1285–1312.
Maume, D. J. (1999a). Glass ceilings and glass escalators: Occupational segregation and race and sex differences in managerial promotions. *Work and Occupations, 26*(4), 483–509.
Maume, D. J. (1999b). Occupational segregation and the career mobility of white men and women. *Social Forces, 77*(4), 1433–1459.
Ochsenfeld, F. (2012). Gläserne Decke oder goldener Käfig: Scheitert der Aufstieg von Frauen in erste Managementpositionen an betrieblicher Diskriminierung oder an familiären Pflichten? *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 64*(3), 507–534.
Reskin, B. (1988). Bringing the men back in: Sex differentiation and the devaluation of women's work. *Gender & Society, 2*(1), 58–81.
Reskin, B. (1993). Sex segregation in the workplace. *Annual Review of Sociology, 19*(1), 241–270.
Reskin, B., & Roos, P. (1990). *Job queues, gender queues: Explaining women's inroads into male occupations* (Women in the political economy). Philadelphia: Temple University Press.
Ridgeway, C. L. (2001). Gender, status, and leadership. *Journal of Social Issues, 57*(4), 637–655.
Ridgeway, C. L., & Correll, S. J. (2004). Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender & Society, 18*(4), 510–531.
Sengenberger, W. (1987). *Struktur und Funktionsweise von Arbeitsmärkten: Die Bundesrepublik Deutschland im internationalen Vergleich*. Arbeiten aus dem Institut für Sozialwissenschaftliche Forschung e. V. ISF München. Frankfurt/Main: Campus-Verl.
Smyth, E. (2005). Gender differentiation and early labour market integration across Europe. *European Societies, 7*(3), 451–479.
Sørensen, A. B., & Kalleberg, A. L. (1981). An outline of a theory of the matching of persons to jobs. In I. E. Berg (Ed.), *Sociological perspectives on labor markets* (Quantitative studies in social relations, pp. 49–74). New York: Academic.
Trappe, H., Pollmann-Schult, M., & Schmitt, C. (2015). The rise and decline of the male breadwinner model: Institutional underpinnings and future expectations. *European Sociological Review, 31*(2), 230–242.
Williams, C. L. (1992). The glass escalator: Hidden advantages for men in the 'female' professions. *Social Problems, 39*(3), 253–267.
Williams, C. L. (2013). The glass escalator, revisited. *Gender & Society, 27*(5), 609–629.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Modelling Mortality Using Life Trajectories of Disabled and Non-Disabled Individuals in Nineteenth-Century Sweden

Erling Häggström Lundevaller, Lotta Vikström, and Helena Haage

1 Introduction

The purpose of this study is to investigate how disabilities and the experiences of work and family during early adulthood affected subsequent mortality in past society. As in many other historical demographic studies, this calls for a life course approach and a choice of analysis methods accordingly.

In the last two decades, statistician Gilbert Ritschard has promoted sequence analysis for studying events during extended time spans of individual life, just as the research of he and his colleagues has shown (Oris and Ritschard 2014; Ritschard and Oris 2005; Ritschard et al. 2008; Studer and Ritschard 2016). Although statistical life course analysis has come to predominate within the field of historical demography when there is ample access to data, the method of sequence analysis has been of limited use compared to Cox regression models. While the latter models provide accurate estimates of significant factors determining the single event under study, sequence analysis examines a series of several events that help to grasp the life course as the dynamic process it is.

Researchers increasingly call for a combination of the two methods that can work to complement each other (Courgeau 2018; Kok 2007; Madero-Cabib et al. 2015) and some of them have undertaken such an approach of interest to historical demography (Bras et al. 2010; Dribe et al. 2014; Schumacher et al. 2013). A similar approach, but with logistic regression in place of Cox regression, is proposed in Rossignon et al. (2018). See also the first case study in Eerola (2018). We appreciate this move towards joining methods and the present study is an attempt to test this combination so as to contribute results that reflect life and death among disabled

E. H. Lundevaller (✉) · L. Vikström · H. Haage
Umeå University, Umeå, Sweden
e-mail: erling.lundevaller@umu.se

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_5

people historically. They constitute a group whose demographic experiences have received poor recognition in historical research and who are rarely subject to the statistical use of life course analyses.

Our study aims to detect the life sequences among young adults pertaining to their transition to work and family formation, and then see whether there are significant associations between the mortality risks and the specific life sequences we find in a nineteenth-century Swedish population comprising both disabled and non-disabled men and women.

Our previous findings are primarily based on Cox regression models using a larger population (some 35,000 cases), a selection of which is targeted below ($N = 4,116$), that originate from Swedish parish registers digitised by the Demographic Data Base (DDB), Umeå University. Our mortality results demonstrate that disabilities caused people to have significantly higher premature death propensity ($<$ 54 years of age), in particular if having mental disorders or if male regardless of type of disability (Haage et al. 2016). In another study, our Cox regression results suggest that disability jeopardised the marital propensity in similar ways (Haage et al. 2017). In a recent study, we employ sequence analysis on a series of events expected to occur in the life of young adults: work, marriage and parenthood, also taking some account of outward migration and death (Vikström et al. 2017). We found that the trajectories of disabled individuals did not include work or family to the same extent as those of non-disabled people, and that they rarely migrated, but suffered from premature death ($<$34 years of age). The trajectory findings from conducting sequence analysis and the mortality results obtained through Cox regressions models made us curious about the outcomes from combining the two methods somehow, as increasingly suggested by life course scholars.

From the eighteenth century onward, mortality patterns have been investigated through both macro and micro studies, especially in the Western world (Bengtsson 2004). These studies demonstrate gendered variations in mortality across different time-space contexts and age groups. The Tabular Commission (*Tabellverket*) began population statistics in 1749 (Sköld 2001), and since then we can see that the mortality among Swedish men has been higher than that among women, except for some brief time periods, and mainly among young people (Willner 1999; Fridlizius 1988; Edvinsson 1992). This male excess in mortality persisted throughout the nineteenth century although the gap between the genders decreased.

There are few historical studies on how death hit a larger number of disabled individuals and whether their mortality differed from general or gendered patterns. Our own research reveals that disability jeopardised the survival of individuals in nineteenth-century Sweden, but with some variation by type of disability and gender. Both men and women with mental disabilities and men with any type of disability ran the highest premature death risks compared to their non-disabled peers (Haage et al. 2016; Vikström et al. 2017). For East Flanders, Belgium, 1750–1950, De Veirman (2015) presents statistical life course results of deaf individuals. Comparing their mortality risks with those of their hearing siblings, who constitute a reference group, De Veirman cannot find that deafness significantly influenced survival chances. Olsson (1999) provides some results in her study of disabled

people in nineteenth-century Linköping, a town in central Sweden. Measures of their longevity demonstrate that disabled women on average grew older than their male counterparts, but this did not make their mortality patterns different from the gendered death differentials outlined above.

2 Methods

To investigate how disabilities and the experiences of work and family during early adulthood affected subsequent mortality in past society we, have to operationalise these concepts, find adequate data and apply a suitable statistical approach to analyse the data.

The strategy chosen here is to first note the occurrence and type of disability prior to the age of 15. Second, life trajectories are analysed using sequence analysis between ages 15 and 33 in order to determine homogeneous groups, given their experience of work and family in their early adulthood. Important demographic events that occur in the life of young adults—first occupation, first marriage and first child—are recorded yearly and cause the person's trajectory to change state. From the parish registers we know the date of the events so the dates are discretised to the age of the individual in full years. The reason for choosing young persons is that they were, in the beginning of their transition to adulthood, associated with the central events under study, getting the first job, marrying for the first time and giving birth to the first child. Third, the groups derived are used as explanatory variables in combination with disability and other variables in continuous Cox regressions with mortality as outcome. The individuals are followed from their 33rd birthday as long as the registers permit and it is noted if the period ends with death or if the observation is censored. The duration of this period is counted in days as we have dates for the events.

As indicator of different types of disabilities, notes from the parish registers are used as described in detail in the next section. As indicators of experience of work and family, we have chosen to use the occurrence of first job, first marriage and first child. These are used to create two sets of explanatory factors. Both of these are created by first looking at the yearly combination of these occurrences, giving rise to a preliminary factor variable factor with eight levels. E.g. if a person has experienced the first marriage, first child and first occupation the state will be "Married/child/occupation" for the rest of the observation period.

The first factor variable is constructed using sequence analysis by clustering them into similar groups based on similarities of the trajectories. This is done using Ward hierarchical clustering of the sequences using the optimal matching distances between the sequences of transitions between states (Studer and Ritschard 2016). The method used is implemented in TraMineR in the function seqdist as "OMstran", here used with parameters otto = 0.2 and indel = 1. This method is highly sensitive to the sequencing of the events which is relevant here. Substitution costs are calculated with the method "TRATE" which uses the observed transition rates. The parameter values are chosen so that they give groups that perform well in the Cox regression.

The second factor variable we will call End-state. It is the status just before the 33rd birthday and serves as a comparison to the variable constructed above.

Two control variables indicating urban setting and cohort are also used. Cohort is defined as cohort 1 if the person is born 1820–1829 and cohort 2 if the person is born 1850–1859. The parishes are divided into rural and urban/industrial as will be detailed in the data section.

We also have access to other variables, for example indicating the socio-economic background. We do not use it, since studies have shown this background had little effect on mortality in the period and region under study (Edvinsson and Broström 2012; Edvinsson 1992; Haage et al. 2016).

All our analyses are performed in the statistical environment R, using the package TraMineR (Life Trajectory Miner for R, Gabadinho et al. 2011).

3 Data

To meet the purpose of our study, and to be able to handle the vast information about individuals and impairments reported in the DDB's parish registers, we target a young population in the Sundsvall region. The data-set consists of observations of 4,116 unique 15-year-old individuals born 1820–29 or 1850–59, of whom 74 had marks of impairments before the age of 15 years. In the following subsections, the area and parish registers that are the source of our data is described.

3.1 Area Selected for Analysis

The Sundsvall region is chosen as a research area. (See Fig. 1). This region is a fairly representative selection of the population makeup and the economic structure found elsewhere in nineteenth-century Sweden and north-western Europe. Also in the Sundsvall region the majority of people depended on agricultural production, especially eight parishes in this study, Attmar, Hässjö, Indal, Ljustorp, Selånger, Såttna, Tuna and Tynderö. In another four parishes, Alnö, Skön, Njurunda and Timrå, the socio-economic and demographic structure transformed from the middle of the century onwards, from agriculture to being primarily shaped by the sawmill industry. In these parishes and Sundsvall town, the effects of industrialisation were most evident. Besides the mortality decline typical for the nineteenth century, the large influx of primarily domestic migrants looking for better prospects in this expanding sawmill industry explains the rapid population growth during the latter half of the century (Bergman 2010; Edvinsson 1992; Vikström 2003). In 1840 there were 18,793 inhabitants, a number that had increased to 46,418 in 1880 (Alm Stenflo 1994).

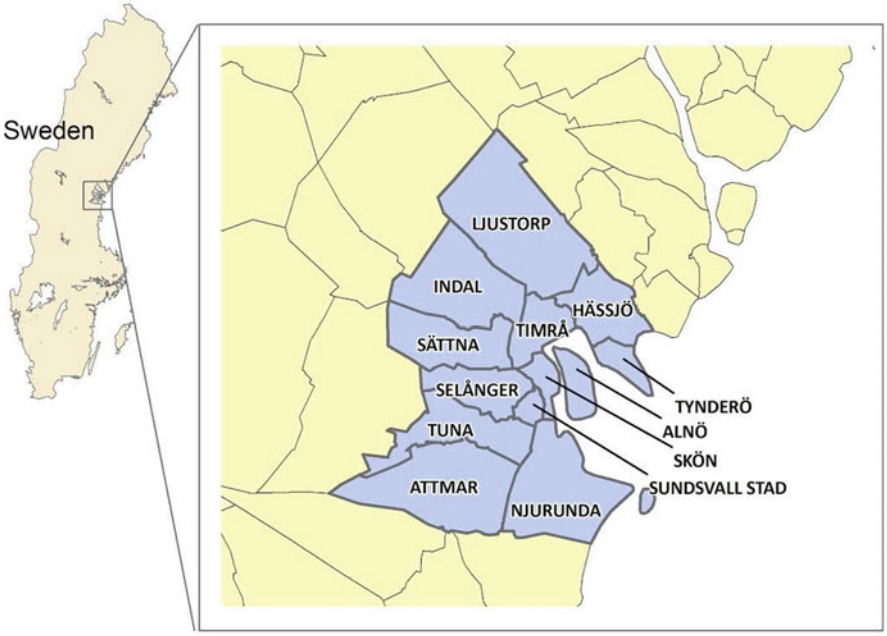

Fig. 1 Map of Sweden and the Sundsvall region and the parishes included in the study. (Source: Demographic Data Base, Umeå University)

3.2 Digitised Parish Registers Indicating Disabilities

The parish registers of the Sundsvall region are digitised and stored by the Demographic Data Base (DDB) at Umeå University, Sweden. They are based on original registers for parishioners' birth, baptism, marriage, outward or inward migration, death and burial and the catechetical examination records. They also provide notes on occupation. As all these registers are digitised and linked on an individual level, they yield a demographic description of each parishioner containing essential life events, such as the start and type of work, marriage, childbearing, relocation or death (Westberg et al. 2016; Vikström et al. 2006). The rich information across an individual's lifetime makes the DDB registers well suited for life course research.

The catechetical examination records (*husförhörslängder*) explain why Sweden's parish registers are exceptionally informative. They go back to the seventeenth century, due to the obligation of the church ministers to keep records of the parishioners' knowledge of the catechism and their reading skills (Nilsdotter Jeub 1993). In these records, the ministers reported events such as occupational changes and also made other notes about the parishioners, such as remarks about their impairments (*lytesmarkeringar*), which indicate the presence of disabilities. Parishioners whom the ministers recognised as disabled have been manually categorised as such by us, since this information is not consistently coded by the DDB. Further, we account

Table 1 The categorisation of disability based on the notes of impairment in the parish registers from the Sundsvall region, 1835–1844 and 1865–1874

Disability	Description
Blind	Visual defects from weak-sighted, short-sighted to blind
Deaf mute	Hearing or communication dysfunctions, ranging from poor hearing to deaf and from difficulties speaking, stammer to mute
Crippled	Physical dysfunctions e.g. lame, limping, walking on crutches, missing body parts, hare-lipped, small in size or crippled
Idiot	Mental dysfunctions since childhood and lack of full intellectual development as an adult, e.g. foolish, silly or less cognisant (Mindre vetande)
Insane	Mental dysfunctions identified in adulthood and fully developed intellect as a child, e.g. insane, feeble-minded or crazy
Multiple disabilities	Combination of two or more of the above disabilities

only for fairly evident impairments, such as hearing and visual disabilities and a few other types of physical or mental dysfunctions described in Table 1. In the analysis we have merged the two groups that ministers labelled as idiots or insane into a new group that we term "mentally disabled". All other disability groups are merged into one group called "other disabilities" in the Cox regression. This facilitates comparisons both within the group of disabled people and with parishioners in the Sundsvall region who did not have any of these particular impairments reported in the parish registers. Thus, those who were not blind, deaf mute or crippled and so forth we recognise as being non-disabled, even though some of them may have had impairments more vaguely defined by the ministers or perhaps suffered from some illnesses (Drugge 1988; Haage et al. 2017; Rogers and Nelson 2003). They represent the average, or typical, life trajectory of the population living in the same time-space context as did the group of disabled people.

Of course, there is a risk of underestimating disabilities in the parish registers, as this type of documentation was not the primary task of the ministers. However, this is not a big problem, as disabled persons who may have been incorrectly added to the group of non-disabled people will affect the results very little and because we can be fairly certain that those in the group with disabilities were in fact disabled.

In Table 2 the frequencies and mean observation time for the different disability groups are presented. Observation time is the time in years from observation start at the 33 birthday until death or censoring. There are considerable differences between the groups in follow up time due to different mortality and migration patterns.

In Table 3 the frequencies and mean observation time for the different end types are shown. Of all observations, 919 end with death, the outcome we are interested in.

Table 2 Frequency and mean observation time in years after the 15th birthday by disability group

Disability	Frequency	Mean observation time
Non-disabled	4042	16.00
Blind	6	32.11
Crippled	14	20.04
Deaf mute	22	17.48
Mentally disabled	32	14.02

Table 3 Frequency and mean observation time in years by end type

End	Frequency	Mean observation time
Unknown	14	7.56
Registers end	2545	16.15
Dead	919	21.91
Migrated	638	7.26

4 Results

First, results from the sequence analysis are presented, followed by the Kaplan-Meier curves for survival for the different groups derived in the sequence analysis. Lastly, the results of the Cox regressions are presented.

4.1 Sequence Analysis Results

In Fig. 2 the transversal state distribution within each cluster found in the cluster analysis for men is shown. Three typical types are found. Type 1 gets a job but does not get a family quickly. Type 2 gets a job, gets married and has a child rather quickly. Type 3 is characterised by not getting a job or family until late in the observation period or not at all.

The pattern for women shown in Fig. 3 is a bit different. Type 1 starts a family but does not get a job. Type 2 gets a job early and then most get a family. The third type is, as for the men, characterised by not getting a job or family until late in the period or not at all.

These identified types below are used as stratification variables to make Kaplan-Meier curves and in the Cox regression as explanatory variables to see if they affect mortality.

4.2 Kaplan-Meier Curves

The x-axis in the Kaplan-Meier curve is years after the 33rd birthday. p-values in the caption of the figures refer to log-rank or Mantel-Haenszel tests (Harrington and Fleming 1982).

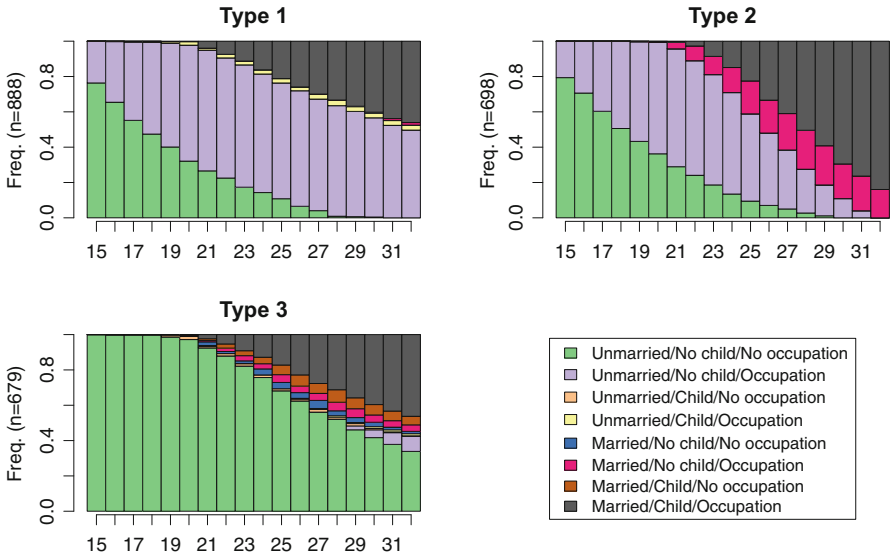

Fig. 2 Transversal state distributions for the types found for men

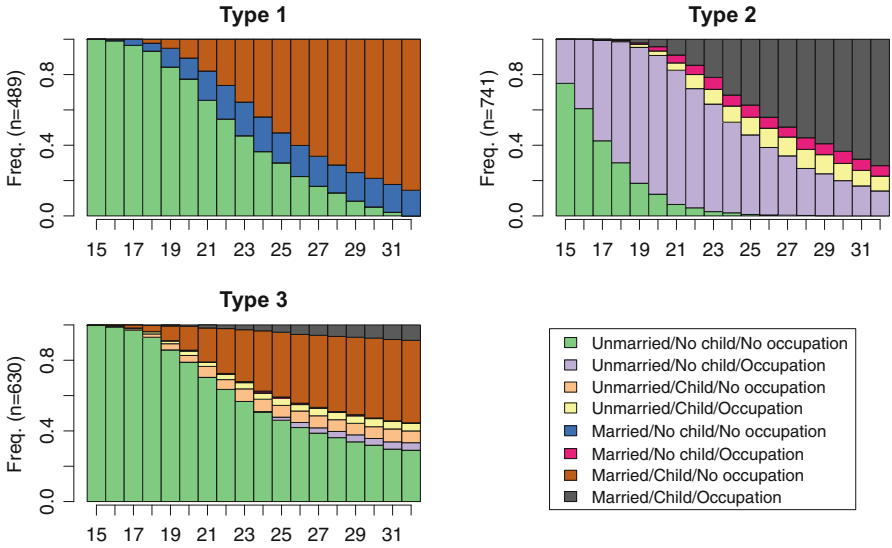

Fig. 3 Transversal state distributions for the types found for women

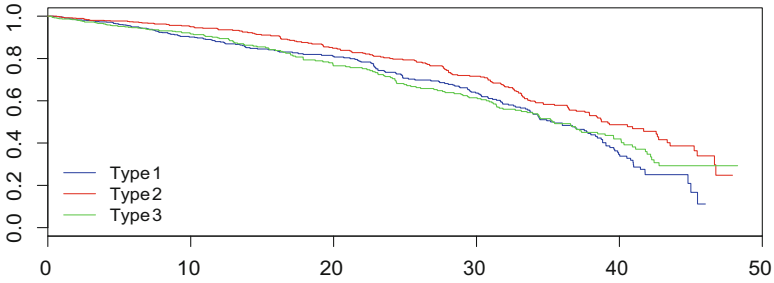

Fig. 4 Kaplan-Meier by type according to sequence analysis, men, p-value: 0.001

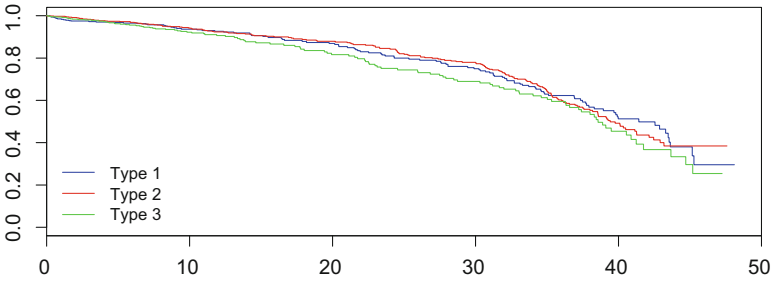

Fig. 5 Kaplan-Meier by type according to sequence analysis, women, p-value: 0.235

Figures 4 and 5 show the Kaplan-Meier curves for the types extracted for men and women respectively in the sequence analysis. The curves for men are significantly different with a lower mortality for type 2, those that achieved both work and family. The curves for women are not significantly different but there is a tendency for type 3, those that remained without family and work longest, to have a slightly lower curve indicating higher mortality.

4.3 Cox Regression Results

Four different models are analysed with Cox regression, two for men and two for women. The genders are analysed in different models as it is likely that mortality patterns differ. Within gender the models differ depending whether End-state or Type is used as explanatory variables. Reference level for End-state is married/child/occupation and for Types, Type 1. Variables controlling for urban or rural setting (Rural/Urban) and cohort inclusion (Cohort) are also used. Reference levels for these variables are Rural and Cohort 1. Proportionality tests for all models indicate no significant violations of the proportionality assumption.

The results are shown in Table 4. The significance test for the complete factor variables (Disability, End-status, Type, Urban and Cohort) as shown in Table 5

Table 4 Cox regression estimates as hazard ratios. Reference levels are Non-disabled, Married/child/occupation, Men Type 1, Women Type 1, Rural parish and Cohort 1

	Hazard ratios for mortality			
	Men	Women	Men	Women
	(1)	(2)	(3)	(4)
Mentally disabled	1.516	3.620**	1.446	3.512**
Other disability	1.064	0.903	1.036	0.883
Unmarried/No child/No occupation	1.097	1.115		
Unmarried/No child/Occupation	1.273**	1.097		
Unmarried/Child/No occupation	0.0004	1.167		
Unmarried/Child/Occupation	1.572	1.001		
Married/No child/No occupation	0.596	0.905		
Married/No child/Occupation	1.113	0.882		
Married/Child/No occupation	1.017	1.022		
MenType 2			0.689***	
MenType 3			0.916	
WomenType 2				1.022
WomenType 3				1.203
Urban/Industrial	1.113	1.151	1.112	1.147
Cohort 2	1.029	0.813	1.004	0.796
Observations	2,262	1,854	2,262	1,854

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 5 p-values for the factors in the four models

Model	(1)	(2)	(3)	(4)
Factor:				
Disability	0.59	0.11	0.66	0.12
Cohort	0.85	0.26	0.98	0.22
End-status	0.49	0.99		
Type Men			0.00	
Type Women				0.35
Rural/Urban	0.29	0.24	0.29	0.25

shows a significant result at the 5% significance level only for the type of men variable derived from the sequence analysis. However, comparing levels within factor, mentally disabled women have a significantly higher mortality than non-disabled women in Model 2 and 4. For men, having a job, wife and family, implied significantly lower mortality risk than if only having a job in Model 1.

5 Discussion

This study combines sequence analysis with event history analysis to examine how past trajectories in terms of work and family determined mortality among disabled and non- disabled individuals in nineteenth-century Sweden. We find that sequence

analysis is a useful tool to encapsulate information from people's life histories and then make use of this information in Cox regressions to gain evidence on how past trajectories shaped subsequent events in life, here mortality. While the results from using end states did not render any significant outcome, two major findings stand out from our examination of the variables constructed by sequence analysis. First, it seems as if experiencing work and family during young adult life (between 15 and 33 years of age) significantly affected subsequent mortality, but only for men. The trajectories that consist of job, marriage and children, often perceived as the desirable life path, seem beneficial for men. This suggests that men's survival was more sensitive to having a job, a spouse and family than was the case for women. Our second major finding regards the effect of disability, or rather the low impact of it. While mental disabilities among women made them run significantly higher mortality risks, such an effect was insignificant among women having other types of disabilities. The effect of disability was also insignificant among and all men. However, it is possible that some of the negative effect of disability is captured in the variables based on past trajectories. There may be yet another reason contributing to the absence of significant effects of disability on mortality and it concerns a possible selection bias. It must be borne in mind that all individuals under observation have survived their 33rd birthday. This implies that the persons most frail in the disability group, which we expect to be more vulnerable than the non-disabled group, have already died before reaching this age, probably leaving the strongest and healthiest disabled persons for us to study. Moreover, according to the "healthy migrant theory" (Abraido-Lanza et al. 1999), frail people tend to relocate to a low extent, just as did the disabled individuals we analyse in comparison to the high level of migrants in the non- disabled group. If frail people from the latter group did not move and thus are included in our follow up, this partly explains the low mortality difference between the two groups. However, the life-course results we obtain through a combination of sequence analysis and event history analysis suggest that disability cannot only be associated with the disadvantages indicated above with high mortality risks. Obviously, getting a job, spouse and family during young adulthood helped to limit nineteenth-century individuals' death risks later in life even if disability was part of it.

Acknowledgements This study is part of a project headed by Lotta Vikström that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 647125), 'DISLIFE Liveable Disabilities: Life Courses and Opportunity Structures Across Time'. The study is also part of another project led by Lotta Vikström, 'Experiences of disabilities in life and online: Life course perspectives on disabled people from past society to present', funded by the Wallenberg Foundation (Stiftelsen Marcus och Amalia Wallenbergs Minnesfond). The authors wish to thank the three anonymous reviewers and the editors for providing suggestions on how to advance our analysis. We are also grateful to all the support regarding both research and data that we enjoy from our colleagues at the Centre for Demographic and Ageing Research (CEDAR) and the Demographic Data Base (DDB), Umeå University, Sweden.

References

Abraido-Lanza, A. F., Dohrenwend, B. P., Ng-Mak, D. S., Turner, J. B. (1999). The Latino mortality paradox: A test of the "salmon bias" and healthy migrant hypotheses. *American Journal of Public Health, 89*(10), 1543–1548.

Alm Stenflo, G. (1994). *Demographic description of the Skellefteå and Sundsvall regions during the 19th century.* Umeå: Demographic Data Base.

Bengtsson, T. (2004). Living standards and economic stress. *Life under pressure: Mortality and living standards in Europe and Asia, 1700–1900* (pp. 27–59). Cambridge: The MIT Press.

Bergman, M. (2010). Constructing communities: The establishment and demographic development of sawmill communities in the Sundsvall district, 1850–1890. Technical report, Umeå: Umeå University.

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.

Courgeau, D. (2018). Do different approaches in population science lead to divergent or convergent models. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

De Veirman, S. (2015). *Breaking the silence: The experiences of deaf people in East Flanders, 1750–1950: A life course approach*. Ph.D. thesis, Ghent University.

Dribe, M., Manfredini, M., Oris, M. (2014). The roads to reproduction: Comparing life-course trajectories in Preindustrial Eurasia. In C. Lundh, Kurosu, S., et al. (Ed.), *Similarity in difference: Marriage in Europe and Asia, 1700–1900*. Cambridge, MA: MIT Press.

Drugge, U. (1988). Om husförhörslängder som medicinsk urkund. Psykisk sjukdom och förstånd-shandikapp i en historisk källa. *Scriptum. Rapportserie utgiven av forskningsarkivet vid Umeå universitet, 8*.

Edvinsson, S. (1992). *Den osunda staden: Sociala skillnader i dödlighet i 1800-talets Sundsvall*. Ph.D. thesis, Umeå universitet.

Edvinsson, S., & Broström, G. (2012). Old age, health and social inequality: Exploring the social patterns of mortality in 19th-century Northern Sweden. *Demographic Research, 26*, 633–660.

Eerola, M. (2018). Case studies of combining sequence analysis and modelling. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Fridlizius, G. (1988). Sex-differential mortality and socio-economic change: Sweden 1750–1910. In A. Brändström & L.-G. Tedebrand (Eds.), *Society, health and population during the demographic transition* (pp. 237–272). Stockholm: Almqvist & Wiksell International.

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.

Haage, H., Häggström Lundevaller, E., Vikström, L. (2016). Gendered death risks among disabled individuals in Sweden: A case study of the 19th-century Sundsvall region. *Scandinavian Journal of History, 41*(2), 160–84.

Haage, H., Vikström, L., Häggström Lundevaller, E. (2017). Disabled and unmarried? Marital chances among disabled people in nineteenth-century Northern Sweden. *Essays in Economic & Business History, 35*(1), 207–238.

Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika, 69*, 553–566.

Kok, J. (2007). Principles and prospects of the life course paradigm. *Annales de démographie historique* (1), 203–230.

Madero-Cabib, I., Gauthier, J.-A., Le Goff, J.-M. (2015). The influence of interlocked employment–family trajectories on retirement timing. *Work, Aging and Retirement, 2*(1), 38–53.

Nilsdotter Jeub, U. (1993). *Parish records: 19th century ecclesiastical registers*. Umeå, Umeå University: Information from the Demographic Data Base.

Olsson, I. (1999). *Att leva som lytt: Handikappades levnadsvillkor i 1800-talets Linköping*. Ph.D. thesis, Tema, Linköpings universitet.

Oris, M., & Ritschard, G. (2014). Sequence analysis and transition to adulthood: An exploration of the access to reproduction in nineteenth-century East Belgium. In *Advances in sequence analysis: Theory, method, applications*(pp. 151–167). Springer, Cham.

Ritschard, G., Gabadinho, A., Muller, N. S., Studer, M. (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management, 1*(1), 68–90.

Ritschard, G., & Oris, M. (2005). Life course data in demography and social sciences: Statistical and data-mining approaches. *Advances in Life Course Research, 10*, 283–314.

Rogers, J., & Nelson, M. C. (2003). "lapps, Finns, Gypsies, Jews and Idiots" modernity and the use of statistical categories in Sweden. *Annales de démographie historique, 1*(105), 61–79.

Rossignon, F., Studer, M., Gauthier, J.-A., Le Goff, J.-M. (2018). Sequence history analysis (SHA): Estimating the effect of past trajectories on an upcoming event. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Schumacher, R., Matthijs, K., & Moreels, S. (2013). Migration and reproduction in an urbanizing context. Family life courses in 19th-century Antwerp and Geneva. *Revue Quetelet, 1*, 51–72.

Sköld, P. (2001). *Kunskap och kontroll: Den svenska befolkningsstatistikens historia*. Almqvist & Wiksell.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179*(2), 481–511.

Vikström, L. (2003). *Gendered routes and courses: The socio-spatial mobility of migrants to nineteenth-century Sundsvall*. Umeå: Umeå University, Demographic Data Base.

Vikström, L., Haage, H., Häggström Lundevaller, E. (2017). Sequence analysis of how disability influenced life trajectories in a past population from the nineteenth-century Sundsvall region, Sweden. *Historical Life Course Studies, 4*, 97–119.

Vikström, P., Brändström, A., Edvinsson, S. (2006). Longitudinal databases: Sources for analyzing the life course. *History and Computing, 14*, 109–128.

Westberg, A., Engberg, E., Edvinsson, S. (2016). A unique source for innovative longitudinal research: The POPLINK database, historical life course studies. Vol. 3, 20–31.

Willner, S. (1999). *Det svaga könet?: Kön och vuxendödlighet i 1800-talets Sverige*. Ph.D. thesis, Linköpings Universitet.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Sequence History Analysis (SHA): Estimating the Effect of Past Trajectories on an Upcoming Event

Florence Rossignon, Matthias Studer, Jacques-Antoine Gauthier, and Jean-Marie Le Goff

1 Introduction

In many research questions framed within the life-course paradigm, the estimation of the effect of a previous trajectory on an upcoming event is of central interest. While this paradigm recognizes that structural constraints influence choices and outcomes, it also acknowledges the significant impact of past individual trajectories. Many previous studies addressed such kinds of issues. For instance, Madero-Cabib et al. (2015) modeled the influence of past occupational trajectories on the timing of retirement. Studer (2012) studied the effect of past working and financing conditions on the chances of obtaining a PhD among teaching assistants at the University of Geneva. Lundevaller et al. (2018) investigated how different work and family trajectories during early adulthood affected mortality risks during late adulthood in Sweden in the nineteenth century. Finally, Eerola and Helske (2016) studied how trajectories of partnership formation and parenthood predict depression scores (see also the first case study in Eerola 2018, in this bundle). In all these examples, the past processes under study consist of categorical states unfolding over time, such as family or occupational statuses. These kinds of research questions might also

F. Rossignon (✉)
Swiss Federal Statistical Office, Neuchâtel, Switzerland
e-mail: florence.rossignon@bfs.admin.ch

M. Studer
NCCR LIVES and Geneva School of Social Sciences, University of Geneva, Geneva, Switzerland

J.-A. Gauthier · J.-M. Le Goff
NCCR LIVES and University of Lausanne, Lausanne, Switzerland
e-mail: jacques-antoine.gauthier@unil.ch; jean-marie.legoff@unil.ch

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_6

emerge when studying linked-life domains, another core principle of the life-course paradigm. For instance, one may be interested in estimating the effect of past family trajectories on the chances of obtaining a promotion among female managers.

From a methodological point of view, some event history models have tackled this issue by including a few summary indicators of the past trajectory (e.g., time spent in a given state, or a dummy variable indicating whether a specific event has already occurred or not) (Blossfeld et al. 2007). This approach allows estimation of the effect of these past trajectory indicators on the chances of experiencing the event under study. However, this process is often limited as it might fail to identify the key dimensions of the previous trajectory affecting the event. First, these key dimensions might depend on the trajectories themselves. In that case, it becomes spurious to decide a priori which relevant past trajectory indicators should be included in the model. Second, trajectories represent complex objects with many different dimensions. It might therefore be difficult to identify the most relevant ones. For instance, life-course scholars stress the importance of three sub-dimensions, each requiring several indicators to be included in the analysis. These indicators refer to the timing, the ordering, and the duration of states and of transitions (Scott and Alwin 1998). Finally, there might be many interaction effects between these sub-dimensions, making the selection of relevant indicators of past trajectories even more difficult.

In this article, we develop an innovative method combining Sequence Analysis and Event History Analysis which we call Sequence History Analysis (SHA). Its aim is to tackle the aforementioned methodological challenges and the method works in two steps. We start by identifying typical past trajectories of individuals over time by using Sequence Analysis. As trajectories are considered as a progression over time, where events and life stages accumulate, individuals are likely to move from one cluster to another over time. We then estimate the effect of these typical past trajectories on the event under study using discrete-time models. SHA is presented in the first part of this paper.

In the second part of this article, we use the proposed methodological approach in an original study of the effect of past childhood co-residence structures on the chances of leaving the parental home in Switzerland. In western countries where nuclear families and neolocality prevail, the departure from the parental home is a crucial step and an indicator of the transition to adulthood. It is often a prerequisite to achieving other family life transitions, such as co-residency and becoming a parent (Mulder 2009; Schizzerotto and Lucchini 2004). Furthermore, the departure from the parental home has significant consequences for important policy areas, such as the demand for housing (Ermisch and Di Salvo 1997) and the risk of poverty among young people (Iacovou and Aassve 2007). In this context, identifying the determinants of the early departure from the parental home of young adults is of prime interest. Among these determinants, many sociological theories stress the importance of family configuration, as well as the whole individual trajectory preceding home-leaving. This is a key concern in Switzerland where the

number of divorces has experienced a strong increase over the past 40 years (Swiss Federal Statistical Office 2016), with a divorce rate that reached 52.6% in 2005. A significant number of studies showed the impact of lone- and step-parenthood on early departure (Holdsworth 2000; Bernhardt et al. 2005). Consequently, some studies focused on the co-residence structure in which a young adult lived at a specific moment of his/her life, often at the time of the youth's final home-leaving (Mitchell et al. 1989; Chiuri and Del Boca 2010). However, few studies looked back at the effect of the whole co-residence trajectory. This is mostly due to the lack of detailed life history records of co-residence structures during childhood (Aquilino 1991; Goldscheider and Goldscheider 1998; Blaauboer and Mulder 2010) and to the lack of a proper methodological framework to estimate the influence of early co-residence trajectories on the departure from the parental home.

The structure of the article is as follows. First, we will present the methodological features of SHA. We will then apply it empirically using social science data. Based on data from the LIVES Cohort Study (Elcheroth and Antal 2013), our analyses showed that it is not only the occurrence of an event such as parental divorce that increases the risk of leaving home, but also the order in which changes to the preceding family co-residence structure occurred. Two features have a significant influence on leaving home: the co-residence structure itself and the arrival or departure of siblings from the parental home.

1.1 Sequence History Analysis: A Combination of Sequence Analysis and Event History Analysis

Several methods are available to estimate the effects of a set of covariates on the hazard rate of a given event. This approach uses a discrete-time representation of the data: the so-called person-period file (Allison 2014). In this format, one observation is generated for each individual i at each time point t. Since the time t is assumed to be observed on a discrete scale, a finite set of observations is generated. The time ranges from the start of observation (typically 0) until the end of the observation period of the ith individual.

Before going into details, let us present a small example that will help us to clarify the presentation of a person-period file. In this example, (Table 1), we are interested in estimating the effect of past cohabitation trajectories on the chances of leaving the parental home among a cohort of young adults. The individual 72 is a woman. She left home after 6 time periods. She also has the following cohabitation trajectory: BP-BS-BS-BS-LS-LS where BP stands for biparental household, BS for biparental household and siblings, and LS for lone-parent household and siblings. The corresponding person-period file therefore reads as follows:

Table 1 Example of a person-period file

ID	Time	Departure from the parental home	Cohabitation status
72 (Woman)	15	0	BP
72 (Woman)	16	0	BS
72 (Woman)	17	0	BS
72 (Woman)	18	0	BS
72 (Woman)	19	0	LS
72 (Woman)	20	1	LS

1.2 Sequence History Analysis: Operationalizing Previous Trajectories

There are two different interpretations of the aim of Sequence Analysis. It may be seen as a way to identify ideal-typical trajectories. It can also be considered as an effective means to reduce the complexity of trajectories into a few main types of sequences. Both approaches are interesting in our context, because one might typically expect to observe many different individual past trajectories. Sequence Analysis can therefore be used to operationalize the concept of past trajectories by reducing this complexity or as a way to identify ideal-typical past trajectories.

Since the 1990s–2000s, the research trend in Sequence Analysis has been structured around a core program including a limited number of methodological options (Gauthier et al. 2014). Generally, Sequence Analysis works in three steps. First, trajectories are coded as sequences of states. Second, the distances between each pair of sequences are computed and gathered into a distance matrix. Finally, a cluster analysis is conducted on this matrix. It gathers together similar sequences while separating dissimilar sequences. The result is a categorical covariate that can be used in subsequent analyses. Let us briefly discuss these three steps in our case.

In the first step, we rebuild the past trajectory at each time point, i.e., for each observation of each individual i at time t. Taking our previous example a step further, Table 2 presents two ways of modeling past family trajectories. First, rebuilding the past trajectory at each time point for each individual in our person-period file is done by considering, at each time point t, the trajectory leading to the current position. As such, the length of the trajectory logically increases by one for each additional time unit. Sometimes, we are only interested in the previous trajectory, excluding the present state. Thus, the last column reconstructs for each individual i at each time point t, past trajectory until $t - 1$. These past trajectories can therefore be interpreted as past trajectories until all possible present times. There are, thereby, t trajectories of varying lengths for each individual. Indeed, since the duration from the starting time is not the same at each time point, the past trajectories considered grow over time.

In a second step, we need to choose a distance measure to conduct Sequences Analysis. This measure defines which criteria should be taken into account to

Table 2 Two different ways of reconstructing past trajectories

ID	Time	Departure from the parental home	Past trajectory	Past trajectory excluding present
72	15	0	BP/15	BP/14
72	16	0	BP/15-BS/1	BP/15
72	17	0	BP/15-BS/2	BP/15-BS/1
72	18	0	BP/15-BS/3	BP/15-BS/2
72	19	0	BP/15-BS/3-LS/1	BP/15-BS/3
72	20	1	BP/15-BS/3-LS/2	BP/15-BS/3-LS/1

Table 3 Creation of a typology of past trajectories

ID	Time	Departure from the parental home	Past trajectory excluding present	Typology
72	15	0	BP/14	Biparental household
72	16	0	BP/15	Biparental household
72	17	0	BP/15-BS/1	Early arrival of siblings
72	18	0	BP/15-BS/2	Early arrival of siblings
72	19	0	BP/15-BS/3	Early arrival of siblings
72	20	1	BP/15-BS/3-LS/1	From biparental to lone-parent household (with siblings)

compare two trajectories. According to Studer and Ritschard (2016), the choice of a dissimilarity measure should be based on its sensitivity to timing, sequencing, or duration. For instance, if age is thought to be an important property of the past trajectory, one should emphasize timing. This would be interesting if age at parental divorce is believed to be of key importance. Conversely, if we want to focus on the path, i.e., the states through which an individual goes, a distance measure sensitive to sequencing should be chosen. Finally, if the time spent in each state is important, a distance measure sensitive to duration should be used.

In a third step, after having computed the distances between sequences, a typology is built using cluster analysis. This step results in a categorical covariate in our person-period file. Taking back our previous example and assuming that the cluster analysis identified three groups: (1) "Biparental household", (2) "Arrival of siblings", and (3) "From biparental to lone-parent household (with siblings)," our person-period file would be read as shown in Table 3. It can be noted that a given individual can belong to different clusters over time. In other words, the type of past trajectory an individual belongs to may change over time. It stems from the fact that our unit of analysis is one person-period, not one individual. In the next step, we will use this information to estimate the effect of a past trajectory on the event under study.

In these last two steps, we analyze sequences of different lengths.[1] Sequence Analysis should not be used to analyze sequences of different length when this difference results from incomplete or censored data, because Sequence Analysis implicitly assumes that the processes under study are fully observed.[2] However, in our case, the differences in sequence length do not result from incomplete data. They result from meaningful differences in the length of the process leading to the current situation. However, as the length of the previous trajectory and age are often closely related, we strongly recommend to always control for age.

1.3 Event History Analysis: Estimating the Effect of Typical Past Trajectories on the Event Under Study

Event history analysis is an suitable tool to understand how events are produced and how they are conditioned by other explanatory variables, which may or may not vary over time (Allison 2014). As such, once the typology of previous trajectories is created, we propose to estimate its effect on a given event using a discrete-time event history model. More precisely, at each time of observation since the starting time, the trajectory is introduced as a time-varying covariate. Consequently, applying Sequence History Analysis to our previous example, we could estimate the chance of obtaining a promotion according to the type of previous family trajectories.

Two factors should be taken into consideration when specifying the model. First, our typology of past trajectories might be linked to their length. If this is the case, we recommend adding the length of the previous trajectory or a transformation of it to the model. Consequently, the effect of the typology will no longer be related to the length of the trajectories, which could lead to misleading interpretations.[3]

Second, two interpretations of the past trajectories' effect can be made. It could be related to the ability of the typology to summarize the main information of the current situation. For instance, the effect of the past family trajectory on the chances of receiving a promotion could be related to the characteristics of the current family situation, such as being married or having a child at time t. It could also be linked to the individual history, such as the age when the individual got married or if he/she was married before having a child or not. We can distinguish these two situations by adding simple indicators of the current situation to the model. If the effect of the past trajectory types remains significant, one may conclude that "individual history

[1]The length of the previous trajectories typically depends on age.

[2]By clustering incomplete sequences, we often end up with one (or several) clusters of incomplete trajectories, which cannot be interpreted. If this is not the case, we implicitly predict the end of the sequence, which might also be problematic.

[3]From a general point of view, even though this is not specific to the proposed methodology, it is generally recommended to add some timing information to the model as the hazard rate is usually not constant over time.

matters" for the issue under study. Aside from these two kinds of information, the usual control variables should be added to the model. The latter are research-specific and we therefore do not discuss them in more detail.

To sum up, we propose a methodological approach to estimate the effect of a previous trajectory on an upcoming event. This methodology functions in three steps: (1) building previous trajectories in a person-period file, (2) running Sequence Analysis, and (3) estimating the effect of typical past trajectories on an upcoming event using a discrete-time model. After having presented this methodology, we now turn to its application. Therefore, we provide an empirical example displaying the effect of childhood co-residence trajectories on the risk of leaving home. The analyses will be based on life history calendar data.

2 Empirical Application: Childhood Co-residence Trajectories and Leaving Home

In this section, we apply Sequence History Analysis to assess the influence of childhood co-residence trajectories on the probability of leaving the parental home in Switzerland. Previous research has demonstrated the multiple effects of previous co-residence trajectories on the departure from the parental home (Mitchell et al. 1989; Aquilino 1991; Sandefur et al. 2008; Blaauboer and Mulder 2010). There are also some reasons to believe that the number of siblings living in the same household is likely to affect the probability of young adults leaving the parental home (Mitchell et al. 1989; Aquilino 1991; Gierveld et al. 1991; Avery et al. 1992; Buck and Scott 1993).

First and foremost, growing up with two biological parents—which is still the most common form of living arrangements in Western Europe—is linked with closer family bonds and longer stays in the parental home (Mitchell et al. 1989; Aquilino 1991; Mitchell 1994; Goldscheider and Goldscheider 1998).

Second, several studies showed that children of divorced parents tend to leave the parental home earlier than those of intact families (Goldscheider and Goldscheider 1998; Cherlin et al. 1995; Juang et al. 1999; Holdsworth 2000; Bernhardt et al. 2005; Zorlu and Gaalen 2016). As noted by several authors, this effect might be more related to low family socio-economic background than to the absence of one of the parental figures (Bianchi 1987; Mitchell et al. 1989; McLanahan and Carlson 2004; Kiernan 2006, for instance). Aquilino (1991) showed that young adults who grew up in a single-parent household from birth do not have a higher hazard of leaving home than those who grew up in an intact family. Therefore, the stability of co-residence structure could also have an impact on the timing of leaving home.

Third, children from step-parent families tend to leave home earlier than young adults from intact families (Mitchell et al. 1989; Aquilino 1991; Kiernan 1992; Goldscheider and Goldscheider 1998). Among various explanations, Goldscheider and Goldscheider (1998) stress the difficulty of welcoming a new parental figure,

step-siblings, and/or half-siblings into one's home. Other studies have shown that severe conflicts and disagreements within step-families play a significant role in early nest-leaving (Gaehler and Bernhardt 2000; Gossens 2001).

Fourth, there might be some circumstances in which both intact and non-intact families may no longer be able to maintain their households. In such situations, both children and parents might seek shelter in someone else's household, in most cases in the houses of grandparents (Aquilino 1991). This type of family arrangement is often referred to as "extended family." Therefore, as having to move back in with relatives is usually the result of financial difficulties, such situations might push children to get a job and establish an independent household earlier.

There is some evidence that having siblings might also influence the departure from the parental home. Individuals with many siblings were found to have a higher likelihood of leaving home (Mitchell et al. 1989; Aquilino 1991; Gierveld et al. 1991; Avery et al. 1992; Buck and Scott 1993). This may be explained by the fact that individuals who grow up with a large number of siblings have a higher risk of feeling "overcrowded" in their parental home and of suffering from a lack of physical space for privacy. First-born children have a higher risk of leaving home at an earlier age than any other children, except if they are an only child (Bianchi 1987). Indeed, Holdsworth (2000) has shown that an only child will tend to stay longer at home in order to take care of their parents.

3 Data

We use data from the LIVES Cohort Study (FORS and NCCR LIVES 2015), a panel survey whose first wave was conducted from mid-October 2013 to the end of June 2014 (Elcheroth and Antal 2013). The sample includes 1691 respondents, of whom 415 were Swiss and 1276 were from a foreign background. The sample is composed of people aged 15–24 on January 1st 2013 and who began in a Swiss school before the age of 10. Second-generation immigrants are over-represented in the sample and particular attention is paid to offspring of low- or middle-skilled migrants who mainly hail from Southern Europe or from the Balkan Peninsula. The sampling design of the survey is quite innovative in the sense that it combines a stratified random sample with two iterations of controlled network sampling, with random selection within the personal networks (Brändle 2017, for more details).

The life history calendar allows for the collection of detailed information regarding the co-residence trajectories of each respondent. This information was recoded into five statuses—combining information on parents and siblings—namely living with: (a) both parents (without siblings), (b) both parents and sibling(s), (c) one parent (without siblings), (d) one parent and sibling(s), and (e) other relatives. Unfortunately, living with a step-parent could not be distinguished from the "one parent" situation, since this information was not available in the survey.

The timing of the departure from the parental home was operationalized as the first episode in which an individual does not live with his/her parents anymore.

We assumed that individuals are at risk of leaving home from the age of 15. Consequently, we identified 147 events (i.e., departures from the parental home) for 1637 individuals.[4]

3.1 Control Variables

Several control variables were introduced into the model, such as age (logarithm), sex, ethnic origin (based on the mother's country of birth), place of residence at 14 years old, labor market integration (apprenticeship included)[5] and family socio-economic background.[6] This last is a key control variable since it has been argued that the effect of living with a single parent is linked to differences in economic conditions. However, the large number of missing values for family socio-economic background (70%) forced us to run a model without it. Finally, two additional control variables were included: the occurrence of parental disruption and the presence of siblings. These variables were included to verify that the effect of the previous trajectory, as measured through our method, is not only related to the current situation of the household.

4 Analysis

Sequence History Analysis works in three steps. We start by recoding all past trajectories in a person-period file. We then conduct a Sequence Analysis on these recoded past trajectories. Finally, using a discrete-time model, we estimate the effect of these past trajectories on the probability of leaving home. Let us present these steps in more detail using our empirical example.

[4]This means that 55 individuals had missing data in at least one of the variables.

[5]The Swiss education system is largely an apprenticeship-based system of education (Thomsin et al. 2004). Almost two thirds of every cohort of students attend a vocational and training program (VET) (Swiss Federal Statistical Office 2015) In principle, an apprenticeship contract is signed between the apprentices and an approved firm in which the former will spend about two thirds of their time following a practical vocational training. The rest of the time is spent in a vocational school.

[6]In most cases, the occupation of the father when the respondent was 15 was taken as the benchmark to define the family's socio-economic background. When this information was unavailable, the occupation of the mother when the respondent was 15 was used.

4.1 Sequence Analysis: Operationalizing Previous Co-residence Trajectories

Here, we consider previous co-residence trajectories for each individual i at each time t, from birth until the current age of each individual at the time of the survey. Since we are interested in the hazard of leaving home after the age of 15, the past trajectories have a starting length of 14 for all individuals. The final length of the past trajectory corresponds to the occurrence of the departure from the parental home minus one time unit (cf. Table 2, past trajectories excluding present state) or to the end of the observation period.

We then ran Sequence Analysis on these past trajectories excluding the present state to identify ideal-types of past trajectories. All past trajectories were included in the analyses at the same time. The use of Sequence Analysis involves choosing an appropriate distance measure and a clustering algorithm. We are interested in estimating the effect of a previous family history on the departure from the parental home. This previous history is strongly linked to the order of the stages through which an individual goes. We have therefore chosen a distance sensitive to differences in sequencing. When sequencing is of central interest, Studer and Ritschard (2016) suggest using optimal matching on sequences of distinct successive states (DSS). In our case, we used a constant substitution cost of 2 and an indel cost of 1. The DSS is obtained by considering the succession of states without considering the duration of each state. For instance, the sequence "S-S-M-M-M" is recorded "S-M." We therefore focus only on sequencing (the information about timing and duration is not used).

We then clustered the sequences in the following way. Two specific groups were created "manually" to match a precise definition: two whole trajectories spent with both parents either with or without siblings. Each resulting group represents respectively 40.5% and 4.1% of the sample. Although small, this latter group (i.e., only child with both parents) is relevant as being an only child is expected to have a significant influence on the risk of leaving the parental home. We then used the PAM ("Partitioning Around Medoids") algorithm to cluster the remaining past trajectories. This algorithm aims to obtain the best partitioning for a data set into a predefined number k of groups (Kaufman and Rousseeuw 2005; Studer 2013). Based on the best average silhouette width, we kept six groups. The result is a final typology of eight groups (the two manually-constructed groups and the six clusters on the remaining sequences).

All statistical analyses conducted in this article use the R Software and environment (R Core Team 2016), along with the TraMineR package (Gabadinho et al. 2011) for sequence analysis and the WeightedCluster package (Studer 2013) for cluster analysis. The final typology of co-residence trajectories is presented in Fig. 1. When we observe these ideal-types of past trajectories, we see that the percentages presented are based on the person-period trajectories and that individuals can switch between clusters over time. For instance, we expect that some individuals will start by being in the cluster "Both parents and siblings" before going to the cluster "Late departure of siblings."

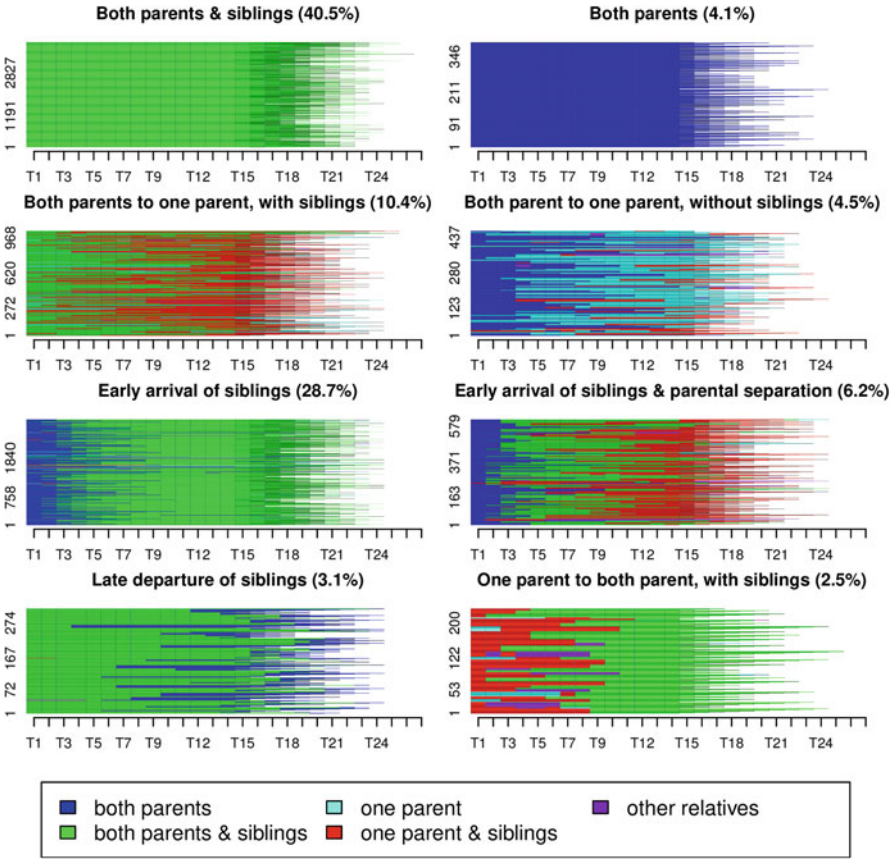

Fig. 1 Trajectories of past co-residence structures

- Both parents and siblings (40.5%)—Trajectories spent entirely with both parents and siblings. As this cluster represents the most common trajectory, it was used as the reference category in the regression models.
- Early arrival of siblings (28.7%)—Trajectories of oldest children who experienced the arrival of younger siblings during their early childhood.
- Both parents to one parent (with siblings) (10.4%)—Trajectories characterized by a transition from a biparental to a lone-parent household, in both cases in the presence of siblings.
- Early arrival of siblings and parental separation (6.2%)—Trajectories of older siblings who experienced the arrival of younger siblings during their adolescence and a subsequent parental disruption.
- Both parents to one parent (without siblings) (4.5%)—Trajectories characterized by a parental disruption without siblings.

- Both parents (4.1%)—Trajectories spent entirely with both parents, but without siblings.
- Late departure of siblings (3.1%)—Trajectories characterized by the departure of siblings. These trajectories are probably those of younger children.
- One parent to both parents (with siblings) (2.5%)—Trajectories initiated by a co-residency with one parent only and siblings before the second parent joins the household later. This might be a typically common trajectory of a migrant family. Fathers first migrate alone, leaving their wives and children behind. After a few years, mothers and children will also migrate, reuniting the family. Consequently, children will live their first years in a "lone-parent household" before moving to a biparental household.

4.2 Event History Analysis: Estimating the Effect of Typical Past Trajectories on the Event Under Study

After having identified the previous typology of trajectories, we estimate their effect on the risk of leaving home using a discrete-time model (cf. Table 4). We do it by running a logistic regression on the person-period file. Individuals with missing data were not included in the models (4% in models 1,2, & 3 and 70% in model 4).

Four models were estimated. The first model includes the past trajectories and the control variables. In the second model, aggregated indicators of parental divorce and presence of sibling(s) were included to assess whether the effect of past trajectories remains significant in the presence of these aggregated indicators. The third model is computed without the past trajectories to estimate its statistical power. Finally, the last model is composed of the past trajectories, the control variables, and the family socio-economic background factor.

A first sign of the overall importance of the past trajectory covariate can be asserted by looking at the Bayesian Information Criterion (BIC).[7] We decided to use this criterion because it is the most conservative and penalizes complexity more than the Akaike Information Criterion (AIC). According to the BIC, the first model is the most parsimonious. The BIC value of the second model is also very close to that of the first model. Both models include the ideal-typical past trajectories. From a statistical point of view, there is therefore an added value to including the past ideal-typical trajectories in the model.

[7] $\mathrm{BIC} = -2\ln(L) + \ln(N) * k$, where L is the likelihood, $-2\ln(L)$ is equal to the deviance, ln is the logarithm, and k represents the number of parameters (i.e., coefficients). Raftery (1995) argues that the N can be estimated in three different manners when it is used for event history models: the number of observations (person-period), the number of individuals, or the number of events. According to the recommendations made in this article, we used the last option which is the least conservative. This option is also coherent with the calculation of the BIC for survival continuous-time models (i.e., Cox models) in which N represents the number of observed events.

Table 4 Logit models predicting probability of first home-leaving

	Model 1	Model 2	Model 3	Model 4
Intercept	-9.79 (0.67) ***	-10.79 (0.77) ***	-9.20 (0.66) ***	-10.76 (1.10) ***
Co-residence configurations:				
Both parents & siblings (ref.)				
Both parents	0.09 (0.56)	0.96 (0.65)		-0.05 (0.83)
Late departure of siblings	1.05 (0.35) **	1.91 (0.48) ***		0.92 (0.49) +
Early arrival of siblings	0.34 (0.25)	0.40 (0.26)		0.75 (0.35) *
Both parents to one parent (without siblings)	1.62 (0.34) ***	2.06 (0.45) ***		2.59 (0.53) ***
Early arrival of siblings & parental separation	0.66 (0.35) +	0.66 (0.48)		0.32 (0.60)
One parent to both parents (with siblings)	0.32 (0.58)	0.37 (0.59)		0.71 (1.13)
Both parents to one parent (with siblings)	0.80 (0.28) **	1.01 (0.30) ***		1.01 (0.40) *
Age (ln)	3.10 (0.28) ***	3.14 (0.29) ***	3.12 (0.28) ***	2.96 (0.42) ***
Women	0.52 (0.18) **	0.54 (0.18) **	0.44 (0.18) *	0.83 (0.27) **
Ethnic origin: Switzerland (ref.)				
Eastern Europe	-0.99 (0.26) ***	-1.02 (0.26) ***	-1.10 (0.26) ***	-0.75 (0.43) +
South-Western Europe	-0.88 (0.29) **	-0.88 (0.27) **	-0.93 (0.28) ***	-0.98 (0.42) *
North-Western Europe & North America	0.69 (0.33) *	0.74 (0.33) *	0.61 (0.33) +	0.56 (0.51)
Other continents	-0.29 (0.31)	-0.24 (0.30)	-0.15 (0.30)	-0.11 (0.44)
Labor market integration	0.52 (0.22) *	0.56 (0.22) *	0.51 (0.22) *	0.89 (0.31) **
Place of residence: Large population centers (ref.)				
Periurban & metropolitan centers	0.22 (0.34)	0.19 (0.35)	0.11 (0.34)	0.23 (0.50)
Touristic municipalities	0.48 (0.51)	0.39 (0.51)	0.46 (0.50)	0.58 (0.65)
Middle- and small-sized population centers	0.17 (0.22)	0.09 (0.23)	0.13 (0.22)	0.62 (0.33) +
Periurban & commuting municipalities	0.26 (0.36)	0.35 (0.36)	0.07 (0.35)	0.44 (0.56)
Outlying municipalities	0.31 (0.28)	0.24 (0.28)	0.19 (0.28)	0.56 (0.39)
Underrepresented places of birth	-0.19 (0.28)	-0.14 (0.28)	-0.29 (0.26)	-0.14 (0.46)
Divorce		0.09 (0.37)	0.54 (0.23) *	
Siblings		0.89 (0.33) **	-0.03 (0.22)	
Family socioeconomic status Qualified manual professions (ref.)				
Top management				-0.35 (0.81)
Academic professions & senior management				1.11 (0.47) *
Liberal professions				0.47 (0.49)
Other self-employed				-0.14 (0.60)
Intermediate professions				0.27 (0.52)
Skilled non-manual professions				0.88 (0.38) *
Unskilled non-manual & manual professions				0.48 (0.50)
Nb obs.	8700	8700	8700	3456
Nb ind.	1624	1624	1624	506
Nb events	142	142	142	77
Deviance	1096.1	1088	1118.2	550.23
BIC	1200.2	1201.9	1212.4	724.57

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In all models, most coefficients of past co-residence trajectories are significant and go in the same direction. While the differences between the trajectories "both parents" and "both parents with siblings" are not significant, we observe a significant effect concerning the changes in the presence of siblings. Staying in the parental home after the departure of siblings increases the probability of leaving home. The "early arrival of siblings" also increases the hazard rate of leaving home. Besides, living with only one parent—with or without siblings—is associated with a higher risk of leaving home. There is no significant difference between the categories "early arrival of siblings and parental separation" and those who grew up with two parents. The same applies for young individuals who started by living in a lone parent household with siblings before moving to a bi-parental household.

We can see in Table 4 that the types of past co-residence trajectories are more informative than the aggregated indicators of parental divorce and of the presence of siblings. Indeed, in the second model that includes covariates related to these two aspects, the effects of childhood co-residence trajectories remain significant. These results show that behind the effect of having siblings or of having experienced a parental divorce, the past trajectory, i.e., the personal history, does matter.

The effect of control variables confirms our hypotheses. The hazard rate of leaving the parental home rises with increasing age. The departure from the parental home is also significantly influenced by the ethnic origin. Second-generation immigrants from Eastern or South-Western countries are less likely to leave home than Swiss natives. Conversely, having a Northern-Western European or a North-American background increases the risk of leaving home. Obtaining a first job significantly increases the likelihood of leaving home. Women are more likely to leave the parental home than men. However, we did not find a significant effect of the place of residence. Residents of middle and small centers have a higher probability of leaving home than inhabitants of large population centers, but this effect is only significant at the 0.1 level. Lastly, respondents whose parents had an academic profession or a senior management position when they were adolescents leave home more often than those whose parents had a qualified and manual profession. Children of skilled and non-manual workers are also more likely to leave the parental home.

5 Discussion

The occurrence of divorce does not play a strong role in the departure from the parental home, nor does that of having siblings. Conversely, childhood co-residence patterns influence the ways in which young adults leave the parental home. More precisely, the occurrence, the timing, and the sequencing of the events have a specific effect on home-leaving. For instance, two features have a significant and strong impact on the departure from the parental home: the lone parenthood configuration and the arrival and departure of siblings.

Having spent some years in a lone-parent household has a positive impact on the risk of leaving home. It does not seem to matter much if it occurred in the presence of siblings or not, as both situations lead to an increase in the likelihood of leaving home. In addition, these effects remained significant when controlling for the family socio-economic background. In other words, the negative effect of lone parenthood is not only explained by lower socio-economic background.

Having siblings matters when it comes to leaving the parental home. However, behind that simple fact, our method showed that birth order and arrivals or departures of siblings matter more. Moreover, when the family socio-economic background is taken into account, being an only child significantly increases the odds of leaving the parental home. Additionally, the departure from the parental home of siblings (most probably older siblings) encourages the remaining siblings to leave home. This could be interpreted as an imitation of the first-to-go individuals' behavior.

In this study, we measured a link between childhood co-residence structures and departure from the parental home. However, the underlying mechanism linking these two concepts was not explained. For example, some longitudinal information regarding the relational quality in households was not available and could therefore not be included in the model. Moreover, the respondents were quite young. Consequently, only a small proportion of them had left the parental home at the end of the observation period. Hence, the observed effects could be mainly related to differences among early home-leavers.

6 Conclusion

The aim of this paper was two-fold. First, we proposed a methodological framework to estimate the effect of an unfolding trajectory on an upcoming event. We then applied the proposed approach to an original study of the effect of past co-residence trajectories on the departure from the parental home in Switzerland.

The results obtained with the combination of Sequence Analysis and Event History analysis provided results that would not have been obtained if each method had been used separately. However, the combination of Sequence Analysis and Event History Analysis was not an easy task as these two methods are based on very different approaches to life-course data. Sequence Analysis is based on a holistic approach of life course trajectories. The overall trajectory of each respondent is examined and compared with that of the other individuals. Consequently, Sequence Analysis aims to investigate the progression of individuals over their life course. Conversely, the focus in Event History Analysis is rather the investigation of the probabilistic distribution of life course events over time according to individual characteristics (Tuma and Hannan 1984; Mayer and Tuma 1990; Courgeau and Lelièvre 1992). The aim of Sequence History Analysis is thus to resolve this conflict between the two approaches. The proposed framework may have a much broader field of application. For instance, it could allow us to study how previous

professional trajectories are linked with the risk of dying at each age. We believe that Sequence History Analysis is a very promising tool. This method allows the combining of two traditions of investigating longitudinal data in life-course research: the holistic approach of Sequence Analysis and the processual approach of Event History Analysis.

Further work is needed to develop this approach more fully, to address its weaknesses and build on its strengths. For example, the proposed framework does not allow the drawing of causal interpretations of the results. We cannot state that parental divorce "causes" early departure from the parental home. In this respect, quasi-experimental designs, such as propensity-score matching, could be an interesting lead to follow. It would also be interesting to know to what extent the individuals have experienced relocation in their past. It might make young adults more likely to move out of the parental home and less hesitant about setting up their own independent households. Consequently, if possible, further analysis should take this factor into account.

Acknowledgements This paper 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 (Grant number: 51NF40-160590).

References

Allison, P. D. (2014). *Event history and survival analysis* (Vol. 46). Los Angeles: Sage Publications.

Aquilino, W. S. (1991). Family structure and home-leaving: A further specification of the relationship. *Journal of Marriage and the Family, 53*(4), 999–1010.

Avery, R., Goldscheider, F., & Speare, A. (1992). Feathered nest/gilded cage: Parental income and leaving home in the transition to adulthood. *Demography, 29*(3), 375–388.

Bernhardt, E., Gähler, M., & Goldscheider, F. (2005). Childhood family structure and routes out of the parental home in Sweden. *Acta Sociologica, 48*(2), 99–115.

Bianchi, S. M. (1987). Living at home: Young adults' living arrangements in the 1980s. In *Annual Meeting of the American Sociological Association, Chicago*.

Blaauboer, M., & Mulder, C. H. (2010). Gender differences in the impact of family background on leaving the parental home. *Journal of Housing and the Built Environment, 25*(1), 53–71.

Blossfeld, H.-P., Golsch, K., & Rohwer, G. (2007). *Event history analysis with Stata*. New York: Psychology Press.

Brändle, K. (2017). The geography of social links among a young cohort in Switzerland. *LIVES Working Papers* (58), 1–33.

Buck, N., & Scott, J. (1993). She's leaving home: But why? An analysis of young people leaving the parental home. *Journal of Marriage and the Family, 55*(4), 863–874.

Cherlin, A. J., Kiernan, K. E., & Chase-Lansdale, P. L. (1995). Parental divorce in childhood and demographic outcomes in young adulthood. *Demography, 32*(3), 299–318.

Chiuri, M. C., & Del Boca, D. (2010). Home-leaving decisions of daughters and sons. *Review of Economics of the Household, 8*(3), 393–408.

Courgeau, D., & Lelièvre, E. (1992). Analyse des données biographiques en démographie. In L. Coutrot & C. Dubar (Eds.), *Cheminements professionnels et mobilités sociales* (pp. 59–70). Paris: La Documentation Française.

Eerola, M. (2018). Case studies of combining sequence analysis and modelling. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Eerola, M., & Helske, S. (2016). Statistical analysis of life history calendar data. *Statistical Methods in Medical Research, 25*(2), 571–597.

Elcheroth, G., & Antal, E. (2013). Echantillon de cohorte LIVES. Première vague. Technical report, University of Lausanne, Lausanne.

Ermisch, J., & Di Salvo, P. (1997). The economic determinants of young people's household formation. *Economica, 64*(256), 627–644.

FORS and NCCR LIVES. (2015). LIVES Cohort Panel, Wave 1. Data available at URL: http://forscenter.ch/fr/our-surveys/swiss-household-panel/datasupport/telecharger-les-donnees-4/cohort-w1/

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.

Gaehler, M., & Bernhardt, E. (2000). The impact of parental divorce, family reconstitution, and family conflict on nest-leaving in Sweden (pp. 6–8). Rostock: Max Planck Institut für Bevölkerungsforschung.

Gauthier, J.-A., Bühlmann, F., & Blanchard, P. (2014). Introduction: Sequence analysis in 2014. In Ph. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 1–19). New York: Springer.

Gierveld, J. D. J., Liefbroer, A. C., & Beekink, E. (1991). The effect of parental resources on patterns of leaving home among young adults in the Netherlands. *European Sociological Review, 7*(1), 55–71.

Goldscheider, F. K., & Goldscheider, C. (1998). The effects of childhood family structure on leaving and returning home. *Journal of Marriage and the Family, 60*(3), 745–756.

Gossens, L. (2001). Transition to adulthood: Developmental factors. In M. Corijn & E. Klijzing (Eds.), *Transition to Adulthood* (pp. 27–42). The Netherlands: Springer.

Holdsworth, C. (2000). Leaving home in Britain and Spain. *European Sociological Review, 16*(2), 201–222.

Iacovou, M., & Aassve, A. (2007). *Youth poverty in Europe*. New York: Joseph Rowntree Foundation.

Juang, L. P., Silbereisen, R. K., & Wiesner, M. (1999). Predictors of leaving home in young adults raised in Germany: A replication of a 1991 study. *Journal of Marriage and the Family, 61*(2), 505–515.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Hoboken: Wiley.

Kiernan, K. (1992). The impact of family disruption in childhood on transitions made in young adult life. *Population Studies, 46*(2), 213–234.

Kiernan, K. (2006). Non-residential fatherhood and child involvement: Evidence from the Millennium Cohort study. *Journal of Social Policy, 35*(4), 651–669.

Lundevaller, E. H., Vikström, L., & Haage, H. (2018). Modelling mortality using life trajectories of disabled and non-disabled individuals in 19th-century Sweden. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Madero-Cabib, I., Gauthier, J.-A., & Le Goff, J.-M. (2015). The influence of interlocked employment–family trajectories on retirement timing. *Work, Aging and Retirement, 2*(1), 38–53.

Mayer, K. U., & Tuma, N. B. (1990). *Event history analysis in life course research*. Madison: The University of Wisconsin Press.

McLanahan, S., & Carlson, M. S. (2004). Fathers in fragile families. In M.E. Lamb (Ed.), *The role of the father in child development* (Vol. 4, pp. 368–396). Hoboken: Wiley.

Mitchell, B. A. (1994). Family structure and leaving the nest: A social resource perspective. *Sociological Perspectives, 37*(4), 651–671.

Mitchell, B. A., Wister, A. V., & Burch, T. K. (1989). The family environment and leaving the parental home. *Journal of Marriage and the Family, 51*(3), 605–613.

Mulder, C. H. (2009). Leaving the parental home in young adulthood. In A. Furlong (Ed.), *Handbook of youth and young adulthood: New perspectives and agendas* (pp. 203–210). New York: Routledge.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–164.

Sandefur, G. D., Eggerling-Boeck, J., & Park, H. (2008). Off to a good start? Postsecondary education and early adult life. In R.A. Settersten Jr., F.F. Furstenberg Jr. & R.G. Rumbaut (Eds.), *On the frontier of adulthood: Theory, research and public policy* (pp. 292–319). Chicago: University of Chicago Press.

Schizzerotto, A., & Lucchini, M. (2004). Transitions to adulthood. In R. Berthoud & M. Iacovou (Eds.), *Social Europe: Living standards and welfare states* (pp. 46–68). Cheltenham: Edward Elgar.

Scott, J., & Alwin, D. (1998). Retrospective versus prospective measurement of life histories in longitudinal research. In J.Z. Giele & G.H. Elder Jr. (Eds.), *Methods of life course research: Qualitative and quantitative approaches* (pp. 98–127). California: Sage Publications.

Studer, M. (2012). *Étude des inégalités de genre en début de carrière académique à l'aide de mèthodes innovatrices d'analyse de donnèes sèquentielles*. Ph.D. thesis, University of Geneva, Geneva.

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers*(24) (pp. 1–32).

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society A, 179*(2), 481–511.

Swiss Federal Statistical Office. (2015). Secondary education II – Synoptic table. Retrieved from http://www.bfs.admin.ch/bfs/portal/fr/index/themen/15/04/00/blank/uebersicht.html, 2015-09-15.

Swiss Federal Statistical Office. (2016). Divorces and divortiality since 1876. Retrieved from: https://www.bfs.admin.ch/bfs/en/home/statistics/population/marriages-partnerships-divorces/divortiality.assetdetail.325774.html, 2017-12-04.

Thomsin, L., Le Goff, J.-M., & Sauvain-Dugerdil, C. (2004). Genre et étapes du passage à la vie adulte en Suisse. *Espace Populations Sociétés. Space Populations Societies, 11*(1), 81–96.

Tuma, N. B., & Hannan, M. T. (1984). *Social dynamics models and methods*. Orlando: Academic Press.

Zorlu, A., & Gaalen, R. (2016). Leaving home and destination of early nest leavers: Ethnicity, spaces and prices. *European Journal of Population, 32*(2), 1–25.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part III
The Sequence Network Approach

Network Analysis of Sequence Structures

Benjamin Cornwell

1 From Sequence Pathways to Sequence-Networks

Many social sequences can be analyzed as independent linear chains of events. For example, a given firm's trajectory of growth can be analyzed as a set of developmental stages that may be similar to, but not connected to, other firms' experiences. A common goal in recent analyses of such phenomena is to describe multiple chains and compare them to each other, then classify them or describe their degree of dissimilarity (e.g., see Aisenbrey and Fasang 2010; Piccarreta 2017; Studer et al. 2011). Expanding on the interest of scholars in the relational nature of social phenomena, some social scientists (e.g. Bearman et al. 1999; Bearman and Stovel 2000; Bison 2014; Cornwell 2015) have begun to explore sequences in terms of sets of intersecting events that constitute a larger network of pathways, in which multiple sequence chains are inextricably entwined and not considered as separate. These larger structures of intersecting sequences can be called "sequence-networks." See also Hamberger (2018) in this bundle.

Researchers who study sequence-networks are often interested in characterizing the overall landscape of interconnectedness and intersection—considering these structures as whole standalone entities, opportunity structures (Merton 1996), or social systems (Parsons 1951). Visually speaking, this involves seeing sequences as a vast, integrated web which, in two-dimensional space, resembles a roadmap. The ordered nature of the phenomena that make up these structures naturally invites the use of existing sequence-analytic tools such as discrepancy analysis (Studer

B. Cornwell (✉)
Department of Sociology, Cornell University, Ithaca, NY, USA
e-mail: btc49@cornell.edu

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_7

et al. 2011) and parallel coordinates plot (Bürgin and Ritschard 2014) to help describe them. This has also prompted more focused investigation of the structure of whole sequence-networks, as well as the role that particular elements within those structures play in tying those networks together.

Toward that end, this chapter presents a set of explicitly network-analytic methods for understanding the larger structure that reflects the connections that exist between the individual sequences that one observes within a given setting. This approach expands on other useful sequence-oriented tools. The notion that social life is a set of sequenced phenomena still holds. But the question now shifts to collectivist questions such as "Is the set of action pathways forged by group A more complex than that forged by group B?" I argue that network methods can offer unique insight into these kinds of issues and therefore provide a useful supplement to existing sequence methods. The goal of this chapter, then, is to outline some affinities between network analysis and sequence analysis, and how the former can be used to enhance the latter. To illustrate, this chapter maps out the complex network of sequence pathways by which individuals get through everyday life. To highlight the capacity for network methods to help shed light on the overall structure of these larger maps, it then compares the complex landscape of interconnected pathways that characterize the everyday lives of younger and older adults, respectively.

2 Sequence Pathways in Everyday Life

What does a sequence-network look like? A context that invites this approach is that of individual-level behavior that unfolds on the hourly or daily time scale. How individuals order everyday action has long been an issue in the social sciences (e.g. Bales 1951). Microsequential activity (Gershuny 2000) is composed of series of successive acts (e.g., talking, eating) that form temporally ordered chains. These sequences contain activities that span a specific period of time, usually a 24-h day. Social actions that occur in a given individual's schedule on a given day—e.g., "I took a shower, got dressed, and then went to work"—are temporally connected and therefore provide the basis for larger chains of social action. These are not trivial sequences, as the way people sequence their activities during the course of the day provides a key basis for social organization, and reflects their positions within a larger system of social roles and social norms (e.g. Giddens 1984; Merton 1968; Parsons 1951). Thus, understanding how sets of individuals order their everyday activities—and how those activities are related to each other—is vital to understanding larger social structure.

2.1 Activity Sequences in Networks

Assume that we are interested in the activity sequences reported by g individuals. Assume also that we have activity sequence data from each of these individuals that describe which of the k elements (e.g., number of different types of activities) each individual was doing at each of t time points (e.g., 1,440 min in a given day). Here, the activity data for a given set of individuals are observed in one-minute intervals. An element in a given person's activity sequence may be: "Working between 12:00 p.m. and 12:01 p.m." That particular time-activity will then be followed by another time-activity, such as: "Eating between 12:01 p.m. and 12:02 p.m."

These sequence data can be treated as network data. Every network has two aspects: Nodes, or vertices, and the links, or ties, that associate specific pairs of nodes with each other (Wasserman and Faust 1994). Here, the activities that occur at specific times—in other words, the activity-time elements are treated as the nodes in the network. The "links" between them are temporal connections, which are inferred from their adjacency in the sequence of events that is observed in one or more individuals' activity sequences.

2.2 Organizing the Data as a Sequence-Network

The ordered activity sequence data that are contained in a typical time-diary dataset can be recorded in a rectangular matrix, which we will call $\boldsymbol{A}$. In $\boldsymbol{A}$, the g individuals are arranged down the rows of the matrix and the $k \times t$ activity-time combinations described above are arranged along the matrix's columns. In network terms, this is known as an affiliation or "2-mode" network matrix (Borgatti and Everett 1997; Wasserman and Faust 1994).[1] The left side of Fig. 1 shows how this matrix would be constructed using a simple hypothetical sample of 10 individuals (1 thru 10), who are observed for three time periods (1, 2, and 3) for evidence regarding which of four activities (A, B, C, and D) they were doing at each time point. Each cell contains information about whether the individual in the corresponding row reported doing the specific activity at the specific time that is designated in the corresponding column. For example, a "1" will appear in the upper-left-most cell if individual 1 reported doing activity A in time period 1. That cell would contain a "0" otherwise.

Note that matrix $\boldsymbol{A}$ does not yet show how the time-activity elements are linked to each other in a chain. In network-analytic terms, sequential order can be captured in what is referred to as a "directed" network. One can transform matrix $\boldsymbol{A}$ (which

[1]The network framework is very flexible and can be used in different ways. One might fashion a sequence-network that is composed of multiple sequences from a single individual. For example, individuals' activity sequences from different days might interest, thus providing insight into the emergence of routine in everyday life. Alternatively, one can use networks to show how different activities that occur in potentially non-adjacent time periods link together for a given individual.

Matrix ***A***: Which individuals did which activities at which time points

Subjects	Time-Activities											
	Time											
	1				2				3			
	Activity				Activity				Activity			
	A	B	C	D	A	B	C	D	A	B	C	D
1	1	0	0	0	1	0	0	0	1	0	0	0
2	1	0	0	0	1	0	0	0	0	1	0	0
3	1	0	0	0	0	1	0	0	0	0	0	1
4	0	1	0	0	0	0	1	0	1	0	0	0
5	0	1	0	0	0	0	1	0	0	0	1	0
6	0	0	1	0	0	0	1	0	0	0	1	0
7	0	0	0	1	1	0	0	0	1	0	0	0
8	0	0	0	1	1	0	0	0	0	1	0	0
9	0	0	0	1	0	1	0	0	0	0	0	1
10	0	0	0	1	0	0	0	1	0	0	0	1

Matrix ***B***: Frequencies of transitions between successive time-activities in the activity sequences shown in matrix ***A***

Time-Activities	Time-Activities											
	1A	1B	1C	1D	2A	2B	2C	2D	3A	3B	3C	3D
1A	-	-	-	-	2	1	0	0	-	-	-	-
1B	-	-	-	-	0	0	2	0	-	-	-	-
1C	-	-	-	-	0	0	1	0	-	-	-	-
1D	-	-	-	-	2	1	0	1	-	-	-	-
2A	-	-	-	-	-	-	-	-	2	2	0	0
2B	-	-	-	-	-	-	-	-	0	0	0	2
2C	-	-	-	-	-	-	-	-	1	0	2	0
2D	-	-	-	-	-	-	-	-	0	0	0	1
3A	-	-	-	-	-	-	-	-	-	-	-	-
3B	-	-	-	-	-	-	-	-	-	-	-	-
3C	-	-	-	-	-	-	-	-	-	-	-	-
3D	-	-	-	-	-	-	-	-	-	-	-	-

Fig. 1 Sequence-network matrices ***A*** and ***B***, describing a hypothetical set of ten individuals who engaged in three potential activities (A, B, and C) at four specific time points (1, 2, 3, and 4)

is an undirected affiliation matrix) into a directed matrix that reflects the ordered relationships among the activities. The cells in the resulting matrix reflect how many times a given pair of successive time-activities—for example, 1A and 2B—occur in order across the different sequences. The cells of this new matrix are valued, so that relationships between pairs of successive activities—for example, "Working between 12:00 p.m. and 12:01 p.m." and "Eating between 12:01 p.m. and 12:02 p.m."—are expressed in terms of the number of individuals who reported this transition.

The example from the left side of Fig. 1 is carried over into the right side, which shows the contents of matrix ***B*** for this group. For the purposes of constructing the directed sequence-network, we null all of the blocks along the diagonal (which would represent connections between elements in the same time period, or simultaneous activities), the blocks below the diagonal (which would represent time in reverse), and most of the blocks above the diagonal (which represent links between elements in non-adjacent time periods). This leaves us with the partial, asymmetric matrix that is shown, which contains only the shaded regions. The valid cells in matrix ***B*** tell us in how many individuals' sequences a given activity at a given time is followed by another specific activity at the subsequent time. This is therefore a "valued" network matrix. For example, this matrix would suggest that of these 10 people, two transitioned from doing activity D at time 1 to doing activity A at time 2.

Note that matrix $\boldsymbol{B}$ essentially consists of a set of nested first-order transition matrices, but one could expand to include higher-order transition matrices in the nulled regions above the diagonal. There are also numerous ways one could manipulate the values in matrix $\boldsymbol{B}$, such as by dichotomizing at a given level or by normalizing by row and/or column values. The methods presented here are intentionally left simple and accessible. For an overview of the analysis of temporal networks, see Batagelj et al. (2014).

The following sections describe how the network-analytic framework can be utilized to understand this complex system and what it can reveal about the sequential structure of social phenomena that conventional sequence methods cannot reveal.

3 Analyzing Sequence-Network Structure

As described above, the network framework treats sequences as integrated sets of intersecting pathways. The adoption of this framework does not imply an abandonment of sequence concepts. Rather, it shifts the focus from the individual chains and transitions to the larger structure that results. This system has properties that, once analyzed, reveal additional information that should be of interest to sequence analysts. This section describes some of the properties one might be interested in along these lines (along with some discussion of sequence approaches that share a similar focus), and how to compare those properties across sets of sequences. Note that all of the analyses that were conducted here were executed using Stata 14.2 (StataCorp 2015), but most of these analyses can be conducted using publicly available network software, such as Ucinet (Borgatti et al. 2002) or in R.

3.1 Describing Sequence-Network Structure

I begin by highlighting some affinities that exist between some network concepts and sequence ideas. Table 1 provides a non-exhaustive list of sequential properties (in the middle column) that can be analyzed using an integrated network framework. To give a simple example and to help fix ideas, the concept of *network diameter* reflects the distance between the most distant pair of nodes in the network (that is, how many "steps" they are from each other in terms of the presence of intervening nodes in the observed sequence). In a sequence, this corresponds to the length of the longest observed sequence. Thus, assuming that all of the observed sequences include the same number of time periods, the network diameter is simply the number of time periods being observed.

Network size is the most fundamental network-level property, as it reflects how many nodes are involved. In the sequence context, network size is equivalent to

Table 1 Affinities between key network concepts and sequence concepts, and a comparison of their measurement in the observed sequence-networks of 219 younger and 688 older adults in the ATUS, respectively

Network concept	Corresponding sequence phenomenon	Observed levels in sequence-networks[a]	
		Younger adults	Older adults
Diameter—The largest geodesic distance between any two nodes in the network	The length of the longest observed sequence (number of time periods observed)	1,440	1,440[b]
Size—The total number of nodes in the network	The boundary of the observed element universe (number of unique activity-times observed)	15,489	14,181** (13,442, 15,107)
Density—The proportion of connections that are observed among all pairs of nodes that could be connected[c]	The extent to which alternative possible sequence pathways emerge at a given time (as opposed to a circumstance in which activities at a given time tend to converge on just one activity at the subsequent time)	0.045	.040*** (.037, .042)
Centralization—The extent to which certain ties are stronger than others	The extent to which particular sequence pathways are more common (some transitions between pairs of specific activity-times are experienced by many, others by few)	35.535	40.250*** (38.782, 41.887)
Homophily—The extent to which nodes are connected to similar types of nodes as opposed to different types of nodes	The extent to which particular sequence pathways are composed of like elements (activities tend to remain the same over successive time periods)	0.986	.989*** (.988, .990)

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

[a]Values for younger adults are based on their 219-person network. Values for older adults are derived from the bootstrapping procedure involving 1,000 samples, each with 219 randomly selected older adults. Corresponding 99% confidence intervals for older adults appear in parentheses below the mean.

[b]This value does not differ between groups because the same number of time periods was observed for both groups.

[c]Network density is the total number of different types of transitions that were observed divided by the total number that could have occurred. Between any two successive time points, there are $17 \times 17 = 289$ possible transitions (e.g., eating → working). With 1,440 time points, that means $1{,}439 \times 289 = 415{,}871$ possible transitions per group. This is the denominator for calculating network density.

the number of unique activity-time elements that are observed. This can also be expressed as a proportion of the possible element universe. With respect to activity sequence-networks, a greater number of nodes means that there are more possible pathways through the day, and thus more sequence complexity.

Network density is a property that refers to the extent to which all of the nodes which could be connected within a network are indeed connected. In a sequence context, greater density means that more of the available pathways which could be followed are indeed being followed in reality (to some extent). With regard to activity sequences, greater density means that alternative pathways between time periods are being followed by different people—as opposed to a circumstance in which the sequence are characterized by dominant pathways that many people follow. All else equal, lower levels of density mean that there are numerous pathways that are not being followed.

A related concept is *centralization*. Centrality can refer to the number of connections a given node has or how dominant a given pathway is. Those that are more central tend to be more influential or prominent in the context in question (Wasserman and Faust 1994). For example, a more central node corresponds to an activity that serves as a junction between many other preceding and subsequent activities. This corresponds to what (White 1995) refers to as "publics." Likewise, central ties are those "wide" pathways on which numerous actors are observed at a given time (i.e., common transitions). "Centralization" is the network-level extension of this idea, as it refers to a network in which pathways tend to converge on central events or where certain pathways are more dominant than others. Here, I focus on path centralization, which can be measured simply as the standard deviation around the average width of paths that are observed in the network. (Note that there are numerous other measures of centrality—such as betweenness centrality—that may be employed to understand sequence structure and how particular elements fit into it. But due to space constraints I focus here only on path centralization.)

A final concept is *network homophily*, which refers to the extent to which like elements (e.g., people of the same race) form connections with each other as opposed to with unlike others (McPherson et al. 2001). Sequence-networks are "partitioned" networks, meaning that the nodes have different properties—where temporally adjacent elements may or may not be similar. In this context, homophily can refer to the extent to which sequence elements tend to be followed by like elements in subsequent time period (i.e., a "run" or "spell"). There are numerous ways to measure this, but one straightforward approach is to simply calculate the proportion of times a given sequence element (e.g., sleeping) at one time point is followed by the same element at the next time point, as opposed to transitioning to a different element (e.g., eating).

There are numerous other potentially relevant network concepts one could consider, but measuring the concepts that are described in Table 1 will help to demonstrate how a network approach can help to describe overall sequential structure. To be sure, there are more conventional sequence-analytic approaches one could take to measure related properties of sequence-network behavior. Dissimilarity-based

discrepancy approaches can be used, for example, to assess the diversity of pathways followed by different groups (e.g., see Studer et al. 2011). Likewise, something akin to the homophily measure that was just described can be derived using a measure of the mean of the sequence entropies for subjects belonging to different social groups. Indeed, just about any measure that is derived from matrix $\boldsymbol{B}$ can be derived from a first-order Markov model.[2]

Another benefit of treating sequences as networks is that it facilitates a different type of visualization of sequence pathways. Visualization aids in the comparison of the larger systems of action that characterize different groups. One conventional approach is to present a sequence index plot (e.g., Brzinsky-Fay et al. 2006; Gabadinho et al. 2011) or, more germane to the notion of linking elements across time points, parallel coordinate plots (Bürgin and Ritschard 2014). For recent developments in the visualization of sequences, see e.g. Fasang and Liao (2014) and Gabadinho and Ritschard (2013). Network diagrams that use spring-embedding algorithms (Mrvar and Batagelj 1996, e.g., see) can be useful as well by allowing the positioning of particular activity-times within the plot to vary on the vertical axis, thus highlighting how their relationships to other particular types of elements (e.g., another type of activity) depends on the element position/time. This reflects a core network-analytic argument, which is that the significance and meaning of a given node or element depends on the nodes or elements to which it is connected and its position within the larger structure (Wellman 1983). This will also be illustrated momentarily, in a representation of the daily action sequences of younger and older individuals, respectively.

3.2 Comparing Sequence-Networks

A key concern for many sequence analysts is whether these properties of sequential action differ between groups. For example, do older adults tend to enact more homogenous activity sequences that are characterized by dominant pathways than do younger adults? Conceptually, this involves testing whether older adults' sequence-networks are smaller, evince less variation with respect to tie strength, and have lower density.

But the network framework poses an interesting challenge in this respect, as theoretically a given group (e.g., older adults) is characterized by only one sequence-network. If comparing the structure of sequence pathways for younger individuals to that of older individuals, for example (as will be done later in this chapter), there are only two structures to compare. This is akin to attempting to compare two roadmaps. Prima facie, it seems that this makes any statistically definitive comparison of sequential patterns between the groups difficult. Yet, as will be shown below, with a nonparametric approach, these comparisons are relatively easy to conduct.

[2]I am indebted to an anonymous reviewer for pointing out these very useful affinities between network analysis and sequence analysis concepts and methods.

The first step is to identify some feature of the sequence-networks that one wants to compare. For example, as just discussed, we may be interested in the extent to which the two groups differ with respect to the total number of elements that are involved in their sequence-networks (i.e., sequence-network size).

A challenge emerges when the groups being compared differ in size. This makes the estimates of their key network properties incomparable, as network properties automatically vary by group size. For example, the greater the number of individuals in a group, the larger and less dense the network will be (Borgatti and Everett 1997). Because each group is a different size, any apparent differences in network properties often merely reflects group size differences. Therefore, the procedure will involve comparing the value of a given measure that is calculated for one group to the value of the same measure for the second group that would be expected if that second group were the same size of the first group.

It is relatively easy to use a bootstrap approach to obtain these comparison values (Efron and Tibshirani 1993). Begin by treating the smaller of the two groups being compared as the baseline group. For example, if the first group includes 500 individuals and the second group includes 300, then begin by recording the properties of the network of the second group of 300, and treat these as baseline estimates.

The next task is to obtain comparable estimates for the first group, which contains 500 people. To do this, we draw a random subset of the same size (300 people) from the sample of 500 people in the second group, without replacement. We then calculate the network measure of interest for that random subset, and record it as a first comparable observation. For precision, we do this 1,000 times using different combinations of 300 randomly chosen respondents, thus generating 1,000 comparison estimates. We then average these 1,000 estimates together to assess the "typical" value of the measure in question for 300-person sets from within the larger second group. We use the percentile method (Mooney and Duval 1993) to obtain a 95% confidence interval around this estimate. We can thereby assess whether the observed values of various network measures for the smaller baseline 300-person group fall within the confidence interval that we see for the random subsets that are drawn from the 500-person group. This makes it possible to determine whether there are statistically significant differences in the structure of the sequence-networks of people in the groups being compared.

4 Illustrative Analysis: Activity Sequencing by Age

To demonstrate the potential analytic value of networks for studying sequential social phenomena, a network analysis of the activity sequences that emerge in different age groups is presented below. There are good reasons to analyze how age shapes sequential daily behavior. Most importantly, the notion of "successful aging" is an influential concept in efforts to conceptualize the lives of adults who remain physically, cognitively, and psychologically healthy, as well as physically

active and socially engaged throughout the life course (Rowe and Kahn 1987). A substantial body of research (Adams et al. 2011, e.g.,) shows that older adults who remain physically and socially active lead healthier and happier lives in general.

Attempts to study this aspect of successful aging have typically involved examining age differences in generic estimates or summaries of activity tendencies or levels. Empirical measures capture overall levels of activity—for example, frequencies of social group participation. One problem with such measures is that they lack temporal context. Rates of various types of activity may be higher for members of a given group on the whole, and yet be lower than another group's at certain times of day. The more fundamental problem is that the conventional approach treats activities as independent events, as if they do not occur within a larger chain of activities. As discussed above, activities are elements in a larger chain or path. As numerous scholars have argued, how people experience activity sequences can be just as consequential as their levels of certain activities (Bales 1951; White 1995). Some sequences, for example, include more transitions between activities, which in turn implies greater movement between social roles (Cornwell and Watkins 2015). The higher rates of switching that result require that people move from automatic to deliberative modes of cognition, reduce their dependence on pre-established routines, and heighten their awareness of environmental stimuli and cues (Hitlin and Elder Jr. 2007). The higher rates of cognitive decline among older adults thus provides some motivation for examining the extent to which older adults experience more or less complex activity sequences.

Unfortunately, few studies have analyzed activities in the context of their sequencing—and fewer still that consider this in the context of aging. Research that uses the conventional approach to assessing older adults' activities has generally supported the general assumption that later life is a period of relative simplicity. Do the activity pathways people tend to take through the day in later life indeed wind through a less complex network of action possibilities?

4.1 The Activity Sequence Data

I study this using time diary data that are collected in the annual American Time Use Surveys (ATUS). The ATUS is a nationally representative telephone survey that has been conducted every year since 2003 to assess individuals' work schedules and community involvement. For the sake of illustration, the following analysis relies on data from 2015 only.

To obtain a sample, the ATUS begins by drawing a random sample of households from those leaving the Current Population Survey (CPS) rotation each month. An eligible person from the household (who is at least 15 years old) is randomly selected from the household to be interviewed by phone. In 2015, the response rate was 48.5%, yielding a total of 10,905 respondents (Bureau of Labor Statistics and U.S. Census Bureau 2016).

The main goal of the ATUS is to collect 24-h recall diaries from these respondents, yielding a full account of their activities on the day immediately preceding the interview. ATUS interviewers start by asking respondents about the beginning of the previous day: "So, let's begin. Yesterday [e.g., Thursday], at 4:00 a.m. What were you doing?" They then work forward through the day, collecting information about: (1) what the respondent was doing; (2) the times each activity began and ended; (3) where each activity occurred; and (4) whom the respondent was with. This analysis focuses on the activity data.

In the interview, respondents provide finely grained accounts of how they spent the day in question, in time intervals as small as one minute. All activities except those that occur during periods of paid work are coded using a detailed list of hundreds of possible activities. For the purposes of this analysis, these are collapsed to the ATUS's first-tier coding scheme, which includes 17 possible activities.[3] Because weekday and weekend activities are so different, and because weekdays are bound to provide more structure for most people, this analysis ignores individuals whose diaries were collected on weekends (roughly half).

The goal is to examine any differences in the activity sequences of younger and older adults. This can be done by defining two groups of people who fall into different age ranges and constructing their sequence-networks separately.[4] The two age groups are constructed using twenty-year ranges. For the younger age group, the time diaries of those ATUS respondents who were between 25–44 years of age are included (25 being a typical starting point in studies of time use among working adults.) For the older group, those between 65–84 are included (65 being a typical starting point for studying retirees). It is likely that any differences that are observed between the younger and older age groups will reflect difference in employment status and thus paid work activity. In addition, the data on work activity in the ATUS is fairly poor and non-specific, which will artificially delimit the activity codes for workers. To ensure that this is not the case, this study only includes those who are not in the labor force. Taking all of these restrictions into account, 1,106 respondents remain in scope. Finally, 199 of these individuals are eliminated because they have at least one missing activity code during the 24-h period in question, resulting in a final valid sample of 907 individuals.

The sample includes 219 non-working people between 25–44 years old and 688 non-working people between 65–84 years old. Because the younger age group is smaller (a common sampling problem), it is treated as the baseline group against which the bootstrap samples for the larger, older group are compared.

[3]These activities are: personal care, household (HH), caring for/helping HH members, caring for/helping non-HH members, work/work-related, education, consumer purchases, professional/personal care services, HH services, government services/civic obligations, eating/drinking, socializing/relaxing/leisure, sports/exercise/recreation, religious/spiritual activities, volunteering, telephone calls, and travel.

[4]One can develop more comparable group by implementing a rigorous matching process, whereby members of the two groups are not only different by age, but also similar with respect to other attributes (apart from employment status, which is already controlled here) that could shape activity sequences, such as socioeconomic status and income.

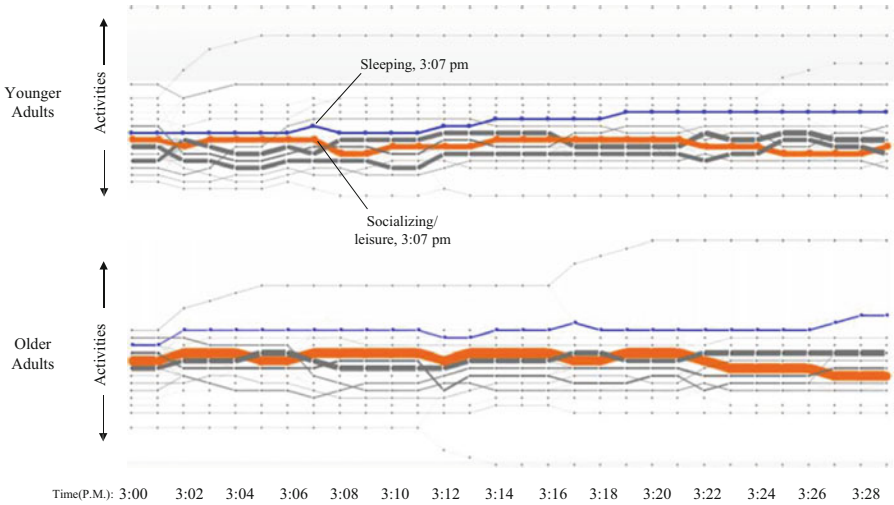

Fig. 2 Sequence-networks representing one half hour in the afternoon among younger and older adults, respectively. Both networks are drawn from equal size groups of 219 individuals. The tiny circles (nodes) represent activities that were reported as occurring during specific one-minute intervals, and the lines represent transitions between those activities from one time point to the next. Blue lines represent pathways characterized by sleep. Orange lines are pathways characterized by socializing/leisure. The thickness of lines is proportionate to the number of respondents who reported those exact activity transitions at that time

4.2 Sequence-Network Analysis Findings

This kind of analysis does not focus on the types of activities people engage in, how that varies over time, or which types of sequences people follow. Rather, it addresses *how sequence* pathways intersect in larger networks. This section presents the unique kinds of findings that such an analysis adds to sequence studies.

A network diagram for the younger group is shown in the top half of Fig. 2. This diagram was created using Pajek (Mrvar and Batagelj 1996), which places points adjacent to each other to the extent that they involve similar actors. For example, when many of the people who are engaged in a given activity at 3:00 p.m. are engaged in another activity at 3:01 p.m., those two activities will appear close together one step to the right. If, on the other hand, very few of the people who are engaged in the 3:00 p.m. activity are engaged in a given activity at 3:01 p.m., those two activities will be separated by substantial space on the vertical axis. This, in turn, means that the diagram adjusts the placement of activities at each time point such that those which fall on pathways that tend to intersect at that time point or adjacent time points are placed more closely together in the diagram.

For illustrative purposes, this diagram shows only a half-hour portion of the 24-h period, but it holds the same structural pattern as other time periods.[5] This network is composed of nodes (small dots)—which represent activities (e.g., leisure) that occurred during specific minute intervals (e.g., 3:00 p.m. to 3:01 p.m.)—and lines, which represent transitions between those activities. Time flows from the left side of the diagram (i.e., the left-most nodes represent activities that took place during the 3:00 p.m. to 3:01 p.m. interval) to the right side of the diagram (3:29 p.m. to 3:30 p.m.). Pathways and activity time-points that involve spells of socializing/leisure (a common activity during this time) are marked in orange, while pathways that involve spells of sleep are marked with blue.

Between the time intervals that are represented here, there are several thick horizontal lines, which represent popular types of activities that many people reported throughout the half hour that is presented here. For example, this figure shows that there were numerous people in both groups who exhibited spells of socializing and leisure during this time (indicated by the thick orange lines in both panels of the diagram). In general, these more common paths contain people who were engaged in one of three common activity types: (1) leisure/socializing/recreation; (2) household activity; and (3) taking care of household members (e.g., children). Thinner strands that occur throughout the period represent people who were doing less common activities or who were switching between these and other types of activities during this time period. For example, spells of sleep were relatively uncommon during this time period.

Note that there is also considerable switching between activity types during this period. An example of this is marked in the younger adults' (top) panel. The fact that the thickest horizontal lines (which represent the most common types of activities) lay relatively close together in the network diagram attests to the fact that there was a lot of switching between these activities throughout the time period. In other words, the network diagraming software modulates the vertical location of the activities vis-à-vis each other such that those activity pathways which are more interconnected via switches appear closer together. This visual connection is one of the advantages that come with using network-diagrammatic algorithms to place particular sequence elements relative to particular sequence positions. A good example is the link between socializing/leisure (orange) and sleeping (blue). As the afternoon wears on, sleeping becomes more decoupled (i.e., it becomes less compatible) with the more common activities of this time period, including socializing/leisure.

[5] A half hour is used as the analytic time frame because it provides enough granularity to allow a close visualization of activity pathways but without the indecipherable jumble of lines and nodes that would appear if a longer period of time were used. This particular time period (midafternoon) is chosen because so few people were asleep.

The measures describing the structure of the younger adults' complete network for the full 24-h period are provided in Table 1. This network is composed of the 15,489 unique activity-times that these respondents reported from among the $1{,}440 \times 17 = 24{,}480$ possibilities (63.3% of all possible combinations) available. There were 18,683 different types of transitions observed between these 15,489 activity-times, out of $1{,}439 \times 17 \times 17 = 415{,}817$ possible types of transitions between all pairs of successive time points. Accordingly, the density of this network is .045, or 4.5% of the possible level of interconnectedness among pairs of elements between successive time points.

The network for younger adults is also highly centralized. The average path width for this age group is 16.9, and the standard deviation around this mean is 35.5, which is the centralization estimate. This means that path widths tended to vary quite a bit from each other, with some paths being very dominant (witness the long tails at the beginning and end of the sequence-network depicted in Fig. 2) and others being singular and poorly populated. It is evident from the diagram that much of the path centralization reflects the predominant norm of sleeping through the late night and early morning hours.

Finally, the homophily estimate is .986. This means that if we observe one of these younger adults at a given time, s/he will be doing the same activity during the next minute 98.6% of the time. In short, their sequence chains contain long spells of the same activity. This high estimate is due largely to the high granularity of the ATUS data, which contains one observation every minute. (Time diary data that come in longer time period increments will automatically have lower homophily estimates.)

A key question is how these aspects of the structure of the older adults' sequence-network compare to these estimates for younger adults. Results of the bootstrapping procedure are presented in the right-most column of Table 1, along with 99% confidence intervals around the mean estimates for older adults. The main findings are that older adults' sequence-networks have an average of 14,181 unique activity-times, which is significantly smaller than the observed size of the younger adults' network ($p < .01$). The older adults' network also tends to be less dense, with an average of 4.0% of all possible activity pathways being realized ($p < .001$).

In addition, centralization remains high in the older adults' sequence-network. But the average level of centralization here (40.250) is significantly greater than we see in the younger adults' network ($p < .001$) indicating that there is less variation around the average path width for this group. Finally, the average homophily estimate for older adults is .989, which is not substantially different but certainly significantly greater ($p < .001$) than the level of homophily in the younger adults' network.

The differences in structure that were just described can also be seen by visually comparing the network diagram for the older sample, which is shown

in the bottom half of Fig. 2.[6] The diagrams reveal the same patterns that are evident in the statistical comparison. The younger adults' network includes more activity times overall and, thus, more possible pathways. The network begins and ends with more activities (15 and 14, respectively) than does the older adults' network, in which only 13 different types of activities are represented at both the beginning and end. Furthermore, there are a greater number of moderately trafficked pathways in the younger adults' networks. In contrast, the older adults' network contains a single dominant pathway (leisure/socializing/recreation) and a secondary moderately sized pathway (household activity). This reflects the greater centralization around dominant pathways within the older sample that is revealed by the bootstrap procedure.[7] In sum, the older adults' network is composed of fewer, wider pathways, which means that they report doing fewer activities and are more likely to follow a few more common activity sequences.

5 Discussion and Conclusion

Increasingly, sequence-oriented scholars are interested in larger structures that consist of multiple intersecting chains of events, or sequence-networks. This chapter has argued that one can gain additional insight into these structures by combining network-analytic with sequence-analytic methods. Scholars have identified several affinities between the network framework and the sequence framework (Bison 2014; Cornwell 2015). This chapter expanded on this work by showing how several key network concepts—including network diameter, size, density, centralization, and homophily—correspond to key sequence concepts—such as sequence length and the extent of variation in the popularity of transitions between different sequence elements. In addition, some network methods, such as visualization of network diagrams, can reveal new things about sequential social phenomena.

This chapter demonstrated the sequence-network approach via an analysis of the structure of the network of activity pathways that are followed by individuals in different age groups. Older adults' pathways are less complex, as they involve fewer

[6]For older adults, a representative network is derived from the reports of 219 in-scope older respondents. The first network that fell within one standard deviation of the means with respect to network size, density, centralization, and homophily was chosen. The observed levels for these measures for this network were 13,929, .039, 40.741, and .989, respectively.

[7]Supplemental analysis (not shown) shows that the structure of this network evolves somewhat over time. At some time points (e.g., nighttime) there is more centralization in both groups. But the degree of centralization tends to be greater among younger adults than it is among older adults in the morning hours (partly because older adults tend to wake up earlier), and this pattern shifts at midday. Thus, a disaggregation of the network estimates by time would reveal significant heterogeneity in levels of observed difference by age.

activities at different times, fewer transitions between different activity-times, are more centralized around fewer and more dominant sequence pathways, and involve less switching between different activities. In light of what we know about the health benefits of complex activity (e.g., Adams et al. 2011), these findings suggest a new area of concern for social gerontology and research on well-being. These findings suggest that older adults inhabit less complex action systems, which may have negative consequences for their social connectedness as well as their cognitive functioning. On the other hand, it is suggestive of greater levels of organization and predictability in everyday life (Zerubavel 1981), which may provide psychological comfort (Giddens 1984) in an otherwise difficult period of the life course. These substantive findings call for further study. But more work is needed to identify which network measures and methods are most useful for examining sequential social action.

I close by reiterating that the network methods that are discussed here are not substitutes for sequence analysis—they are supplements. These methods are a first step in an emerging effort to develop new approaches for understanding how sets of sequences intersect in larger, complex systems. There are several conventional sequence methods already in use which can be used to analyze some of the sequence-related attributes that I have analyzed in this paper. But this paper proposes a new language and analytic framework for the study of sequential phenomena which has as-yet-unknown levels of correspondence with existing sequence methods. I argue that the most insightful analyses of sequential phenomena will come when researchers combine network and sequence methods together.

Acknowledgements Early versions of this chapter were presented at colloquia hosted by the Departments of Sociology at Duke University, Columbia University, and the University of Oxford in 2015 and 2016. I thank the many members of those audiences who commented and asked questions—and especially Peter Bearman, Erin York Cornwell, Jonathan Gershuny, Sang Kyung Lee, James Moody, Gilbert Ritschard, Matthias Studer, Oriel Sullivan, Giacomo Vagni, and four anonymous reviewers—for suggesting changes that improved this chapter.

References

Adams, K. B., Leibbrandt, S., & Moon, H. (2011). A critical review of the literature on social and leisure activity and wellbeing in later life. *Ageing and Society, 31*, 683–712.

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*, 420–462.

Bales, R. F. (1951). *Interaction process analysis: A method for the study of small groups*. Cambridge: Addison-Wesley.

Batagelj, V., Doriean, P., Ferligoj, A., & Kejzar, N. (2014). *Understanding large temporal networks and spatial networks*. Chichester: Wiley.

Bearman, P. S., Faris, R., & Moody, J. (1999). Blocking the future: New solutions for old problems in historical social science. *Social Science History, 23*, 501–33.

Bearman, P. S., & Stovel, K. (2000). Becoming a nazi: A model for narrative networks. *Poetics, 27*, 69–90.

Bison, I. (2014). Sequence as network: An attempt to apply network analysis to sequence analysis. In P. Blanchard, F. Bhülmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 231–248). New York: Springer.

Borgatti, S. P., & Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks, 19*, 243–269.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *UCINET 6 for windows*. Harvard: Analytic Technologies.

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with Stata. *The Stata Journal, 6*(4), 435–460.

Bureau of Labor Statistics and U.S. Census Bureau. (2016). American time use survey user's guide. User's guide, Bureau of Labor Statistics and U.S. Census Bureau, Washington, DC

Bürgin, R., & Ritschard, G. (2014). A decorated parallel coordinate plot for categorical longitudinal data. *The American Statistician, 68*(2), 98–103.

Cornwell, B. (2015). *Social sequence analysis*. New York: Cambridge University Press.

Cornwell, B., & Watkins, K. (2015). Sequence-network analysis: A new framework for studying action in groups. *Advances in Group Processes, 32*, 31–63.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. London: Chapman and Hall.

Fasang, A. E., & Liao, T. F. (2014). Visualizing sequences in the social sciences: Relative frequency sequence plots. *Sociological Methods & Research, 43*(4), 643–676.

Gabadinho, A., & Ritschard, G. (2013). Searching for typical life trajectories applied to childbirth histories. In R. Levy & E. Widmer (Eds.), *Gendered life courses – Between individualization and standardization. A European approach applied to Switzerland* (pp. 287–312). Vienna: LIT-Verlag.

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.

Gershuny, J. (2000). *Changing times: Work and leisure in postidustrial society*. Oxford: Oxford University Press.

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. University of California Press, Berkeley, CA.

Hamberger, K. (2018). Relational sequence networks as a tool for studying gendered mobility patterns. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Hitlin, S., & Elder Jr., G. H. (2007). Time, self, and the curiously abstract concept of agency. *Sociological Theory, 25*, 170–91.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*, 415–444.

Merton, R. K. (1968). *Social theory and social structure*. New York: Free Press.

Merton, R. K. (1996). Opportunity structure. In *On social structure and science* (pp. 153–161). Chicago: University of Chicago Press.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A non-parametric approach to statistical inference*. Newbury Park: Sage.

Mrvar, A., & Batagelj, V. (1996). *Pajek 64-XXL (version5.01). Program for large network analysis*. University of Ljubljana, Ljubljana.

Parsons, T. (1951). *The social system*. New York: Free Press.

Piccarreta, R. (2017). Joint sequence analysis: Association and clustering. *Sciological Methods & Research, 46*, 252–287.

Rowe, J. W., & Kahn, R. L. (1987). Human aging: Usual and successful. *Science, 237*(4811), 143–149.

StataCorp. (2015). *Stata statistical software: Release 14*. College Station: StataCorp LP.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research, 40*, 471–510.

Wasserman, S., & Faust, K. (1994). *Social network analysis*. New York: Cambridge.
Wellman, B. (1983). Network analysis: Some basic principles. *Sociological Theory, 1*, 155–200.
White, H. C. (1995). Network switchings and Bayesian forks. Reconstructing the social and behavioral sciences. *Social Research, 62*, 1035–1063.
Zerubavel, E. (1981). *Hidden rhythms: Schedules and calendars in social life*. Chicago: University of Chicago Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Relational Sequence Networks as a Tool for Studying Gendered Mobility Patterns

Klaus Hamberger

1 Introduction

In recent years social network analysis and sequence analysis have become increasingly attentive to one another. This mutual interest stems from a profound affinity in their objects of study. In fact, sequences and networks are not so much separate domains as complementary dimensions of the same social reality—they belong together as time and space (see Abbott 1997). On the one hand, sequences are not just isolated linear series of events. A given event may participate simultaneously in several sequences, and a given type of event may occur several times in the same sequence. Put in context, sequences turn out to be mutually intersecting and potentially nonlinear walks in a comprehensive network of events or event types. On the other hand, social networks are more than static compilations of relations between individuals. Social relations are essentially processual, being created, transformed and dissolved in time, and each of these transformations constitutes an event in one or more sequences. Considered in their dynamic aspect, social networks turn out to be the cumulative results of a series of interconnected sequences. Sequence analysis thus logically leads to network analysis and vice versa.

One reason for the long-standing separation of the two approaches resides in the fact that they do not involve the same type of elements. The networks constructed from sequences are networks of events linked by ties of causation or succession. These event networks have been introduced in the study of narrative (Bearman et al. 1999; Bearman and Stovel 2000), in life course studies (Bison 2014; Fitzhugh et al. 2015) and in time use studies (Cornwell and Watkins 2015; Cornwell 2015, 2018). By contrast, the historical core domain of social network analysis has

K. Hamberger (✉)
Laboratoire d'Anthropologie Sociale, Ecole de Hautes Etudes en Sciences Sociales, Paris, France
e-mail: klaus.hamberger@ehess.fr

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_8

been constituted by networks of (individual or collective) agents linked by ties of friendship, kinship, exchange, and so on. While sequence analysis has been recently applied to the study of the dynamics of agent networks (Stark and Vedres 2006, 2012), the sequences thus constructed consist of network positions (or positional states) rather than of events.

Clearly, the two approaches just described are complementary. Sequences consist of events that are linked to each other in the experience or narration of an agent, and social networks consist of agents who are connected to each other via their participation in a common event. Event networks and agent networks thus can be derived from one and the same bimodal network of agents and events. An integrated approach should analyze both kinds of networks simultaneously and study their mutual interaction, as each event transforms the agents' social environment, which in turn shapes the path of future events.

A natural candidate for exploring the potential of this integrated method is the domain of migration and mobility. On the one hand, mobility events are among the classical subjects of life course studies and sequence analysis. On the other hand, social network approaches, including longitudinal perspectives, are increasingly used in migration studies (Lubbers et al. 2010). However, these approaches generally deal with the relevance of migrants' social networks to social integration or transnational support, rather than with the migration process itself. Studies of the ways in which social networks influence migration processes (Liu 2013) remain rare, and rarer still are studies that examine how social networks and mobility interact (Wissink and Mazzucato 2018).

This paper adopts an integrated method which uses social network analysis both to define the alphabet of the sequences (as a set of types of relations) and to represent the sequences themselves (as oriented networks of relational types).

We will apply this method to examine the impact of gender on the patterns of local and regional mobility in South-east Togo. In this region of West Africa, both men and women are highly mobile, though for different reasons. The mobility of Togolese women is closely connected to their traditional role in trade (Cordonnier 1987) and the long-standing practice of migration of girls and young women to the capital Lomé as foster children and/or as domestic workers (Pilon and Ségniagbéto 2014). This pattern of non-matrimonial female mobility is complemented and reinforced by a virilocal post-marital residence pattern combined with traditionally high divorce and remarriage rates (Locoh and Thiriat 1995). Male mobility, on the other hand, has historically taken the form of circular labor migration, be it as farmhands, tenant farmers or lumberjacks to the plantations and forests of Western Togo, Ghana or Ivory Coast (in the case of the older generation), or more recently as unskilled workers to urban centres on the west African coast between Abidjan and Lagos. Even when rural emigration becomes long-term, the paternal home in the village frequently remains a refuge in case of unemployment, illness, and for retirement in old age, a pattern that is widespread in colonial and postcolonial West Africa (see e.g. Cordell et al. 1996).

The network approach proposed in this paper is intended to improve understanding of the logic underlying these broad tendencies by examining the micro-morphology of male and female migration trajectories and relational environments.

2 Method

2.1 Basic Concepts

Every event sequence can be represented as a path in a network, where nodes represent events and arcs represent immediate temporal succession. In Fig. 1 (left), for example, two simultaneous sequences (blue and red), are represented as paths through five events of four different types (A-D), two of which (B2 and B5) are common to both. By gathering similar events into larger classes, any such path will induce a sequence network, the nodes of which are not single events but event classes. Sequence networks of this sort are a variant of the networks used in life course analysis, such as life history graphs (Butts and Pixley 2004) or state transition graphs (Bison 2014). Since events of a given class can occur more than once in these networks, sequences now will no longer be linear paths but potentially nonlinear walks. Thus, in Fig. 1 (right), the networks corresponding to the sequences ABCBB (blue) and CBDAB (red) contain loops (BB), mutual dyads (BCB) and circuits (BDAB).

Events can be classified according to many different criteria, each giving rise to another kind of sequence network. For example, a classification by place of destination will result in the geographical network of migration routes. Here we are interested in people's movement in social space, so we shall classify events according to the relations that link the individuals involved.

Agents can participate in events in a variety of roles (such as "migrant", "host", and so on), and may be related to each other by relations that are polyadic ("A is child of B with C") and composite ("father's sister", "mother's employer" etc.). Complicated as they may appear, networks of this sort are not unfamiliar to

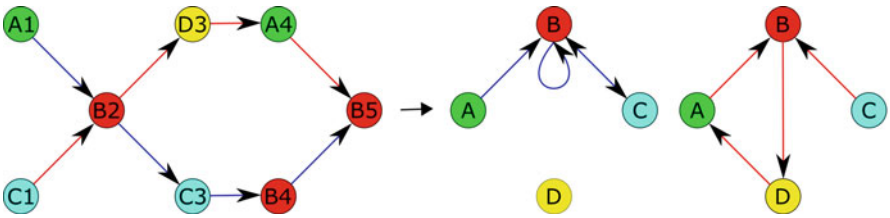

Fig. 1 Event networks (left) and sequence networks constructed from them (right). Nodes represent events (left) or event types (right), arcs represent immediate succession. Sequences are represented as (linear) paths in networks of events and as (nonlinear) walks in networks of event classes (sequence networks)

social network analysis: kinship networks are commonly conceptualized as bimodal networks where "individuals" are linked to "families" by a variety of roles such as "father", "mother" and "child" (Batagelj and Mrvar 2008; Hamberger et al. 2011, 2014), and kinship relations comprise composite chains of parent-child or marriage ties. Now what we call "families" in kinship network theory are not so much sociological units but series of life course events (marriages and births) related to the same couple. We thus can conceive of "family events" (to which individuals are linked as "parent" or "child") in the same way as of "mobility events" (to which they are linked as "host" or "migrant") and vice versa. This similarity between kinship and mobility networks is more than formal. Marriage and divorce usually involve a change of residence for one of the partners, and every birth is in a sense a mobility event creating the starting point of an itinerary. Kinship and migration networks can accordingly be modeled as parts of an integrated bimodal network (see Fig. 2), where individuals and (mobility or family) events are represented as two different sets of nodes, and where the various roles that individuals play in these events are represented by different sets of edges.

The unimodal projections of this integrated bimodal network yield the one-mode networks usually studied in life course and kinship studies: networks of events linked to each other by the involved individuals, and networks of individuals linked to each other by shared events. For a given individual (ego), we can in this manner derive both the event sequence that constitutes his or her itinerary and the personal network that results from this itinerary.

The anchorage of both personal networks and event sequences in a single, multiplex bimodal network makes it possible to classify individuals and events by means of the relational circuits that emerge as mobility links combine with chains of other links (kinship, friendship, employment, and so). For example, individuals C and H in Fig. 2 are linked to each other as migrant and host (via mobility event 2), but also as brother's wife and husband's sibling (via birth events 3a and 3b and marriage event 2), thus forming a relational circuit (marked blue in Fig. 2). Thus, we can first classify the individuals of ego's personal network by the relational chains that link them to ego (for example, host H as "brother-in-law"), and then classify the events of ego's itinerary by the relational classes that characterize individuals in certain roles (for example, mobility event 2 as "move to brother-in-law"). This relational classification of events in turn allows us to merge similar nodes of ego's itinerary, thus transforming it into a sequence network. Since the alphabet of sequence networks of this kind consists in classes of relations rather than in exogenous attributes (such as the motive for mobility events or the social class of the host), we distinguish them as relational sequence networks.

The specificity of relational sequence networks resides in using network techniques both for analyzing and for constructing sequences. They can be applied to any kind of sequences whose elements are pertinently characterized in terms of relations. In addition to migration studies, they may, for example, prove useful in the sociology of interactions, the morphology of folktales or the anthropology of ritual. The fact that we shall use them to study the impact of what may at first sight appear as an attribute of the individual—gender—does not contradict this

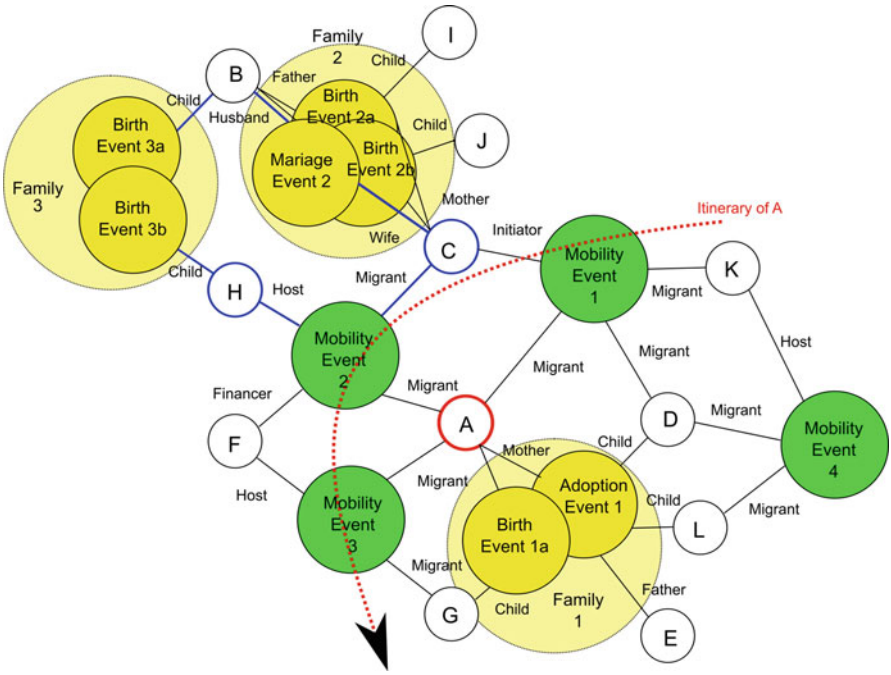

Fig. 2 An integrated kinship-mobility network. White nodes represent individuals; green and yellow relational nodes represent mobility and family events, respectively (family event nodes can be merged to form conventional family nodes, light-yellow). Edges linking individual to relational nodes correspond to the individual's roles in the event. Itineraries are represented as (dotted red) paths through successive mobility events

thoroughly relational orientation. Actually, the results of our analysis corroborate the now widely accepted approach that views gender itself in relational terms.

2.2 Data

The study is based on data collected in a field survey conducted in rural South-east Togo between 2010 and 2015.[1] The survey was centered on Afagnan-Gbléta, a large village whose core area has a population of around 4800 inhabitants (census

[1]Most of the interviews were conducted by the author in collaboration with Toussaint Yakobi and Komi Malou Kakanou during four stays (2009-10, 2010, 2012, and 2014-15). Forty-two interviews during autumn 2010 have been conducted by Ibitola Tchitou (URD, Lomé University) and the same collaborators.

conducted by author in 2014).[2] It is located in the maritime region of Togo, a very densely populated area (while it covers 10% of the national territory, 42% of the Togolese population live there). The region falls within the zone of West Africa with the highest intra-regional migration. Afagnan-Gbléta is situated close to the border of Benin, and the capital Lomé, at 81 km distance, can be reached within two hours by "bush taxi". The village is located close to a major West African migration route connecting the cities of Abidjan, Accra, Lomé, Cotonou, and Lagos. Short geographical distances, affordable transport, easy border crossings in the ECOWAS region and extended social and kinship networks all facilitate high transnational mobility.

The sample has been established in a two-step procedure. The first 60 interviews were conducted with persons drawn at random from the adult (>15 years) residents of Afagnan-Gbléta (based on the village census effectuated in 2004/5) who were still alive and resident in Togo at the beginning of the study in 2010.[3] In the second step, we conducted interviews with all persons mentioned by the initial interviewees as having played a role in their migration biographies or as having been hosted by them, as far as these persons were alive, reachable within the perimeter of South-east Togo, and willing to talk with us. Of the 876 living persons (from a total of 1111 persons named), we were able to interview 328 persons on their own and on their non-adult children's trajectories, which resulted in 449 additional biographies. The total snowball sample thus contains 509 migration biographies (260 male and 249 female).

Migration biographies were collected through retrospective semi-directive interviews. We asked interviewees to recall all migratory events leading to a change of residence for more than three months[4] from birth to the present (time of the interview, last update of information in 2014/15), and to describe the context, motive and course of each migratory event. In particular, we collected information (name, relation and contact) on the persons who received ego (hosts), accompanied ego (co-migrants), initiated ego's displacement (initiators) and financed ego's journey (financers). All other persons mentioned in the context of the migration event were equally noted (others). This set of names was complemented by a list of the interviewee's parents, spouses and children. Finally, we asked each interviewee to name all those persons for whom he or she had acted as a host.

This dataset is completed by a genealogical and residential dataset (about 50,000 individuals including the deceased) collected during three censuses of the village of Afagnan-Gbléta (2004/5, 2009/10 and 2014/15) as part of ongoing ethnographic

[2]According to 2010 official census data, Afagnan-Gbléta counts 12,411 inhabitants (5916 males and 6495 females) when all its hamlets are included.

[3]Information on present life status and residence were obtained after the draw, so that there were 76 draws in all (35 women/41 men) because 6 (3/3) had died, 5 (1/4) had moved abroad, one woman refused the interview, and 4 young men appeared in the sample due to errors concerning their age.

[4]This restriction was meant to exclude short-term stays for funerals, visits or holidays. In fact, the problem was rather that even year-long stays, for example at the natal home after childbirth, were often considered by interviewees as temporary absences not to be mentioned, and discovered only afterwards when dealing with the children's itineraries.

fieldwork (see Hamberger 2011). Most of the remote kinship relations (such as "father's mother's brother's daughter") and mediated non-kinship relations (such as "friend's employer") were thus not (only) directly reported by interviewees but computed from data stemming from numerous different oral sources. The anonymized dataset is available on the Kinsources repository (permanent link: https://www.kinsources.net/kidarep/dataset-275-watchi-2017.xhtml).

The total snowball sample has only an auxiliary value for the analysis. Besides the bias in favor of local and circular migration trajectories owing to the geographical limits of the survey, the snowball technique introduces a systematic bias in favor of connected and mobile persons, who have a higher chance of playing a role in the itineraries of the first interviewees. All comparative data concerning the overall sample thus are likely to overestimate mobility and network cohesion. Moreover, not being composed of independent observations, the snowball sample cannot be analyzed with standard statistical techniques. Most of our analyses are therefore restricted to the random sample of 60 itineraries. The remaining 449 itineraries serve essentially to construct the ego-centered social spaces and sequence networks of these 60 cases. For the purpose of illustration, four of these cases will be considered in further detail below.

2.3 Software Tools

All analyses used in this paper have been made using the open source software Puck version 2.3.50 (Hamberger et al. 2014), which can be downloaded at http://www.kintip.net (source code at http://sourceforge.net/projects/tip-puck/). Initially developed for the study of kinship networks, Puck contains, from its 2.2 version onwards, a package (org.tip.puck.sequences) for the study of longitudinal data. The networks of Figs. 3, 6, 7, and 8 have been visualized with the software Pajek (http://mrvar.fdv.uni-lj.si/pajek/) from files produced by Puck. The detailed procedures for reproducing the analyses of the paper are described in the document accompanying the Watchi dataset on the Kinsources repository (see above).

3 Results

3.1 Personal Networks

In order to introduce the analyses presented in the remainder of this paper, let us begin by considering one empirical case example. We have chosen the itinerary of Betty,[5] a middle-aged female domestic worker with transnational experience (Benin

[5]First names of the four example individuals have been changed.

Table 1 Betty's (10683) itinerary as sequence of successive mobility events. Relational roles of hosts and co-migrants are indicated in brackets. Village and quarter names have been replaced by numbers

	Year	Age	Start and end places	Hosts	Co-migrants
0	1961	0	> V1	10690 [MOTHER]	
1	1973	12	V1 > Lomé-1	41315 [RELATIVE]	
2	1985	24	Lomé-1 > Lomé-1	41319 [EMPLOYER] 41320 [EMPLOYER]	
3	1986	25	Lomé-1 > V2	10687 [FATHER] 10690 [MOTHER]	
4	1987	26	V2 > Lomé-2	41318 [SIBLING]	
5	1988	27	Lomé-2 > Lomé-3	10682 [SPOUSE]	
6	1990	29	Lomé-3 > V3	10678 [AFFINE]	
7	1991	30	V3 > Lomé-3	10682 [SPOUSE]	10685 [CHILD]
8	1992	31	Lomé-3 > Cotonou	41321 [SPOUSES_EMPLOYER]	10682 [SPOUSE] 10684 [CHILD] 10685 [CHILD]
9	2000	39	Cotonou > V3	10682 [SPOUSE]	10684 [CHILD] 10685 [CHILD] 10686 [CHILD] 10695 [RELATIVE_UTERINE]
10	2006	45	V3 > Lomé-1	39670 [AFFINE, MASTER]	
11	2007	46	Lomé-1 > Lomé-4	41322 [AFFINE, EMPLOYER]	
12	2007	46	Lomé-4 > Dakar	41323 [EMPLOYER]	
13	2008	47	Dakar > Lomé-1	39670 [AFFINE, MASTER]	
14	2008	47	Lomé-1 > Lomé-5	41324 [EMPLOYER]	
15	2009	48	Lomé-5 > Lomé-6	41325 [EMPLOYER]	
16	2010	49	Lomé-6 > Lomé-6	56354 [LANDLORD]	10684 [CHILD]
17	2013	52	Lomé-6 > Lomé-7	56751 [LANDLORD]	
18	2013	52	Lomé-7 > Mali	56750 [LANDLORD]	

and Senegal, most recently Mali). Table 1 represents the sequence of successive mobility events, each of which links her to one or more hosts,[6] some of them also to co-migrants. Their respective identity numbers are listed in the last two columns of the table.

In order to study the social networks shaping (and shaped by) individual trajectories, we first have to look at the way in which the relations created by a mobility event (linking a migrant to his or her hosts, co-migrants, and so on)

[6]While birth mothers constitute by convention the starting points of all itineraries (and sequence networks), they have not been explicitly coded as hosts at birth so as to avoid trivial results of relational censuses and personal network analyses.

coincide with other direct or indirect relations between these individuals.[7] We have considered seven main types of relations: (1) kinship ties up to the fourth canonic degree for consanguines (ex. third cousins) and up to the first degree for affines (ex. fathers-in-law); (2) non-genealogical kinship (including both kinship ties for which the genealogical chain is unknown and those for which probably no such chain exists, such as the "father's child" relationship established by reference to a common locality of origin); (3) friendship (including relations between colleagues, fellow workers and classmates); (4) apprenticeship; (5) initiation (which includes a more or less prolonged stay at a vodu priest's home); (6) employment; and (7) landlord-tenant relationship. Since the latter generally implies the absence of a preexisting social relation, hosts are classified as "unrelated" only when they have accommodated ego without payment.

The relations linking Betty to her hosts and co-migrants are given in brackets in the last two columns of Table 1. As we can see, Betty has been received by her parents (1 event, without counting the birth event), her brother (1), her husband (3), a relative (1), three affines (4), one of whom is also an employer (1) and another a (future) master (2), various unrelated employers (4), her husband's employer (1), and various landlords who rented her a room (3). She has mainly travelled alone, and has only occasionally been accompanied by one or more children (3), a uterine niece (1), and her husband (1).

Table 2 shows the frequencies of these various kinds of host and co-migration relations for all itineraries of the total snowball sample (509 cases). The numbers indicate the number of itineraries in which the migrant-host or the co-migrant relation coincides at least once with a social relation of the respective type.

A look at the types of hosts and co-migrants occurring in male and female itineraries instantly reveals some marked differences (but also some important similarities). The virilocal orientation of residence in Togolese society is clearly evident: almost all adult women have moved to their husbands or affines on at least one occasion while the reverse trajectory concerns only a small minority of men.[8]

A more remarkable feature is the marked tendency to return to the parental home. This circular pattern holds not only for men who hold land rights in their natal family, but also for women, who systematically return to it before (re)marrying.

For both men and women, relatives outside the immediate family circle are at least as important as parents in the role of hosts—with a slight preference of male and female for agnatic and uterine relatives, respectively.

[7]Though these relations generally precede the mobility event, there are some cases of relations that have been created after the move (e.g. by marrying a host's relative, or by starting an apprenticeship with the host as master).

[8]The definition of 'host' is here bound to land and residence rights rather than to residential precedence. Thus, a wife returning to her husband has been classed as returning to her marital home, while a man returning to his wife has been categorized as returning to his own (or parental, or rented) home. This is a reasonable assumption to make in the rural context but may introduce unwarranted male bias into data collected in an urban setting.

Table 2 Types of relations to hosts and co-migrants occurring in male and female itineraries. Figures indicate the number of (male, female, all) itineraries where a host or co-migrant in the corresponding relation to ego appears in at least one event. Percentages refer to the total number of itineraries concerned (260 male, 249 female, 509 all)

	Migrant-host relations						Co-migrant relations					
Relation type	M	%	F	%	All	%	M	%	F	%	All	%
Own property	52	*20.0*	8	*3.2*	60	*11.8*						
Parents/Parental home	167	*64.2*	149	*59.8*	316	*62.1*	88	*33.8*	79	*31.7*	167	*32.8*
Spouse/Marital home	2	*0.8*	162	*65.1*	164	*32.2*	90	*34.6*	77	*30.9*	167	*32.8*
Sibling	67	*25.8*	56	*22.5*	123	*24.2*	117	*45.0*	98	*39.4*	215	*42.2*
Child	2	*0.8*	14	*5.6*	16	*3.1*	72	*27.7*	116	*46.6*	188	*36.9*
Relative	162	*62.3*	161	*64.7*	323	*63.5*	77	*29.6*	71	*28.5*	148	*29.1*
Agnatic relative	68	*26.2*	54	*21.7*	122	*24.0*	34	*13.1*	27	*10.8*	61	*12.0*
Uterine relative	64	*24.6*	80	*32.1*	144	*28.3*	24	*9.2*	27	*10.8*	51	*10.0*
Affine	37	*14.2*	79	*31.7*	116	*22.8*	44	*16.9*	48	*19.3*	92	*18.1*
Landlord	116	*44.6*	52	*20.9*	168	*33.0*						
Friend	36	*13.8*	12	*4.8*	48	*9.4*	40	*15.4*	8	*3.2*	48	*9.4*
Employer	39	*15.0*	29	*11.6*	68	*13.4*	1	*0.4*	0	*0.0*	1	*0.2*
Master	51	*19.6*	9	*3.6*	60	*11.8*	7	*2.7*	0	*0.0*	7	*1.4*
Vodu priest	5	*1.9*	24	*9.6*	29	*5.7*						
Public (Hotel, Market)	19	*7.3*	5	*2.0*	24	*4.7*						
State or NGO	14	*5.4*	5	*2.0*	19	*3.7*						
Unrelated	16	*6.2*	1	*0.4*	17	*3.3*						
Unknown	10	*3.8*	12	*4.8*	22	*4.3*	9	*3.5*	2	*0.8*	11	*2.2*

Residence with a master during apprenticeship is more relevant for boys than for girls, who are more frequently employed as market helpers or domestic workers.[9] By contrast, it is mainly girls who as initiates spent a (today much reduced) period of residence with a vodu priest. The male predominance in monetary relations to landlords may partly reflect a tendency among interviewees to represent the husband as the contracting party even if both spouses contribute to the rent (we did not verify who pays). By contrast, stays in state- or NGO-owned facilities (caserns, schools, hospitals) or in hotels are largely restricted to the educated male elite.

While the gender inflection of host-migrant relations concerns mainly the marital tie, its impact on co-migration becomes manifest in the parental tie: children move with mothers rather than with fathers. By contrast, travelling with non-kin (friends and fellow workers) is more frequent for men than for women. Predictably (since our data include infant migration), siblings are among the most frequent co-migrants.

After this first inspection of the positions that make up the social space in which a person navigates, let us now take a closer look at the structure of this space. We

[9]In cases where masters and employers were relatives, the work relationship has been given precedence over kinship when constructing sequence networks.

shall confine our analysis to the personal networks of the 60 initial interviewees. Clearly, our picture of these networks is incomplete since neither the dead nor most of those living abroad could be interviewed about their own relations, forcing us to rely on ego's indications for them. This data collection bias reinforces ego's memory bias, so that the resulting networks are in a sense doubly "ego-centered", being most dense in ego's actual spatial and temporal neighborhood.

The nodes of the networks are made up of all persons who played a role as host, co-migrant, initiator or financer in ego's itinerary, but not those persons for whom ego played only the role of a host. Also, the network does not contain those named as "others", whose role consists in providing the general context. Composite kinship links mediated by persons who are themselves in the network have not been represented by direct arcs so as to preserve their mediated character.[10]

Figure 3 shows four example networks, where relation types are indicated by different colors. Besides Betty's personal network (upper left), we have selected the networks of Omar, a male tailor circulating between Afagnan-Gbléta and the Benin-Nigerian frontier; Sosime, an elderly female trader who has returned to Afagnan-Gbléta after an itinerary including transnational commuting between Lomé and Lagos; and Aloesso, a former goldsmith who spent his youth living in his elder brother's household at Lomé.

We are interested in the importance of different relations for the network structure, be it as intermediaries for indirect relations or as nuclei of locally cohesive clusters. The first aspect can be measured by the betweenness centrality[11] of the individuals so related to ego; the second by the fact that these individuals play a locally central role (measured by node degree)[12] in one of the connected components into which the network decomposes if ego is removed. In Fig. 3, the individuals (other than ego) with the highest betweenness centrality have been marked by bold borders, and connected components (after removal of ego) have been encircled. As can be seen, Betty's network can be decomposed into 10 components, centered respectively around her parents (1), her husband (1), an employer (3), a relative (2) and a landlord (3), where all but one of the three last mentioned categories constitute single-element components. The largest component (9 nodes) is the marital component centered around the husband, who is also the individual with the highest betweenness centrality (14%) after ego (85%).

Table 3 indicates the number of itineraries in which each of the considered relations characterizes the most central alter in the whole network or in one of its components (in addition, we have indicated the mean size and the maximal number of the respective components centered on this relation).

[10]For example, if ego has migrated with his or her mother and maternal grandmother, the network will contain the grandmother but no direct link between her and ego.

[11]The betweenness centrality of a node (Freeman 1977) is measured as the normalized number of shortest paths running through it.

[12]The (in-/out-) degree of a node is defined as the number of nodes to which it is linked (by incoming/outgoing arcs).

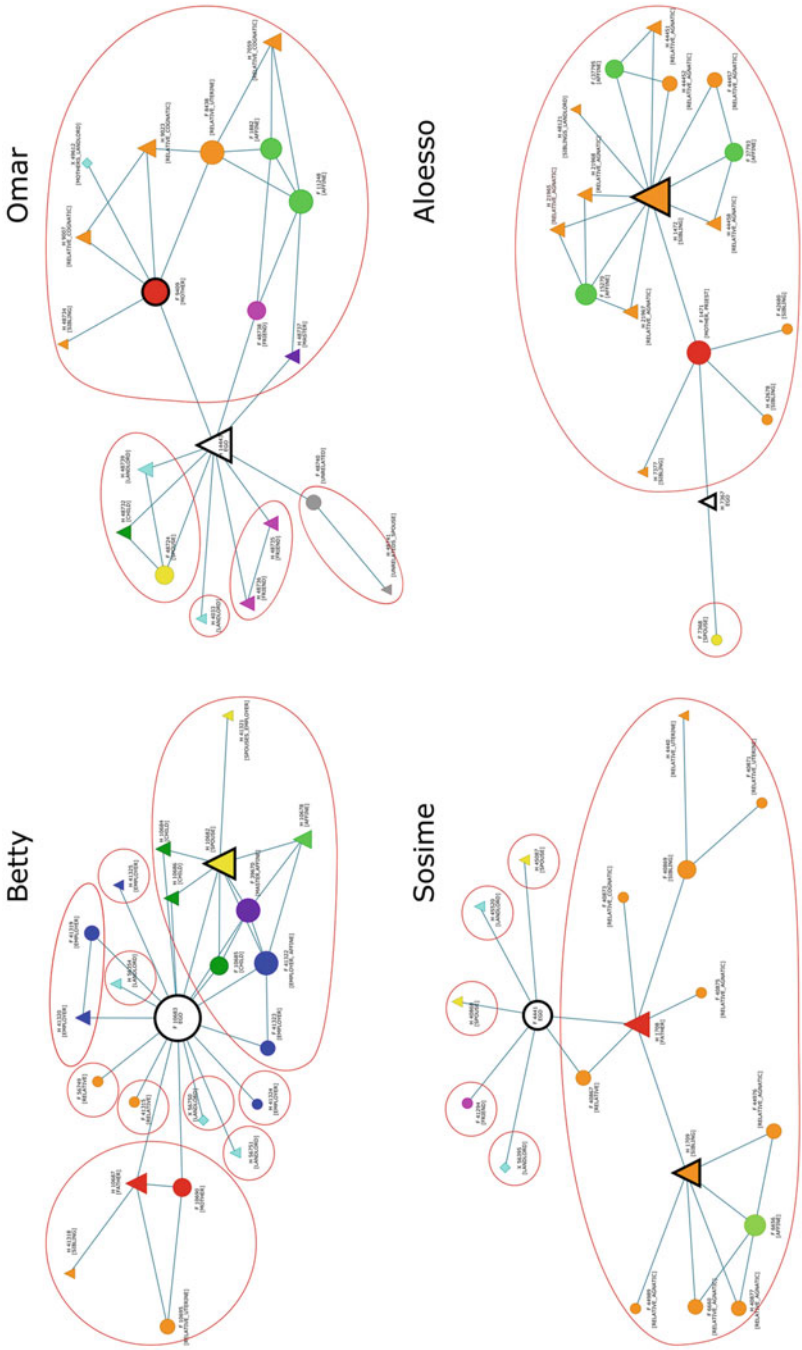

Fig. 3 Four case examples of social networks resulting from migration biographies. Triangles and circles correspond to male and female individuals, respectively. Ego (white) and central alters (colored) are highlighted by bold borders, components are encircled, node size corresponds to degree. The color code is given in Fig. 4

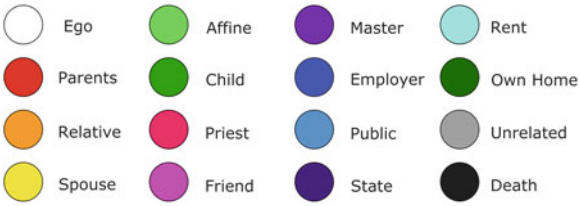

Fig. 4 Color code for Figs. 3, 6, 7, and 8

Table 3 Globally and locally central relations. Figures indicate the (absolute and percentage) numbers of personal networks where the individual(s) most central in the network or in one of its components have the corresponding relation with ego. The table also indicates the mean size and the maximal number of such components; if there is more than one, mean component size refers to the largest. Percentages refer to the total number of itineraries concerned (30 male, 30 female, 60 in all)

	Maximal (non-ego) centrality				Component nucleus							
	Male		Female		Male				Female			
	#	%	#	%	#	%	Mean size	Max nr	#	%	Mean size	Max nr
Father	15	*50.0*	10	*33.3*	10	*33.3*	5.1	1	12	*40.0*	5.5	1
Mother	11	*36.7*	4	*13.3*	7	*23.3*	6.6	1	2	*6.7*	4.0	1
Fa+Mo					5	*16.7*	4.8	1	3	*10.0*	5.0	1
Spouse	3	*10.0*	9	*30.0*	15	*50.0*	4.1	1	23	*76.7*	3.7	3
Sibling	2	*6.7*	2	*6.7*								
Friend	2	*6.7*			9	*30.0*	1.1	7	1	*3.3*	1.0	1
Relative	3	*10.0*	3	*10.0*	9	*30.0*	3.1	2	4	*13.3*	4.3	2
Affine			1	*3.3*					2	*6.7*	8.0	
Landlord	1	*3.3*			14	*46.7*	1.0	9	6	*20.0*	1.0	4
Master	1	*3.3*			5	*16.7*	1.2	2				
Vodu priest			1	*3.3*					2	*6.7*	1.0	1
Employer	2	*6.7*			7	*23.3*	1.6	3	4	*13.3*	1.0	3

We now see that the various relations that we previously examined in isolation operate in quite different ways in organizing ego's social space. On the one hand, there are relations that function as connectors: they give rise to other relations which they mediate, and thus become centers and nuclei of large connected components—this is the case of kinship relations, in particular the parental ties that mediate all others. On the other hand, there are relations that operate as divisors: they neither integrate themselves into an existing social space nor lead to other relations, but tend to form numerous small components. These "atomizing" relations are usually mediated by money, linking ego to landlords, masters or employers. However, some relations of this kind may mediate relations between fellow apprentices or working colleagues (counted as "friends" in our classification).

The extremes of these two types of relations—one centralizing a singular large component, the other leading to multiple marginal isolates—constitute a polarity which to a certain extent characterizes the male mobility pattern. On the one hand,

the kinship core, while important for both genders, is more fundamental for male networks: 50% of male networks, as opposed to 33% of female networks, are centered around the father (who forms the local center of a component in 50% of both male and female networks). On the other hand, the atomized components of rent and wage relations occur both more frequently and in greater number in male than in female networks. The typical male social space is centered around a single large kinship base from which repeated expeditions are undertaken to a variety of rapidly changing social and spatial destinations. The rural-urban circular migration also largely follows this radial pattern, and even those who are in stable employment (teachers, policemen, etc.) are mostly implicated in continuous displacements.

Not all relations, however, fit into this polarity. On the one hand, there are relations that are both unique and isolated, such as the relation of initiate to vodu priest. On the other hand, there is the marital relation, which is both relatively central and gives rise to cohesive components structured around nuclear families, but which, in this region of West Africa, is typically multiple. While men can have several wives at the same time, women generally have several husbands at different times in the course of their lives. Contrary to men, for women each new marriage involves a displacement, and mediates a series of new relations that play a role for their future migrations. Each new husband thus becomes the nucleus of a (frequently large) component of the wife's mobility space.

The gender-specific occurrence of different relations noted previously thus corresponds to different relational logics in the structuring of social space, which is evident in the contrastive overall morphologies of male and female personal networks. On average, female networks are more cohesive than male networks: they contain half as many isolates (1.3 vs 3), their average number of components is lower (2.8 vs 4.1), and their concentration—measured as the Herfindahl index of component shares—is higher (0.58 vs 0.52). However, this concentration does not imply the presence of one dominant giant component, but of two or several components of roughly equal importance. As a corollary, ego's own centrality is generally lower for female than for male egos (0.48 vs 0.62). The more or less (de-)centered position of ego can be expressed more precisely in considering the normalized difference between the (betweenness) centralities of ego and the most central alter. This index of ego's differential centrality ranges from −100% for maximal marginality to +100% for maximal centrality, where 0% indicates a central position shared by ego and one or more central alters. Differential centrality is closely correlated to the concentration of ego's personal network (see Fig. 5). The more disintegrated the network (after removal of ego), the higher ego's centrality (with the migrant teacher as an extreme example), and the more integrated the network, the higher ego's marginality (with the foster child at the other extreme).

While Fig. 5 shows the moderate average tendency of female networks towards higher concentration and lower ego centrality, it also shows a great diversity. To explore further this diversity, let us consider our four case examples, each of which represents different combinations of differential centrality and concentration

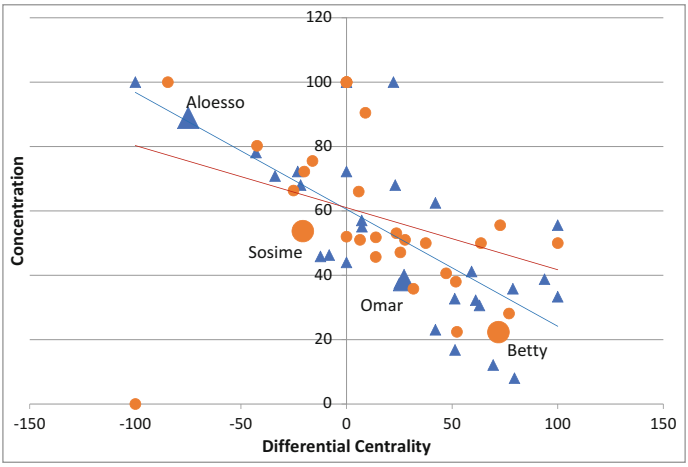

Fig. 5 Differential centrality (x-axis) and concentration (y-axis) of the 60 personal networks of the random sample, including the four example cases. Blue triangles and red circles (and the corresponding blue and red tendency lines) refer to male and female networks, respectively

Table 4 Four case examples (individual attributes and personal network indicators)

	Personal attributes				Personal network indicators				
Case	Gender	Age	Occupa-tion	Family status	Size	Compo-nents (Isolates)	Diffe-rential centrality	Concen-tration	Central alter
Betty	F	49	domestic worker	widow, 3 children	23	10 (7)	72	22.3	Spouse
Omar	M	28	tailor	2 wives, 2 children	20	5 (1)	27.3	38.5	Mother
Sosime	F	79	trader	widow, divorced, no children	19	6 (5)	-20.3	53.7	Father
Aloesso	M	44	farmer (former goldsmith)	1 wife, 3 children	18	2 (1)	-74.9	88.9	Brother

(Table 4 and Box 1). As one can see from Betty's and Aloesso's networks, cases of female solitary long-distant migration (yielding ego-centered, dispersed networks), or of male dependent family migration (yielding alter-centered, concentrated networks), are not unusual.

Betty's network represents the combination of high ego centrality and relatively low network integration typical for persons whose itineraries are shaped by high mobility outside the kinship network, such as teachers or soldiers, but also, as in the present case, female domestic workers. Besides the numerous isolated components she traversed in the course of her professional migrations, her social space also contains several more cohesive groups, one of which, organized around her husband, also channeled her professional itinerary.

Omar's network likewise represents a relatively poorly integrated social network, but ego's relative centrality here is much less pronounced. We again meet the combination of several cohesive components—an initial uterine network centered around the mother, a transnational migrant trader, a component of friends, and a marital component—each of which has been important for professional connections. There are fewer isolated components, also due to ego's younger age. The mother's central position results from the fact that she constitutes the bridge between the two spaces (Togo and Benin) between which ego is oscillating.

Sosime's network combines high network integration with a moderate decentralised position of ego. The high cohesion of the network results from the fact that it is largely composed of (active and passive) fosterage ties with close kin. After having been hosted by paternal relatives in childhood, she has subsequently hosted (and been accompanied by) a whole series of young female relatives, including the daughters of her former hosts, to be again hosted, in old age, by one of her former foster children. However, being childless herself, all these links are mediated through other members of her personal network, so that the high network concentration does not imply a high centrality of ego, but of the relative (her father) who serves as an intermediary. (Since fosterage typically implies a host-migrant relation, we did not include it into the relations of the personal network so as to avoid circularity. Otherwise, Sosime's centrality would have been considerably higher.)

Aloesso's network represents the extreme case of a highly integrated network where ego holds a marginal position (expressed by highly negative differential centrality). In fact, the network's main component is actually not his but his elder brother's network, while the second, marital component has not developed into a larger network, since ego's marriage coincided with a definitive return to sedentary life in the natal village.

Box 1 Personal network analysis of the four example cases

3.2 Sequence Networks

After having examined in a synchronic manner the way personal networks have been shaped by (and have shaped) people's itineraries, let us now turn to the question how these itineraries run through the various positions of these networks. We shall construct the corresponding relational sequence networks by classifying mobility events according to the type of relation or relational chain that links ego to his or her host. To facilitate comparison, we have used a very coarse classificatory grid, which, for example, lumps together all relatives other than the immediate parents into a single category,[13] and generally does not distinguish between different instances of the same relational type (such as "landlord"), with the only exception of spouses. Even so, except for the three individuals who have never left their natal social

[13]The software Puck allows for any kind of precision for the classification of kinship chains.

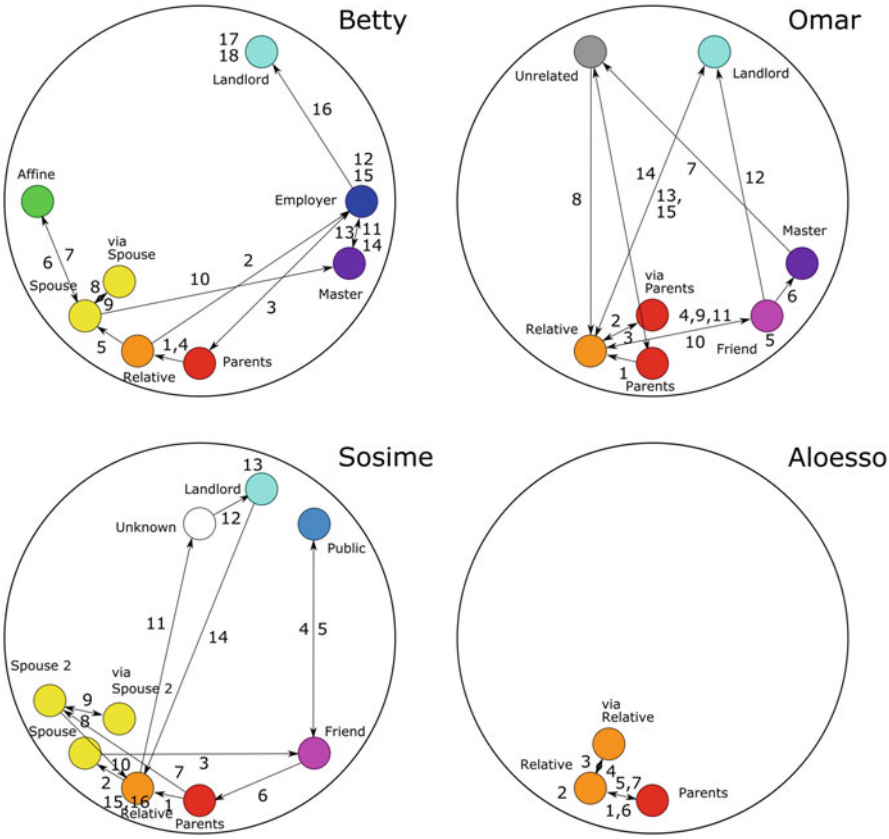

Fig. 6 Relational sequence networks of the four case examples. Nodes represent classes of events according to the type of relation between migrant and host; arcs indicate immediate succession of events of the corresponding type, numbered by order of occurrence in the itinerary. Numbers attached to nodes refer to loops (moves to new hosts within the same relational category). For the color code see Fig. 4

environment, the sequence networks of the 60 individuals of our random sample are all different from each other.

Figure 6 shows the relational sequence networks of our four example individuals, briefly discussed in Box 2. Each relational type is represented by a node. All itineraries start at the parental node (at the bottom), since the host of the first "mobility event" (birth) is by definition the birth mother.

The network representation of these relational sequences allows us to compare them by means of network-analytic indicators. Some of these indicators concern the local position of nodes—such as in- and outdegrees, which measure the diversity of relations that precede or succeed a given relation, or betweenness centralities, which

Aloesso's sequence network is the most simple: an oscillation between the parental home and a close maternal relative (his brother in Lomé), connected to an intermittent oscillation between the brother and a variety of hosts to whom Aloesso was only related by his brother's mediation. The simplicity of the sequence network is the corollary of Aloesso's marginality in the social network, by virtue of which the (orange) "relative" and "via relative" nodes absorb the quasi-totality of events.

Omar's sequence network is more complex: it also starts with a move to non-parental relatives (first his maternal aunt in Benin, later his maternal grandmother and uncles in the village), but he uses these positions as springboards to engage with hosts of various types linked to his professional education and activity, such as masters, compatriots, strangers, and finally landlords.

A quite similar basic pattern can be observed in **Betty**'s case: again starting with a move to a maternal aunt, she soon begins staying with various employers and landlords. However, this professional network is completed by a series of social positions to which she accedes via marriage: her husband, the husband's employers, and the husband's relatives (i.e. her affines). Note that this marital network is not distinct from, but integrated into the professional network: just as was the case with her own relatives during her youth, her husband's relatives are at the same time her employers or mediate her relations to future employers.

The most complex sequence network is **Sosime**'s, who combines a long-term shift from the paternal home (where she is hosted first by her father then by her brother) to several successive husbands with oscillating movements to and from the landlords of the capital and the friends and market places of Nigeria. The brother's home serves as a constant haven of return in these movements, and also as the crossroad where she "picks up" the young female relatives who accompany her through the various stages of her career as a trader.

Box 2 Sequence network analysis of the four example cases

indicate the importance of relations as transitions from or to other relations.[14] Others concern the global network structure—such as the number of dyads or cycles, which correspond to the frequency of direct or indirect returns to relations that appeared earlier in the itinerary. In Betty's case, for example, her frequent but intermittent accommodation by her employers is evident in the fact that the employer relation has both the highest betweenness centrality and the highest outdegree and forms part of the two oriented cycles of her sequence network. (For a detailed discussion of network indicators such as density, centralization or homophily in sequence analysis, see Cornwell 2018).

To which extent do these sequence patterns follow a gendered logic? To answer this question, we can proceed in two alternative ways. The first consists in examining the aggregate male and female networks constructed by merging all sequence networks of people of the same gender. In these networks, line values represent the numbers of individual itineraries in which a given event sequence occurs (see Fig. 7, where line values are indicated by different line widths). Table 5 provides some basic indicators for the network position of the various types of relations.

[14] See above footnotes 11 and 12.

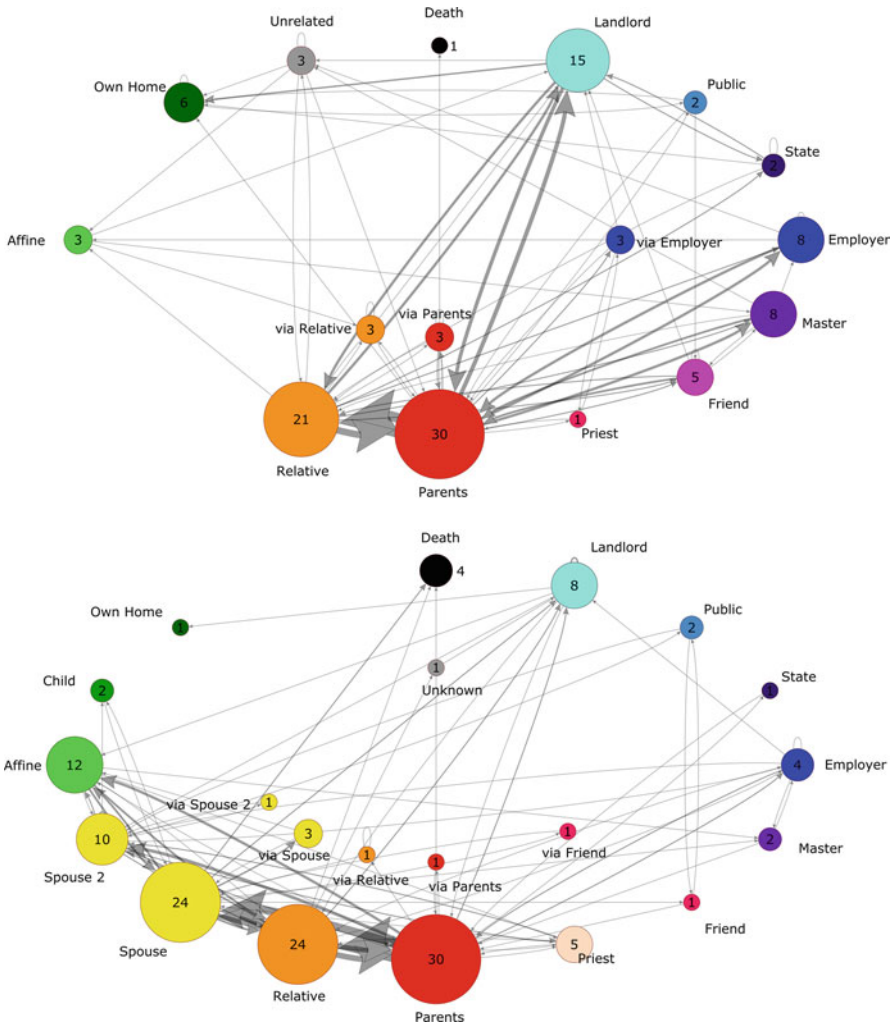

Fig. 7 Aggregate sequence networks of men (top) and women (bottom) (initial random sample, $n = 30$ each). For the color code see Fig. 4

As our four examples have already suggested, both male and female networks are centered around a fundamental axis linking parents and non-parental relatives. Not only are relatives the most frequent successors of parents as hosts (which is partly accounted for by the widespread practice of fosterage), but a person who has been hosted by a relative is also likely to return to the parental home. For both men and women, a stay with relatives provides the springboard for engaging in monetarized relations with landlords, which then tend to replicate themselves. However, men enter two times as frequently into relations with landlords than

Table 5 Basic indicators of the position of social relation types in male and female aggregate sequence networks: number of individual networks containing nodes of this type, normalized in- and outdegree and betweenness centrality of the node in the aggregate network (initial random sample, $n = 60$)

	Occurrences		Indegree		Outdegree		Betweenness	
	M	F	M	F	M	F	M	F
Parents	30	30	75.0	57.1	81.3	57.1	46.24	34.58
Relative	21	24	56.3	38.1	68.8	47.6	17.63	9.8
Landlord	15	8	50.0	28.6	37.5	28.6	8.56	10.07
Spouse		24		42.9		52.4		29.78
Spouse 2		10		42.9		42.9		18.7
Child		2		9.5		4.8		0.45
Affine	3	12	18.8	33.3	18.8	33.3	1.31	5.47
Employer	8	4	25.0	23.8	25.0	28.6	1.42	2.32
Master	8	2	25.0	14.3	31.3	9.5	1.07	
Own home	6	1	37.5	4.8	12.5		2.79	2.06
Friend	5	1	31.3	9.5	31.3	9.5	3.07	0.49
Priest	1	5	12.5	19.0	12.5	19.0		0.45
Public	2	2	12.5	9.5	18.8	9.5	6.67	1.1
State	2	1	18.8	4.8	25.0	4.8	0.12	
Unknown		1		4.8		4.8		
Unrelated	3		31.3		31.3			
Death	1	4	6.3	14.3			1.47	

do women, and are also more than twice as likely to progress from a stay with relatives to a rented apartment. The most significant difference is, of course, the importance of marital homes for female itineraries, this having no equivalent in their male counterparts. Note, however, that these spouse-linked positions are far from representing endpoints: their outdegrees exceed or equal their indegrees; in other words, the positions they lead to are more diverse than those that lead to them, indicating continuing residential mobility. Also note that the parental home, followed by relatives' homes, is the most frequent base from which women (re)join their first and subsequent husband(s)—it is extremely rare for women to move directly from one husband to another.

To sum up, the aggregate female sequence network, rather than being totally distinct from the male network, represents a variant of it, rendered more complex by the integration of one or more additional focal nodes. While the typical male network is largely organized around a triangle formed by the parental home, relatives and the residential market, the female network contains in addition one or more spouse nodes, which may become the nuclei of marital subnetworks. As a consequence, the relative importance of the various focal nodes (as measured by their betweenness centrality) differs between male and female networks. In male networks, the parental home clearly constitutes the dominant central node—as might be expected in a rural environment with agnatic transmission of land rights—

while in female networks, it shares importance with the almost equally central first marital home, followed by subsequent marital homes.

The problem with this kind of comparative macro-analysis is, of course, that it already presupposes that gender is a pertinent classification criterion for merging individual sequence networks, instead of deriving the classification criteria from a comparative analysis at the micro-level of individual sequence networks. This latter perspective characterizes the approach of optimal matching analysis (Abbott 1995) and its more recent network-analytic analogues, where sequence matching is replaced by graph similarity (Butts and Pixley 2004; Butts 2008; Fitzhugh et al. 2015). Since in our relational sequence networks each node type appears, by construction, at most once in each network (the only quasi-exception being first and subsequent spouses, which we have treated as distinct types), we can construct all sequence networks with identical node sets (absent nodes being left unconnected), so that the only edit operations necessary to transform one network into another consist in adding or removing arcs. Graph edit distance thus boils down to the (normalized) number of differently connected pairs of nodes (that is, the number of different cells in the adjacency matrices of the two graphs). Based on this graph distance matrix, hierarchical clustering techniques can be applied to represent the similarities and differences of the relational sequence networks. Figure 8 displays the agglomerative clustering tree of the 60 sequence networks of the initial random sample, constructed by the neighbor-joining algorithm (Saitou and Nei 1987),[15] and spatialized by the Kamada-Kawai spring embedded algorithm implemented in Pajek. The graphs of the individual sequence networks have been plotted into the nodes of the tree, and border colors indicate gender (blue for male and red for female). Our four example cases are emphasized by the use of grey shades.

While the gender difference is neatly brought out by the graph (roughly speaking, female itineraries are located in the upper half, male itineraries in the lower half of Fig. 8), a closer look shows that, beyond the great divide introduced by virilocality (the presence or absence of "yellow" marital nodes in the sequence network), both male and female sequence networks are differentiated into several branches of the tree.

Most sequence networks, irrespective of gender, are based on the parent-relative axis, to which the simplest of them (such as Aloesso's) may reduce. In the "male" case, this basic axis frequently combines either with rent (as for the urban migrant workers), with hosting by the master or employer (as for the rural itinerary masons or woodcutters), or with both. This combination, eventually enriched by marital nodes, may also characterize the sequence networks of female domestic workers (such as Betty). A still more complex variant, characteristic of teachers' itineraries, combines rented accommodation with lodging by the state. Finally, the itinerary

[15]The neighbor-joining algorithm consists in sequentially linking the two nearest neighbors to a newly created "ancestor" node and recalculating the distances of all other nodes to the ancestor node from their distances to the joined neighbor nodes. Pairwise distances are adjusted by the average distance of both neighbors to the rest of the network. In Fig. 8 ancestor nodes are only represented as bifurcation points of outgoing arcs.

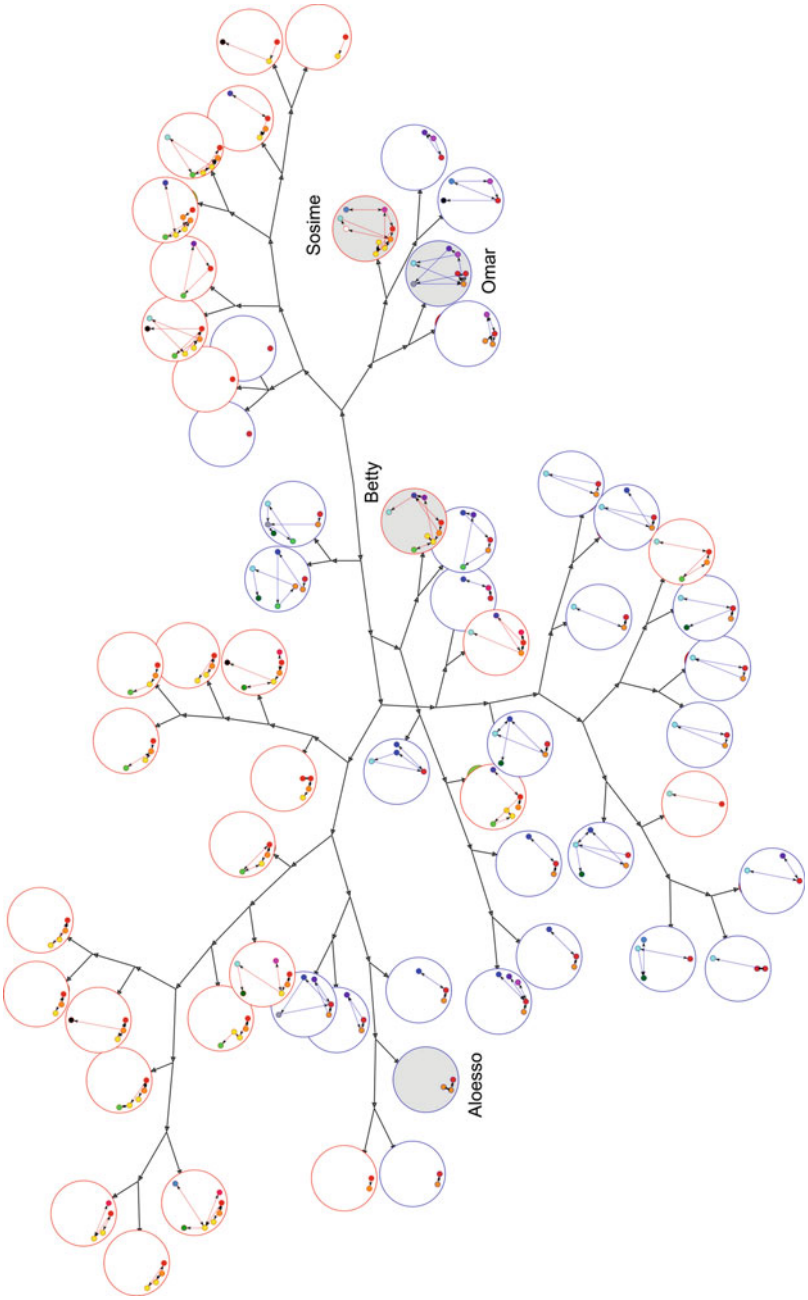

Fig. 8 Clustering tree of the 60 sequence networks of the random core sample, generated by Saitou-Nei neighbor-joining algorithm. Individual sequence networks have been inserted into the nodes of the tree. Visualization by Pajek, using Kamada-Kawai spring embedded algorithm. For the color code see Fig. 4

may involve lodging by friends or unrelated persons, frequently in a transnational context, as exemplified by Omar's case. Again, this pattern may be equally found among female traders (such as Sosime). A quite different type of itinerary is located at the bottom left of the tree, where the parental node is directly linked to the landlord node without passing through a relative's home. This is a minority pattern generally restricted to men who have migrated later in life for professional reasons, and to the few boys whose parents could afford to rent apartments for them while they attended high school. However, we also find this itinerary present in two women who chose not to marry.

While all the predominantly "male" patterns of sequence networks are accessible to women, the "female" upper half of the graph is significantly more homogenous. The upper left part in particular is made up of a large group of highly similar sequences completing the parent-relative-axis by one or more marital or affinal nodes, thus giving rise to several distinct clusters according to the complexity of the marital subnetwork. This pattern may further evolve by integrating stays with landlords, masters or employers (mainly for trade or domestic work), as can be seen on the upper right side (actually a female enclave encompassed by predominantly male patterns). By contrast, we almost never find a female pattern without the "relative" node in it. Women who leave their parental home directly for a rented apartment are even more rare then men (though, also here, the exception proves the rule).

4 Conclusion

Our analysis of the topology of both personal and sequence networks has produced convergent results with regard to the logics underlying gendered mobility patterns in South-east Togo. Male and female sequence networks both rest on a basic kinship axis (linking an "internal" parent pole and an "external" extended-kinship pole). For both genders (though more so for men), the parental home is a central place of (definitive or transitory) return, and for both genders, relatives outside the immediate family circle play an equal (for women even a stronger) role than parents, in particular before the age of marriage. The simplest migration pattern, and the starting point for most male and female itineraries, is an oscillation between parents in the village and relatives beyond.

The main difference between the genders consists in the way their respective social sequence networks develop from this double kinship core. Due to the sexual division of labor which renders wage-labor still an essentially male domain (women are either independent traders or work for relatives as domestic workers and market assistants), and to the virilocal residence rule, according to which women at least initially join their husbands' home, the networks of men tend to evolve through non-kinship links to friends, masters, employers, strangers or landlords, while women's networks grow through the emergence of affinal links to spouses and in-laws.

This difference is not only one of relational circuits but also of network morphology. Male non-kin relations, often radiating from the core in a star-like manner, are generally multiple but socially disconnected (except when they are linked to the kinship complex). By contrast, female affinal relations, though less numerous, are structurally productive: they develop into larger subnetworks that may succeed each other as disjoint spheres or merge into an integrated space of circulation. While male networks thus tend to evolve through a succession of structurally isolated non-kinship links, female networks develop into complex multifocal networks sewn together by marital and affinal ties—that is, the very ties that link together male and female itineraries. In other words, the central source of difference between the sequence networks of men and women is precisely their mutual relation.

Rather than just confirming the macro-tendencies for male and female mobility patterns (as stated in the demographic literature) at the micro-level of individual trajectories, sequence network analysis yields insight into the relational logics that bring these tendencies about. It serves not only to study the differences between gendered social networks, but also to understand gender itself as a relation between networks, that is, not just as an attribute of individuals, but as a structural trait of social space-time. In a more general perspective, the nascent integration of network and sequence analysis may be the first step towards a full-fledged social topology.

Acknowledgements The paper is part of a larger research project conducted with Karin Sohler since 2010. Preliminary results were the subject of joint presentations at the 1st European conference of Social Network Analysis in Barcelona in 2014 and at the 3rd meeting Réseaux-Histoire in Paris in 2015. Though Karin did not coauthor the present paper, its argument owes a fundamental debt to our collaborative work. I am also grateful to Véronique Hertrich, Ismaël Moya, Sorana Toma, Claire Lemercier and Karen Middleton, as well as to the anonymous reviewers of the LaCOSA II conference and for this volume, for their valuable and encouraging comments on earlier versions of this paper.

Fieldwork has been supported by the Laboratoire d'Anthropologie Sociale, the EHESS research fund and the French Cooperation in Togo. The 2010 fieldwork was carried out in cooperation with Ibitola Tchitou and Kokou Vignikin from the Unité de Recherche Démographique of Lomé University. This work would have been impossible without the competence and dedication of my field collaborators Komi Malou Kakanou and Toussaint Yakobi and the active cooperation and help of the interviewees, to whom I present my heartfelt thanks.

The sequence-network module of Puck has been developed with Christian Momon as part of the project "Kinsources" supported by the French National Research Agency ANR (research grant ANR-12-CORP-0008).

References

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21*(1), 93–113.

Abbott, A. (1997). Of time and space: The contemporary relevance of the Chicago School. *Social Forces, 75*(4), 1149–1182.

Batagelj, V., & Mrvar, A. (2008). Analysis of kinship relations with Pajek. *Social Science Computer Review, 26*(2), 224–246.

Bearman, P. S., Faris, R., & Moody, J. (1999). Blocking the future: New solutions for old problems in historical social science. *Social Science History, 23*(4), 501–533.

Bearman, P. S., & Stovel, K. (2000). Becoming a Nazi: A model for narrative networks. *Poetics, 27*(2), 69–90.

Bison, I. (2014). Sequence as network: An attempt to apply network analysis to sequence analysis. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 231–248). Cham: Springer.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology, 38*(1), 155–200.

Butts, C. T., & Pixley, J. E. (2004). A structural approach to the representation of life history data. *The Journal of Mathematical Sociology, 28*(2), 81–124.

Cordell, D. D., Piché, V., & Gregory, J. W. (1996). *Hoe and wage: A social history of a circular migration system in West Africa*. Boulder: Westview Press.

Cordonnier, R. (1987). *Femmes africaines et commerce: Les revendeuses de la ville de Lomé (Togo)*. Paris: l'Harmattan.

Cornwell, B. (2015). *Social sequence analysis: Methods and applications*, Volume 37 of *Structural analysis in the social sciences*. Cambridge: Cambridge University Press.

Cornwell, B. (2018). Network analysis of sequence structures. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Cornwell, B., & Watkins, K. (2015). Sequence-network analysis: A new framework for studying action in groups. In S. R. Thye & E. J. Lawler (Eds.), *Advances in group processes* (Vol. 32, pp. 31–63). Bradford, U.K: Emerald Group Publishing Limited.

Fitzhugh, S. M., Butts, C. T., & Pixley, J. E. (2015). A life history graph approach to the analysis and comparison of life histories. *Advances in Life Course Research, 25*, 16–34.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*(1), 35–41.

Hamberger, K. (2011). *La parenté vodou: Organisation sociale et logique symbolique en pays ouatchi (Togo)*. Paris: CNRS Éditions, Éditions de la Maison des sciences de l'homme.

Hamberger, K., Grange, C., Houseman, M., & Momon, C. (2014). Scanning for patterns of relationship: Analyzing kinship and marriage networks with Puck 2.0. *The History of the Family, 19*(4), 564–596.

Hamberger, K., Houseman, M., & White, D. (2011). Kinship network analysis. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 533–549). London: SAGE.

Liu, M.-M. (2013). Migrant networks and international migration: Testing weak ties. *Demography, 50*(4), 1243–1277.

Locoh, T., & Thiriat, M.-P. (1995). Divorce et remariage des femmes en Afrique de l'Ouest. Le cas du Togo. *Population (French Edition), 50*(1), 61–93.

Lubbers, M. J., Molina, J. L., Lerner, J., Brandes, U., Ávila, J., & McCarty, C. (2010). Longitudinal analysis of personal networks: The case of Argentinean migrants in Spain. *Social Networks, 32*(1), 91–104.

Pilon, M., & Ségniagbéto, K. (2014). Confiage, domesticité et apprentissage à Lomé à la veille de l'indépendance. *Journal des Africanistes, 84*(1), 212–247.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*(4), 406–425.

Stark, D., & Vedres, B. (2006). Social times of network spaces: Network sequences and foreign investment in Hungary. *American Journal of Sociology, 111*, 1367–1411.

Stark, D., & Vedres, B. (2012). Social sequence analysis: Ownership networks, political ties, and foreign investment in Hungary. In J. F. Padgett & W. W. Powell (Eds.), *The emergence of organizations and markets* (pp. 347–374). Princeton: Princeton University Press.

Wissink, M., & Mazzucato, V. (2018). Transit: Changing social networks of sub-Saharan African migrants in Turkey and Greece. *Social Networks, 53*, 30–41.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part IV
Unfolding the Process

Multiphase Sequence Analysis

Thomas Collas

1 Introduction

Many common notions such as lifecycle, adulthood, turning point or ratchet effect are based on the idea that various sets of sequences—especially individual careers—follow regular patterns defined by successive phases. Those phases that all cases encounter in the same order are usually defined by social moments—such as graduation, medical treatment, childbirth, job promotion, new research project, election as a congressman, Oscar nomination, etc.—or by calendar periods—e.g., minutes, months, years or decades. Such a conception of sub-level temporal structures nested in sequences and linked to one another is a cornerstone of life-course studies (Levy and The Pavie Team 2005). Nonetheless, it has rarely been discussed as a methodological parameter of sequence analysis.

Implementing the notion of phase can in particular deepen our understanding of how institutionalised narratives shape social processes (Abbott 2001). Dividing into phases also helps to reduce data complexity. In this chapter, I elaborate on these ideas by presenting visual and metric tools of multiphase sequence analysis (MPSA). Throughout the text, I will develop an example of two-phase sequences drawn from a study of the careers of participants in professional *pâtissier* (pastry cook) competitions in France.

In the first section, I present key properties of multiphase sequences. In the second section, I exemplify and discuss the implications of the division into phases. In the third section, I present tools to render multiphase sequences, introducing sliced representations. In the fourth section, I introduce a dissimilarity measure of multiphase sequences called multiphase optimal matching (MPOM).

T. Collas (✉)
F.R.S.-FNRS – Université de Louvain, Louvain-la-Neuve, Belgium
e-mail: thomas.collas@uclouvain.be

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_9

2 Sequences as Multiphase Structures

The pathbreaking proposition of sequence analysis was to study sequences as wholes (Abbott 1995), constructed as successions of atemporal timeslots.[1] The notion of phase adds an intermediary level to this binary hierarchical structure (sequences and time slots) by introducing discontinuities within these successions.

2.1 Characteristics of Multiphase Sequences

Lesnard and Kan (2011) provide an example of multiphase sequences in their analysis of workweeks. The whole sequences they study are each made up of 672 15-min time slots. Lesnard and Kan represent them as successions of 7 days of 96 15-min time slots with a two-state alphabet (work/non-work). For each sequence, seven successions (days) are nested in a larger succession (a week). This example of multiphase sequences is extremely regular: all sequences are made up of seven phases which contain exactly 96 time slots. It is possible to conceive of more diverse types of multiphase sequences.

First, the phases may be of uneven length within sequences and from one sequence to another. For instance, in France, whenever a new government is formed, many recruits enter the ministerial *cabinets*. Their job careers can be divided into three phases: their positions before entering the *cabinets* (for example, by separating private and public sectors), their positions within the *cabinets* (by differentiating *cabinets* and hierarchical positions) and their positions after they leave these *cabinets* (with the same alphabet as in the first phase). The lengths of the different phases differ because they are not defined according to their duration but with respect to common events that may happen at different moments in a career. Take the first phase for example: some enter a *cabinet* after a 20-year long career whereas others have only worked for three or four years before being recruited. More broadly, the division into phases often implies length differences even for sequences of equal length.

Secondly, some phases may be empty, i.e., of length zero. For example, *pâtissier* competitions suppose a division of careers into two phases: a phase in which competitors are employed as a junior or apprentice *pâtissier*, followed by a phase in which they are a fully-fledged *pâtissier*. Nonetheless, some competitors may quit the trade just after their apprentice period or enter the trade directly as senior workers.

These characteristics lead to a formal definition. Multiphase sequences S_1 and S_2 are successions of n phases such as $S_1 = (\zeta_1^{S_1}, \zeta_2^{S_1}, \ldots, \zeta_p^{S_1}, \ldots, \zeta_n^{S_1})$ and $S_2 = (\zeta_1^{S_2}, \zeta_2^{S_2}, \ldots, \zeta_p^{S_2}, \ldots, \zeta_n^{S_2})$, where $\zeta_p^{S_1}$ is phase p in sequence S_1 and $\zeta_p^{S_2}$ is phase p in sequence S_2. The length of each phase varies from 0 to the length of the sequence in which it is nested. If $n = 1$, we encounter the usual definition of sequences as continuous series of time slots.

[1] A time slot is atemporal since a single state is observed from the beginning to the end of it.

2.2 Two Formal Properties of Phases and Two Methodological Assumptions

Two formal properties of phases are captured by the "turning point—trajectories" model theorised by Abbott (2001, p. 253). In this model, turning points differentiate consistent episodes (trajectories) and link these episodes to the previous and following ones into a larger narrative. The first property is relative consistency: the division of a sequence into phases relies on the assumption that the succession of states within each phase is both consistent and different from successions within other phases.[2] The second property is processual location: a phase is defined by its position within a sequence, hence by its position relative to other phases. Returning to the example of the workweek from Lesnard and Kan (2011), Friday is regarded as a consistent period for work scheduling and its position within the workweek immediately preceding Saturday distinguishes it from other days.

Two crucial methodological assumptions follow these formal properties.

First, as sequences, phases are approached as sites (or locations) of narratives: as successions of time slots each containing one state, phases and sequences are constructed as compartments for modelling narratives.[3] To assert that a phase is relatively consistent and located in a sequence does not imply any assumption about the narrative it contains. "Stage" models in the traditions of Piaget or Parsons (for a synthesis, see Levy and The Pavie Team 2005) define the content of both typical sequences and typical phases within these sequences (the so-called stages). These models assume that each phase is the location of a single and specific kind of narrative (for instance, a certain behavioural development or a type of activity). The notion of phase discussed here helps one to appreciate the relevance of stage models but is not bound to these models. While comparing sequences, one assumes that there are patterns, i.e., types of narratives, to be discovered. While comparing multiphase sequences, one assumes that phase-structured patterns—types of narratives including types of sub-narratives—are to be discovered. As for sequence analysis at large, the only assumption about the content of these sub-narratives lies in the alphabets that are used and in the definition of the boundaries of the set of sequences. These elements limit the universe of possible narratives. Identifying these consistent sub-narratives is a different question from dividing into phases, just as identifying consistent narratives is a different question from delimiting a population of sequences.

Second, distinct phases are dissociated and should not be compared. In the workweek example, it would make little sense to compare one sequence's Tuesday to another's Sunday. Similarly, it does not seem relevant to directly compare contests

[2] As Abbott (2001) and Cornwell (2015, p. 94) indicate, this consistency can be conceived as a set of stable relationships between states as modelled by Markov chain models. A turning point is defined by Abbott (2001, p. 247) as a transition separating stable probability regimes.

[3] By methodological construction and to illustrate the notion of site, a narrative cannot be contained in a time slot, which can only contain one state.

in which apprentice *pâtissiers* compete with contests where senior *pâtissiers* compete. This incommensurability of phases has major implications for the comparison of multiphase sequences. I shall return to this point after discussing the practical operation of dividing into phases.

3 Division into Phases: Reference Frame, Alphabet(s) and Phase-Structure

This section starts with an example of division into phases before distinguishing three crucial aspects of this operation.

3.1 A First Hint: The Extended Example

Careers of participants in *pâtissier* competitions in France offer a case of two-phase sequences. The data are drawn from results of 2060 professional competitions. These competitions consist in making or presenting decorative sculptures (out of sugar, chocolate or ice), cakes or plated desserts before juries of peers.[4]

Here I present data from 777 *pâtissiers* who participated in two to 21 competitions and whose careers of participations in competitions began before 2002 and ended after 1990. Each time slot is a participation in a competition. Each participation is defined by an age category (apprentice, junior, senior), a rank (1st, 2nd, etc.) and a type of competition (preceded or not by screening contests).[5]

To be compared, these careers have been divided into two phases with respect to age categories since some contests are for apprentice and junior competitors only, others for senior competitors. This is a case of institutional division of careers into phases that is verified in the data. Indeed, returns to apprentice and junior competitions from senior competitions are unusual: amongst the competitors who participated in at least two competitions and at least one senior competition, only 7.6% participated in an apprentice or junior competition after competing in a senior competition. Interviews with *pâtissiers* within a larger research project made it clear

[4]For a detailed account, see Collas (2015).

[5]Data were gathered from archives and trade press collections. 2060 rankings covering the period 1933–2012 were coded. *Le Journal du pâtissier* (published since 1978), which is the main source, mentions competitions that are organised in different areas in France, while the other sources mention mainly competitions taking place in Paris. Beginning (before 2002) and end (after 1990) dates were chosen because of source heterogeneity and in order to limit right and left censoring. A first rank in the "*Un des meilleurs ouvriers de France*" competition was not kept since it was always gained at the end of a competitor's career. R software (R Core Team 2014) and the TraMineR package (Gabadinho et al. 2011) were used to visualise sequences, to compute OM-distances, to extract sets of representative sequences and to compute other sequence-related metrics.

that first participation in a senior competition is represented as an entry into a phase of evaluation which is not based on age or scholarship, but on the fact of being identified as a *pâtissier*. Thus, I postulate that participations in competitions preceding participation for the first time in a senior competition are not comparable with subsequent participations.

The dissociated phases are, first, the *ante*-senior phase—including only participations in apprentice and junior competitions and ending with the last participation before a participation in a senior competition or with the last participation if the competitor has never participated in a senior contest—and, second, the senior phase—which begins with the first participation in a senior competition and can include any type of participation. Each phase is defined by its relative position within the sequence and by its postulated internal consistency.

Since many competitors participate only in one type of competition (senior ones on the one hand, junior and apprentice ones on the other), a large fraction of sequences include a phase that does not contain any participation (an empty phase, i.e., of null length): these sequences are made up of participations in competitions during only one phase. The senior phase includes participations in 84.4% of the sequences, the *ante*-senior phase in 44.7% of the sequences.

Division into phases impacts the sequences' states alphabet(s). In that example, division leads to a shorter alphabet and, as a consequence, reduced data complexity: age categories are not taken into account in the definition of each participation since phases already bear this age dimension.

Two dimensions are used to define the alphabet. First, competitions preceded by screening contests are regarded as distinct. Amongst these competitions, two are singular and thus isolated as distinct states in the alphabet: a national plated dessert competition named *Championnat de France du dessert* (CFD) (labelled "C") and the oldest competition preceded by screening contests still organised today named *Un des meilleurs ouvriers de France* (One of the best craftsmen in France exam, labelled "M"). This competition is for senior competitors only. Several other competitions are gathered under the "S" label (for Screening contests).[6] Other types of competitions are labelled "W" (for Without screening contests). Second, each participation is defined by the rank awarded, in three categories: 1st rank (labelled "L", for laureate), 2nd or 3rd rank ("P", for podium) and 4th rank or below ("O", for off-the-podium). For example, a state "LC" indicates a first rank in the *Championnat de France du dessert*.

In order to include long pauses between two participations in the analysis, a state named "4Y" (for four years) was created to indicate every period lasting more than four years and less than eight years between two successive participations. The alphabet contains eleven possible states (Table 1). One, related to a senior competition (OM), is observed only during the senior phase.

[6] *Meilleur Apprenti de France* (Best Apprentice in France), *Meilleur Apprenti du Monde* (Best Apprentice in the World), *Coupe du Monde de la Pâtisserie*, *Grand prix international de la chocolaterie*, World Chocolate Master.

Table 1 States alphabet

Competition	Rank	State
Competition without screening contests	1st	LW
	2nd or 3rd	PW
	4th or below	OW
Championnat de France du dessert, CFD	1st	LC
	2nd or 3rd	PC
	4th or below	OC
Competition preceded by screening contests (other than CFD or MOF)	1st	LS
	2nd or 3rd	PS
	4th or below	OS
Un des meilleurs ouvriers de France, MOF (only in Phase 2)	Unranked finalist	OM
Four years pause between two successive participations		4Y

3.2 Three Aspects of Division into Phases

This case sheds light on three aspects of division into phases.

The first one is the reference frame of the division. Here, the reference frame is endogenous: two phases of participations in competitions are dissociated according to a characteristic of participations in competitions (the first participation in a senior competition). In other cases, the reference frame is exogenous: an external dimension is introduced in relation to a research question. For example, careers of participants in competitions could be divided according to job positions. Taking another example, in many systems, academic careers are structured by a limited number of phases (lecturer, assistant professor, associate professor, etc.). The dissociation of phases of academic activities (e.g., publications) according to these successive academic ranks helps to explore how institutionalised episodes impact scientific outputs. Such a division into phases is also a way to reduce data complexity: what can be regarded as two distinct channels (Pollock 2007; Gauthier et al. 2010)—academic ranks and publications—are reduced to one channel cut into successive phases.

The second aspect is the definition of the alphabet(s). Here, the phase division reduces by a half the number of possible states with a very limited loss of information regarding the age categories of each participation (only the above mentioned 7.6% of senior competitors are affected). But the division into phases can also accompany a definition of several alphabets. Since phases are regarded as consistent episodes within sequences, the number of relevant states for each phase may be quite low, especially when the division is endogenous. Such is the case for the careers in ministerial *cabinets* mentioned above. The types of possible positions during the *cabinet* phase are both more limited and more specific than before and after this phase. As a consequence, two alphabets may be defined. By delimiting a relevant universe of possible states for each phase, the plurality of alphabets reduces data complexity.

Determining the number and level of phases is a third key aspect to phase division. The case developed here is simple: only two successive phases are dissociated. As illustrated by several examples mentioned, this number can be higher, but the number of assumptions about the structure of the sequences rises accordingly. We can also envision sub-phases nested within phases. A tennis match is such a two-level nested structure: points are clustered within games and games within sets. Careers including gradations within ranks (2nd class, 1st class, etc.) such as academic careers in France present such a nested structure.

4 Rendering Multiphase Sequences

Two types of graphical representations help to render multiphase sequences. The event-aligned variety of simple alignment representations is suited for two-phase sequences. Sliced representations, introduced here, generalise the logic of event-aligned representations to *n*-phase sequences.

4.1 Simple Alignment on a Specific Event

Drawing on previous studies (Blanchard 2010; Giudici and Gauthier 2009), Colombi and Paye (2014) discuss a visualisation method that transforms the usual left- or right-aligned representation suited for one-phase sequences into a representation aligned on a specific event. This event is regarded as a turning point between two phases: "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)" (Colombi and Paye 2014, p. 250).

Two steps are followed. First, a relative time axis aligned on the specific event under study is introduced: the temporal scale is negative for states observed before this event, positive for states observed after. The format of each time slot is preserved but the time axis is distorted so as to preserve a social timing according to a supposed turning point. Second, blank time slots are inserted at the beginning and end of each sequence. As a result, the length of every sequence is equal to the length of the longest observed period preceding the studied event summed with the length of the longest observed period following this event.

Taking a toy example, if *A* marries at age 25 and *B* at age 28, their job sequences from 22 to 31 are left-aligned and event-aligned (on marriage date) as shown in Table 2 (E stands for employment, U for unemployment, m on the axis stands for marriage).

Table 2 Different alignments of sequences

Left-aligned sequences, age axis

A	U	U	U	U	E	E	E	E	U	U
B	U	U	E	E	E	E	E	E	E	U
Axis	22	23	24	25	26	27	28	29	30	31

Event-aligned sequences, event-relative axis

A				U	U	U	U	E	E	E	E	U	U
B	U	U	E	E	E	E	E	E	E	U			
Axis	−6	−5	−4	−3	−2	−1	m	1	2	3	4	5	6

Multiphase sequences, left-aligned sliced axis

A	U	U	U				U	E	E	E	E	U	U
B	U	U	E	E	E	E	E	E	E	U			
Axis	s1	s2	s3	s4	s5	s6	m1	m2	m3	m4	d1	d2	d3
Phase	Singlehood						Marriage				Divorce		

4.2 Multiple Alignment by Sliced Representation

Event-alignment is convenient for keeping the continuous representation of left-aligned sequences but such a technique is limited to two-phase sequences. Sliced representations help to render n-phase sequences. In the previous example, if one wants to include divorce in the reference frame of division and if *A* gets divorced at 29, the representation implies three dissociated phases (see bottom subtable in Table 2).

As for event-aligned representations, the temporal scale is relative, but the time axis is indexed on phases, not on a single event. In this example, the axis is left-aligned for each phase. The origin point is the first time slot of each phase. In case of a right-alignment, the origin point is the highest possible length for a given phase. A right-alignment for certain phases and a left-alignment for others can be envisioned. In that sense, the event-aligned representation is a special case of sliced representations.

Figure 1 shows three representations of the same sample of five sequences of participations in competitions. Division into phases underlines the proximity of the last two sequences during the senior phase. While the event-aligned representation highlights the sequences' continuity as the usual left-aligned or right-aligned representations of sequences, sliced representation focuses on the different successions within each phase, not on the sequels and aftermath of a supposed turning point.

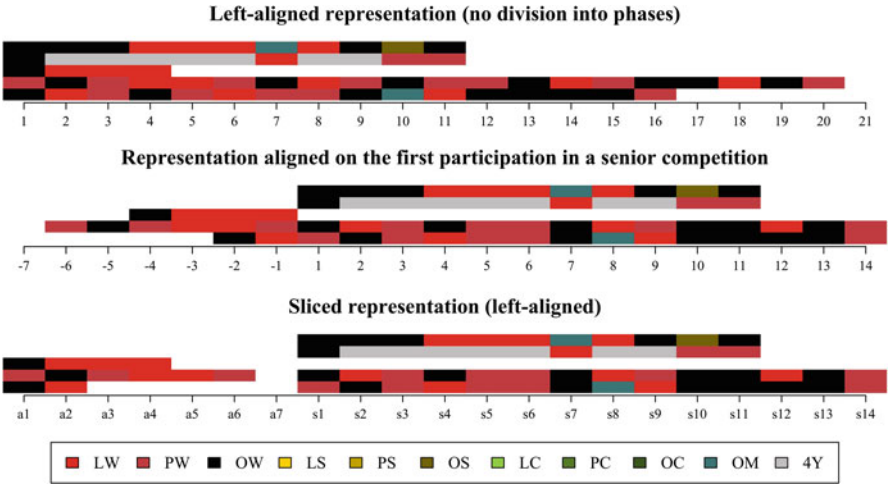

Fig. 1 Three representations of the same careers of participations in competitions (see explanation of state labels in Table 1)

5 Measure and Interpretation of Pairwise Distances Between Multiphase Sequences: Multiphase Optimal Matching

Besides visualisation, dissimilarity measures are commonly used tools for comparing sequences. When measuring dissimilarities between multiphase sequences, phases are regarded both as dissociated incommensurable episodes and as sites of narratives. These are the basic principles of multiphase optimal matching (MPOM) introduced here. This section focuses on optimal matching (OM) for three reasons: it is a seminal and widespread dissimilarity measure in the social sciences; the cost definition operations that OM implies have consequences for MPSA; the principles of OM are adapted to the example under study (this point is discussed below). Nevertheless, MPOM's analytical logic can be extended to other dissimilarity measures when they are applied to multiphase sequences.[7]

5.1 Analytical Logic

MPOM's analytical logic is twofold. First, pairwise distances between sequences are measured with respect to equivalent phases. Time slots belonging to Phase P_1 in Sequence S_1 are only compared with time slots belonging to Phase P_1 in Sequence S_2. Second, each phase is regarded as an ordered set of time slots. Equivalent phases

[7]For a review of dissimilarity measures between sequences, see Studer and Ritschard (2016).

are compared with the three basic operations of OM (Abbott and Forrest 1986): substitution, insertion and deletion. Costs assigned to these operations are defined for each phase.[8]

For each pair of sequences, MPOM measures a distance per phase (distance between S_1 and S_2 on Phase P_1, distance between S_1 and S_2 on Phase P_2, etc.) through OM operations and then a distance between sequences by summing distances between equivalent phases. The matrix of pairwise distances between sequences is the sum of the matrices of pairwise distances per phase. Thus, the contribution of each phase to the distance between two sequences depends on the differences in state composition of this phase from one sequence to another and on its maximal length (the longer a phase is, the heavier its impact on pairwise distance can be, due to the number of insertions and/or deletions necessary to edit one phase into another).

MPOM-measure is a fractal generalisation of OM measure. In the case of a one-level phase division, each phase is approached as OM approaches a sequence. In the case of a higher-level phase division (with multiple levels of phases, phases nested within phases), the same operation is reproduced at each level. Thus, an n-level MPOM-distance implies nested sums. For instance, a two-level MPOM-distance (see the tennis match example above) is a sum of one-level MPOM-distances.

Regarding empty phases, if Phase P_1 is of length $l_1^P = 0$ in S_1 and of length $l_2^P \geq 1$ in S_2, the impact of emptiness on the distance between S_1 and S_2 is equal to the cost of insertion and deletion ($indel$) multiplied by the length of the longest version of Phase P_1 (here l_2). More broadly, the impact of a length difference between two sequences is equal to this difference multiplied by the indel ($|l_2 - l_1| \times indel$).[9]

Thus, MPOM rests on two methodological principles that can be applied to other dissimilarity measures: dissociation of phases and combination of phase pairwise-distances into sequence pairwise-distances.

5.2 MPOM Applied to Careers of Participants in 'Pâtissier' Competitions

Turning to careers of participants in *pâtissier* competitions, optimal matching measure offers an appropriate tool for searching for regular patterns in these data for two reasons. First, these sequences are characterised by a high level of instability from one participation to another regarding ranking (one may rank first, then ninth,

[8]This involves a multiplication of cost-setting operations which may seem dubious since many criticisms of OM have focused on cost-setting operations (Abbott and Tsay 2000). Constant or data-driven substitution costs may be relevant in some cases.

[9]Since division into phases often implies length differences, pairwise distances between phases can be standardised with respect to the maximal possible distance for each phase, which I do not do here precisely in order to take length differences into account.

Table 3 Substitution costs

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1 - LW	0	1.8	1.9	1.95	1.85	1.7	1.7	1.95	1.85	1.7	2
2 - PW	1.8	0	1.7	1.99	1.95	1.92	1.92	1.99	1.95	1.92	2
3 - OW	1.9	1.7	0	2	1.98	1.95	1.95	2	1.98	1.95	2
4 - LC	1.95	1.99	2	0	1.8	1.9	1.9	1	1.8	1.9	2
5 - PC	1.85	1.95	1.98	1.8	0	1.7	1.7	1.8	1	1.7	2
6 - OC	1.7	1.92	1.95	1.9	1.7	0	1	1.9	1.7	1	2
7 - OM	1.7	1.92	1.95	1.9	1.7	1	0	1.9	1.7	1	2
8 - LS	1.95	1.99	2	1	1.8	1.9	1.9	0	1.8	1.9	2
9 - PS	1.85	1.95	1.98	1.8	1	1.7	1.7	1.8	0	1.7	2
10 - OS	1.7	1.92	1.95	1.9	1.7	1	1	1.9	1.7	0	2
11 - 4Y	2	2	2	2	2	2	2	2	2	2	0

then third, etc.); insertion/deletion operations reduce distance due to lags. The second reason is relations between states. A first rank is closer to a second rank than to a tenth rank. Multiple substitution costs lie in the postulate of distance instead of strict difference between states.

So as to limit the number of assumptions, the same substitution costs between states are used for both phases. These costs have been set manually with respect to states' formal closeness: competitions with screening contests are closer to one another than to competitions without screening contests; a first rank is closer to a second rank than to a fourth rank; a second or third rank is closer to a fourth rank than to a first rank (due to the singular situation of ranking first). Minimum substitution cost between two distinct states is 1, maximum cost is 2. Minimum cost is used for a substitution between two states indicating the same position in two competitions preceded by screening contests (e.g., CFD and World Cup). Other substitution costs are set between 1.7 and 2 (Table 3). Several trials intended to take into account the unequal length of sequences without making it the first criterion of distance lead me to set insertion-deletion cost (*indel*) at 1.35 for both phases.[10]

[10]This indel is higher than the minimum substitution cost for distinct states (1) and lower than the minimum substitution cost for states differing in terms of both ranking and screening (1.7). For example, it is less costly to edit one sequence (ABC) into another sequence which is one time slot shorter and otherwise identical (AB) than to edit one sequence into another same-length sequence which is different only for the last time slot (ABC into ABD) only if states C and D are similar with respect to ranking or anterior screening contests.

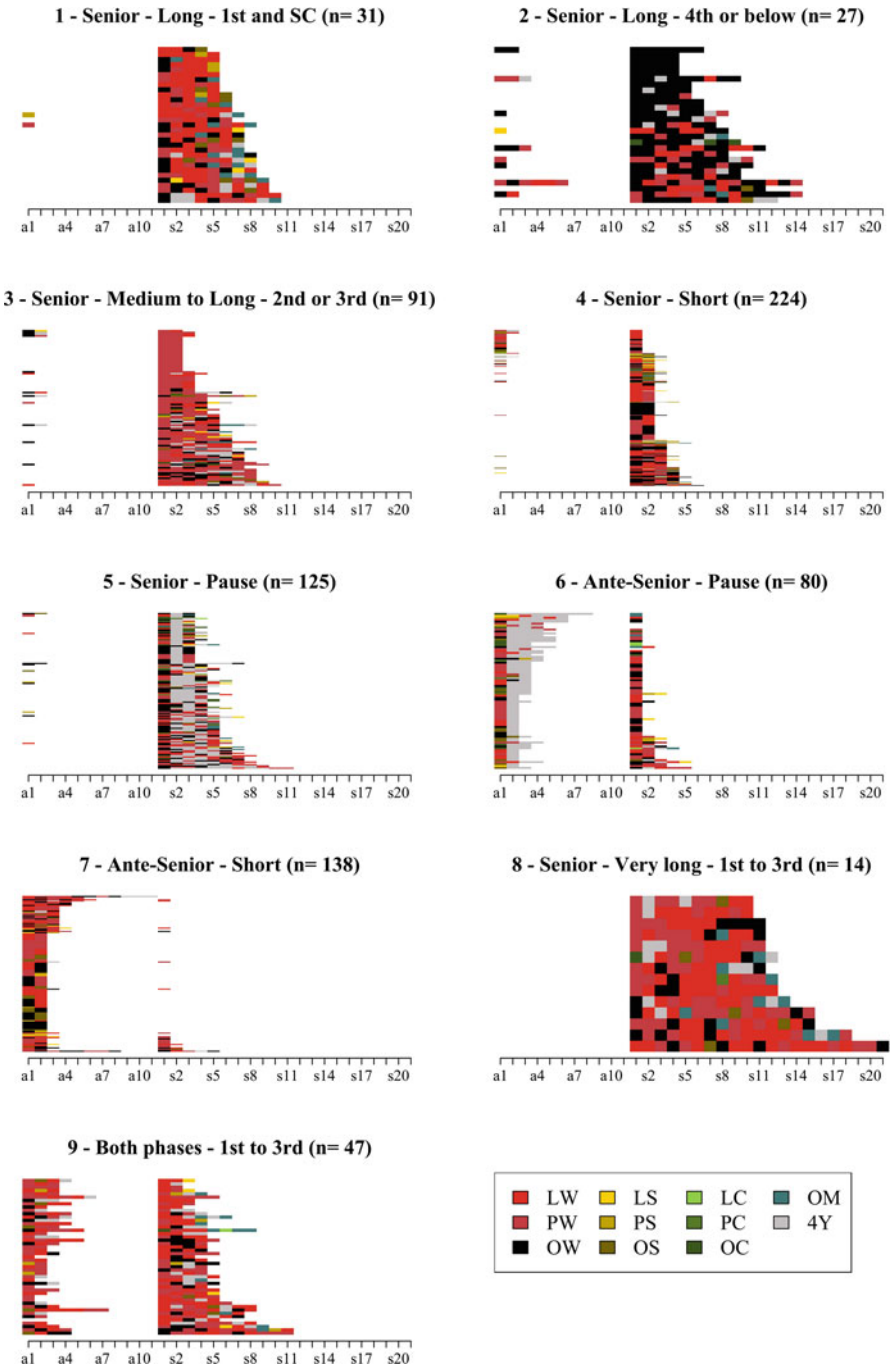

Fig. 2 MDS Sequence Index Plot for each cluster (sorted according to the first factor of multidimensional scaling following Piccarreta and Lior 2010). SC = competitions preceded by screening contests. (See explanation of state labels in Table 1)

Table 4 Sets of representative sequences for each cluster

Cluster	*Ante*-Senior	Senior
1		LW/3-PW/1-LW/1
		OW/1-PW/1-LW/1-OS/1-LW/1
		OW/1-LW/3-PS/1
		PW/1-LW/2-OW/1-OM/1
		LW/1-OW/1-LW/1-OM/1-LW/1-OS/1
		LW/1-PW/1-LW/1-PW/1-OS/1-OM/1
		OW/1-LW/1-OW/1-PW/1-OS/1-LW/1-OS/1
2		OW/5
3		PW/3
4		OW/1-LW/1
5		OW/1-4Y/2-PW/1
6	PW/1-4Y/2	PW/1
7	OW/1-PW/1	
8		PW/1-4Y/1-LW/3-PW/1-LW/2-PW/1-LW/1
		PW/1-LW/1-PW/3-LW/1-OW/1-OM/1-LW/2-OW/1
		PW/1-4Y/2-PW/3-LW/3-PW/2
		PW/2-LW/1-OW/1-LW/1-LW/1-LW/1-PC/1-LW/1-PW/2-OM/1
		PW/1-LW/1-OW/1-PW/2-LW/2-OM/1-4Y/2-OW/1
		OW/1-4Y/1-LW/1-PW/1-LW/1-PW/2-LW/2-PW/1-OM/1-4Y/1-OM/1
9	PW/1	LW/1-PW/1-OW/1-LW/1
	PW/1-LW/	LW/1-4Y/1
	1-PW/1	
	LW/2	PW/1-LW/1-PW/1-LW/1-LS/1
	PW/3	PW/2-LW/1-PW/1-LW/1

N.B.: In each cluster, the distance of at least one sequence out of two from one of the representative sequences is inferior to 30% of the maximum distance in the cluster (for details on the centrality criterion, see Gabadinho and Ritschard 2013). Sequences are sorted by representativeness. Each state is followed by its number of successive occurrences

A comparison between several clustering methods invited me to opt for a nine-cluster Ward's (1963) partition (see Fig. 2 for sequence index plots and Table 4 for representative sequences).[11]

[11] Any distance-based clustering method could be used including the property-based and fuzzy methods addressed by Studer (2018) in this bundle. Here, using the R package WeightedCluster (Studer 2013), several algorithms have been compared for a division into two to ten classes: hierarchical cluster solutions named Ward, single, complete, average (UPGMA), McQuitty (WPGMA) and beta-flexible (flexible-UPGMA) (for a presentation, see Müllner 2013; Belbin et al. 1992) and non-hierarchical partition around medoids algorithm (PAM) (Rousseeuw and Kaufman 1990). PBC, HC, HG and ASW measures have been used to compare the quality of the different clustering solutions (Hennig and Liao 2010). Ward's nine-cluster solution was the most relevant regarding both quality measures and readability. Except for PAM, other algorithms tend to produce a partition between a very heterogenous group containing more than 80% of the cases and two to

Three key elements of interpretation arise: participation in senior competitions, length and tonality (the most frequent state or family of states within the sequence).

Cluster 1 (4% of sequences) gathers sequences mainly characterised by an empty *ante*-senior phase (only two sequences out of 31 do not start with a senior competition) and a long senior phase (mean length is 6.87 time slots) including mainly first ranks and participations in competitions preceded by screening contests (64.7% of participations). Cluster 2 (3.5% of sequences) gathers medium to long senior careers (mean length is 8.63) in which first ranks are rare (less than 11% of participations). Cluster 3 (11.7% of sequences) gathers short to medium length senior careers (mean length is 4.99), mainly characterised by second and third ranks (70% of participations). Cluster 4 (28.8% of sequences) gathers short length senior sequences with no shared tonality. Clusters 5 (16.1% of sequences) and 6 (10.3% of sequences) are defined by at least one four-year pause, respectively during the senior phase and during the *ante*-senior phase. Cluster 7 (17.8% of sequences) gathers sequences defined by a short *ante*-senior phase and an empty senior phase with no shared tonality. Cluster 8 (1.8% of sequences) is made up of very long senior careers (mean length is 13) including mainly first to third ranks (76% of participations). Sequences in Cluster 9 (6% of sequences) share a symmetrical intensity regarding participations in *ante*-senior and senior phases and a relatively low rate of 4th ranks or below.

How far does this clustering take phases into account? First, the Ward two-cluster solution separates Clusters 6 and 7 from the seven other clusters, that is to say careers first defined by *ante*-senior participations from careers first defined by senior participations. Second, Clusters 6 and 7, both characterised by *ante*-senior participations, are distinct from one another with respect to participation in senior competitions. Third, a quarter of the sequences counting one or more ante-senior participations are not clustered in Clusters 6 and 7. In other words, closeness does not only rest on the (non-)emptiness of phases, but also on phases' tonality. Fourth, when, as here, the reference frame of the division is endogenous, that division greatly simplifies the interpretation: once the phases mainly portrayed by each cluster have been identified, interpretation is primarily based on ranking.

5.3 MPOM Compared

MPOM takes cues from two other families of dissimilarity measures that assume a division of sequences into dissociated and incommensurable episodes.

First, MPOM generalises Hamming Distance (Hamming 1950) and Dynamic Hamming Distance (DHD) (see Lesnard 2008, 2014), which measure dissimilarities

nine easy to analyse but very small groups. Quality measures are higher for Ward compared to PAM. A nine-cluster solution is associated with the highest value of HG and HC indexes, the second highest value of PBC and the fourth highest value of ASW.

position-wise. DHD can be approached as a specification of MPOM-distance in which each time slot is a phase (of length 1) and in which substitution costs are derived from transition rates before and after each time slot. This specification is suited for sequences defined by a limited number of states, observed in each sequence and spanning long spells.[12]

Second, compared with Qualitative Harmonic Analysis (QHA) (Deville and Saporta 1983; Robette and Thibault 2008; Robette and Bry 2012), in which sequences are divided into periods that are modelled as bundles of states varying from one another in proportions of time spent, MPOM pays attention to the order of the states within phases within phases.

Thus, the main advantage of MPOM is to take into account a unit nested in a sequence that is more malleable than time slots—its length varies from 0 to the whole sequence's length—and that is studied as a time-ordered structure.

6 Conclusion

This chapter has introduced the idea of multiphase sequences and several tools to study them, especially sliced representations and a multiphase dissimilarity measure (called MPOM) the logic of which can be extended to other dissimilarity measures. Two general issues have been raised and invite further investigation in the development of MPSA.

At the beginning of this chapter, sequences were defined as hierarchical structures, as narratives including sub-narratives and nested into larger narratives. This definition is partly consistent with Dumont's (1980) perspective of hierarchies as nested entities. Any sequence (marital biographies, job careers, dances, etc.) can be approached as a fragment of larger social processes (Abbott 2016) including other fragments of social processes. The approach to sequences as hierarchical structures could be further developed by investigating the variety of relations between nested temporal structures. Hybridisations of network analysis and sequence analysis (Cornwell 2015) may be a possible way to study these relations as a multilevel issue (Lazega and Snijders 2016).[13]

[12]To preserve a division into phases in the comparison of workweeks, Lesnard and Kan (2011) wrap each phase into an atomic time slot, thus describing each phase by a single state. Their two-step method identifies types of narratives through DHD and clustering procedure and then assembles these types in week-sequences analysed with DHD and clustering. Compared to MPOM, this wrapping solution is suited for periods of identical time spread (such as hours or days). Its main limitation is that its second step is based on the heterogenous inputs of a clustering procedure.

[13]Regarding MPOM, the analysis of the articulation of distances between phases and distances between sequences could be further developed since two sequences can be identical along some phases and clearly different along others. A related question is the importance of specific phases in the definition of a whole sequence. Various theoretical perspectives assume that certain phases are more crucial than others (childhood in a whole life-course for example). Such assumptions can orientate the parameters of MPOM by differentiating phases' weights.

A methodological assumption of MPSA is to approach phases as sites of narratives nested within sequences, which are sites of larger narratives. This assumption differentiates phases and the narrative patterns that the division into phases makes it possible to unveil. That echoes other notions centred on temporal substructures, such as subsequence (Elzinga et al. 2008) or motif (Han 2014). There is an open field for research on sub-narratives, their typical position(s) within sequences and their overlaps.[14] MPOM assumes a division into phases prior to the identification of narrative patterns. Different alphabets are defined, thus determining what patterns can be identified. A comprehensive method would manage three different steps: identifying types of narratives, identifying phases, identifying patterns of relation between phases and narratives. In other words, the division into phases could be dynamically revised and preceded by a moment of identification of patterns for different fragments of sequences under study.[15]

These elements invite renewed exploration of the various interrelations and continuities between temporal structures, a major question that sequence analysts have already extensively explored.

Acknowledgements I thank the participants to the LaCoSA 2 conference and the editors and anonymous reviewers of this book for valuable suggestions. I thank Claire Lemercier, Laurent Lesnard, Etienne Ollion and Loretta Platts from whom I received insightful comments on earlier versions of this chapter. I am thankful to Richard Nice for a meticulous proofreading of my English.

References

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21*(1), 93–113.

Abbott, A. (2001). *Time matters: On theory and method*. Chicago: University of Chicago Press.

Abbott, A. (2016). *Processual sociology*. Chicago: University of Chicago Press.

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History, 16*(3), 471–494.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research, 29*(1), 3–33.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge: Cambridge university press.

Belbin, L., Faith, D. P., & Milligan, G. W. (1992). A comparison of two approaches to beta-flexible clustering. *Multivariate Behavioral Research, 27*(3), 417–433.

[14]Regarding overlaps, my purpose was focused on clear-cut phases, divided from the single viewpoint of contractual events (a divorce, a recruitment, etc.), which is simplistic. For instance, the idea that a divorce begins with a contract in couples' histories is debatable. In most cases, the identification of phases as relevant locations of narratives is a challenging issue since many narratives overlap. Nevertheless, the notion of overlap is premised upon the idea of typical locations of narratives. Identifying those typical locations is thus a preliminary step to the study of overlaps.

[15]This identification could rely on tools for detecting consistency within segments, such as the stationarity test proposed by Bakeman and Gottman (1997).

Blanchard, P. (2010). Analyse séquentielle et carrières militantes. Research report, Institut d'études politiques et internationales.

Collas, T. (2015, Unpublished). *La pâte et le décor: Considération et formes professionnelles dans le monde des pâtissiers*. Ph.D. thesis, Sciences Po Paris.

Colombi, D., & Paye, S. (2014). Synchronising sequences: An analytic approach to explore relationships between events and temporal patterns. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 249–264). Cham: Springer.

Cornwell, B. (2015). *Social sequence analysis: Methods and applications*. Cambridge: Cambridge University Press.

Deville, J.-C., & Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series. *Journal of Econometrics, 22*(1–2), 169–189.

Dumont, L. (1980). *Homo hierarchicus: The caste system and its implications*. Chicago: University of Chicago Press.

Elzinga, C., Rahmann, S., & Wang, H. (2008). Algorithms for subsequence combinatorics. *Theoretical Computer Science, 409*(3), 394–404.

Gabadinho, A., & Ritschard, G. (2013). Searching for typical life trajectories applied to childbirth histories. In R. Levy & E. Widmer (Eds.), *Gendered life courses-Between individualization and standardization. A European approach applied to Switzerland* (pp. 287–312). Vienna: LIT.

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.

Gauthier, J.-A., Widmer, E., 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.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Labs Technical Journal, 29*(2), 147–160.

Han, S.-K. (2014). Motif of sequence, motif in sequence. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 21–38). Cham: Springer.

Hennig, C., & Liao, T. F. (2010). Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. Research Report 308, Department of Statistical Science, Department of Sociology, University of Illinois.

Lazega, E., & Snijders, T. (Eds.) (2016). *Multilevel network analysis for the social sciences*. New York: Springer.

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. (2014). Using optimal matching analysis in sociology: Cost setting and sociology of time. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 39–50). Cham: Springer.

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.

Levy, R., & The Pavie Team. (2005). Why look at life courses in an interdisciplinary perspective. In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, & E. Widmer (Eds.), *Towards an interdisciplinary perspective on the life course* (pp. 3–32). Amsterdam-Boston: Elsevier.

Müllner, D. (2013). Fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software, 53*(9), 1–18.

Piccarreta, R., & Lior, O. (2010). Exploring sequences: A graphical tool based on multidimensional scaling. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(1), 165–184.

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.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Robette, N., & Bry, X. (2012). Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 116*(1), 5–24.

Robette, N., & Thibault, N. (2008). L'analyse exploratoire de trajectoires professionnelles: Analyse harmonique qualitative ou appariement optimal? *Population, 63*(4), 621–646.

Rousseeuw, P. J., & Kaufman, L. (1990). *Finding groups in data*. New York: Wiley.

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers* (24).

Studer, M. (2018). Divisive property-based and fuzzy clustering for sequence analysis. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179*(2), 481–511.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Unpacking Configurational Dynamics: Sequence Analysis and Qualitative Comparative Analysis as a Mixed-Method Design

Camilla Borgna and Emanuela Struffolino

1 Introduction

Sequence analysis (SA) was introduced in the social sciences in the 1980s: Andrew Abbott, inspired by the treatment of DNA strings in biology, developed it as a technique to study social processes that unfold over time as sequences of events (Abbott and Forrest 1986). In contrast to the variable-oriented approach of techniques like event-history analysis, which focus on the timing of specific transitions, SA's approach is holistic and case-oriented (Billari 2005). This implies looking at the whole picture of individuals' longitudinal realizations rather than at the effects of variables on single-event outcomes. Against this scenario, processes of change over the life course are regarded as successions of states/actions located within constraining or enabling structures. This reflects (Abbott 1992)'s idea that holistic and analytical approaches can be combined into a multicase narrative methodology, based on which researchers can compare the unfolding of processes to pursue categorization and generalization.

SA has been extensively applied in demography and life-course research to study changes and continuities in individual pathways over time (e.g. Aassve et al. 2007;

The original version of this chapter was revised. A correction to this chapter is available at https://doi.org/10.1007/978-3-319-95420-2_17

C. Borgna (✉)
Collegio Carlo Alberto, Turin, Italy
e-mail: camilla.borgna@carloalberto.org

E. Struffolino (✉)
WZB Berlin Social Science Center – Research Group 'Demography and Inequality' and Humboldt University of Berlin, Berlin, Germany
e-mail: emanuela.struffolino@wzb.eu

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_10

Brzinsky-Fay and Solga 2016; Fasang and Raab 2014; Raitano and Struffolino 2013). Other fields of application include social policy analysis (Frank et al. 2000), democratization research (Wilson 2014), electoral participation studies (Buton et al. 2014), historical sociology (Abbott and Forrest 1986; Mercklé and Zalc 2014), and developmental psychology (Dietrich et al. 2014). The mainstream application of SA relies on optimal matching techniques to measure the distance between sequences that represent the individual realization of a certain process and on the clustering of such sequences to identify typical trajectories. Regression analysis is often applied as a second step to investigate why (and not only how) sequences resemble or differ from each other (Aisenbrey and Fasang 2010). Yet, multivariate regression analysis, with its focus on the net effects of single variables, seems to be at odds with the case-oriented nature of SA. In this chapter, we propose to use qualitative comparative analysis (QCA) as a second, explanatory step for SA. Compared to regression analysis, QCA is more coherent with SA's epistemological framework, because it shares the notion of analytically approaching social phenomena without disregarding their complexity. QCA was introduced in the 1980s (Ragin 1987) and relies on methodological tools that were uncommon in social science until then: in this case, logic and in particular Boolean and fuzzy-set algebra. Its developer Charles Ragin, like Abbott, aimed at bridging the divide between variable-oriented and case-oriented research. However, the joint holistic and analytical perspective of QCA does not concern over-time trajectories, but rather configurations of factors at given time points.

We contribute to the growing literature on mixed-methods research (e.g. Creswell 2009; Teddlie and Tashakkori 2006) by putting forward a novel "sequential mixed-method design" (Hollstein 2014; Teddlie and Tashakkori 2006) that consists of applying a recent innovation in SA—the so-called discrepancy analysis of state sequences (Studer et al. 2011; Struffolino et al. 2016)—as a first step, and crisp-set (Ragin 1987) or fuzzy-set (Ragin 2000, 2008) QCA as a second step. Our proposed framework allows researchers to analyze individual trajectories as a whole and to identify combinations of factors that are systematically linked to variations in the unfolding of such trajectories at potentially critical turning points. This can be read as an answer to the call for a "processual sociology" (Abbott 2016) that can take on the challenge of investigating both the dynamics of social phenomena as lineages of successive events and the complexity of contextual characteristics of "present" moments.

2 Sequence Analysis and Qualitative Comparative as a Sequential Mixed-Methods Design

Qualitative comparative analysis (QCA) is a method that permits systematic comparisons of cases through the highly-formalized tools of logic and set theory. Its perspective on the study of social phenomena is inherently analytical, but at the same time holistic and case-oriented (Berg-Schlosser et al. 2009; Ragin 1987; Schneider and Wagemann 2012).

In the QCA framework, cases are understood as complex entities of interrelated attributes. Empirically, this means that cases are classified as members or non-members of multiple and possibly overlapping sets. For instance, when investigating countries as cases, researchers can label them as "democracies" but also as "industrial economies," "corporatist welfare states" and so on. Each of these labels represents a set to which the country case can either belong or not. Similarly, if cases are represented by individuals rather than countries, researchers could envisage sets such as "working population," "women," "mothers," etc. While belonging to one set always implies not belonging to its negation, different sets are not mutually exclusive, as all mothers are women (subset relation) and some women are part of the working population (set intersection), for instance. In its original formulation—later known as crisp-set QCA—set membership was dichotomous, while the later development of fuzzy-set QCA (Ragin 2000, 2008) permits a more fine-grained assessment: Set membership can vary in a continuum from 0 to 1 (as, for instance, individuals who work few hours a week could be considered as partial members of the set "working population").

QCA can be used as a classification tool, for example for typology building (Berg-Schlosser et al. 2009). However, most of the applications so far have taken an explanatory perspective in assessing the empirical regularities that exist between some factors and an outcome (Marx et al. 2014). Typical fields of application include comparative politics (Schneider 2009), welfare studies (Emmenegger 2011; Vis 2009), policy and administration (Sager and Thomann 2016), sociology of work and education (Borgna 2016; Glaesser and Cooper 2010), and organization research (Fiss 2011). The technique consists of a first step dedicated to identifying (combinations of) conditions in presence of which the cases systematically display a given outcome (a procedure known as truth-table construction). In a second step, redundant elements are removed from these configurations by applying Boolean or fuzzy-set algebra (i.e. truth-table minimization). The resulting Boolean expression represents the logically-minimal combinations of factors that are sufficient for the occurrence of the outcome and are sometimes defined as "outcome-enabling context" (Schneider and Wagemann 2012).

In a nutshell, QCA is a powerful method for systematically comparing cases without ruling out their potentially configurational nature. This complements the approach of SA, which underlines the complexity of cases in terms of sequencing of states and duration of events.

In this contribution, we bring together the strengths of SA and QCA in a sequential mixed-methods design. We propose applying SA in a first research stage, in order to describe both qualitatively and quantitatively the temporal complexity of a given social process. By applying QCA in a second stage, it is possible to shed light on the configurational complexity at given phases along the process and to reduce such complexity to synthetic combinations of explanatory factors. This combination is not a simple juxtaposition of methods but rather a genuine example of cross-fertilization between two methodological traditions that arose from the same desire to "bring cases back" in quantitative analysis (Ragin and Becker 1992). By adopting our proposed research design, researchers can approach complexity from different

angles and thus explore its temporal and configurational dimensions within the same analytical framework. This exploration is a first step towards the identification of what (Abbott 1992, 2009) has called "turning points" along the unfolding of social processes.

Practically, the steps of our proposed framework are the following: (1) sequence construction, where each time point is designated as a categorical state identifying the outcome of interest; (2) discrepancy analysis of state sequences, where we identify the phases when a given factor is mostly or increasingly relevant to explain inter-sequence differences; (3) truth-table construction, where we assess for each of these phases which combinations of factors are systematically associated with the outcome; (4) truth-table analysis, where the logically-sufficient factors for the occurrence of the outcome are identified.

3 Empirical Illustration

To illustrate the added value of our approach, we apply it to the study of women's employment trajectories in divided Germany (1955–1990). The leading research question for this exercise is: what conditions enabled women in East and West Germany to continue their education or employment over early- and mid-adulthood? This question is suitable for our illustrative purposes because there is extensive literature that accounts for the factors shaping women's labor-market participation within the two contexts (Diewald et al. 2006; Rosenfeld et al. 2004). This enables us to evaluate our results against a wide range of established, substantive evidence.

3.1 Background

During the period of division (1955–1990), the German Democratic Republic (henceforth, East Germany) and the Federal Republic of Germany (henceforth, West Germany) differed to a great extent concerning their economic and welfare systems.

The social market economy in West Germany was coupled with the promotion of a pro-traditional male breadwinner model in a corporative welfare state (Engelhardt et al. 2002; Rosenfeld et al. 2004). A highly gendered division of labor was indirectly supported by social policies that did not promote the compatibility of work and family and by a taxation system that penalized working wives (Brückner 2004; Cooke 2011; Sainsbury 1999). Female labor-market participation was around 50%—mostly part-time—and it was discouraged especially for women with children, while for male employment rates were around 80% (Diewald et al. 2006).

In the egalitarian centrally planned economy of East Germany, women's employment was supported by an infrastructure and an ideology that affirmed the right and duty to work. Pro-natalist family policies aimed at improving compatibility of work and family: The normative pressure to have children in one's early twenties was

combined with widespread childcare, and state-controlled resources (e.g. housing or loans) were available to those who started a family (Kreyenfeld 2004). As a result, women's employment rates paralleled men's, reaching 90% also for women with children (Huinink et al. 1995).

When considering enabling conditions for employment, we expect gender to play a major role in the explanation of inter-individual differences along the whole life course, and especially so when specific life transitions occurred. Moreover, given the institutional differences outlined above, we expect gender to be a prominent explanatory factor in the West more than in the East. Finally, within each context, we expect some combinations of factors to be consequential for women's employment. We focus on three drivers that the literature indicates to be relevant to female labor participation (Engelhardt et al. 2002; Kreyenfeld 2004; Rosenfeld et al. 2004): partnership status, number of children, and parental education.

To develop our empirical illustration, we analyze longitudinal-retrospective data from the Starting Cohort Six of the *National Educational Panel Study* (NEPS) (Blossfeld et al. 2011). We selected a subsample of individuals born in East and West Germany (N = 374 and N = 1,695 respectively) between 1944 and 1955, i.e. individuals who experienced most of their early- and mid-adulthood in divided Germany.

3.2 Empirical Analysis

In a first step, we construct individual sequences representing longitudinal employment trajectories of men and women in West and East Germany. We then apply discrepancy analysis of state sequences to estimate the proportion of the variation in employment trajectories explained by gender at each time point in the sequences. This allows us to identify the phases when gender is mostly or increasingly relevant.

In a second step, we focus on these time points and apply fuzzy-set QCA to the sample of women to identify the configurations of factors sufficient for them to be employed or pursuing education in West and East Germany.

3.2.1 Step 1: Sequence Analysis

Sequences' Construction

We construct individual employment trajectories from ages 15 to 40 as sequences of monthly intervals[1]: each of the 300 months is encoded according to a binary definition that distinguishes being in education or employment from being out of education and out of employment (unemployment, inactivity, military service, maternity leave). We adopt this binary definition because both crisp- and fuzzy-set QCA require a two-dimensional outcome. This does not rule out the possibility

[1]All analyses are performed in R.3.2.5 by using the packages TraMineR, TraMineRextras (Gabadinho et al. 2011), and WeightedCluster (Studer 2013). For a discussion on the choice of the distance measure see Studer and Ritschard (2016).

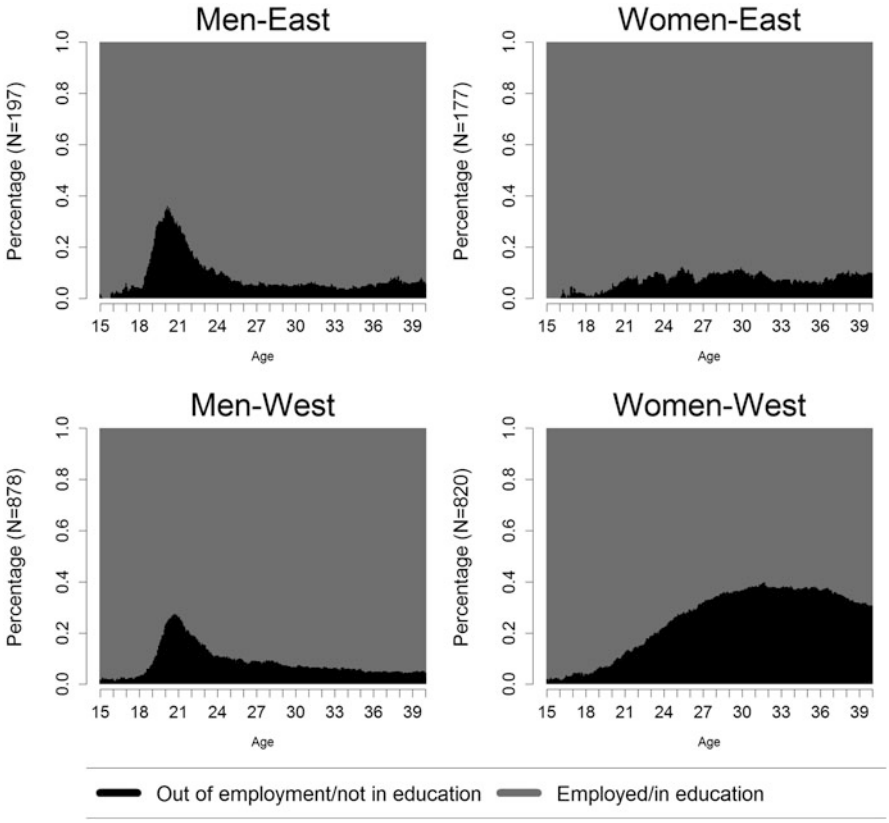

Fig. 1 State distribution plot by gender and context: Percentage of individuals in a certain state at each point in time. (Source: authors' elaboration on NEPS data)

of adopting more elaborate definitions of states, as SA could be combined with multi-value QCA (Cronqvist and Berg-Schlosser 2009). From a substantive point of view, being in employment or in education is often considered a "positive" outcome during early adulthood and one that is associated with greater chances of societal integration later on. We censor our observational window at age 40 because we assume that the most important steps for the establishment of one's career have already occurred during this phase of the life course.

Figure 1 shows the state distribution plot by gender in East and West Germany. For men, periods spent out of employment or education were concentrated in the years around military service in both contexts. In contrast, women's employment trajectories differed according to context: while women in the East exhibited very limited amounts of time spent out of employment over the whole timespan considered, in West Germany women's likelihood of being in employment or education decreases over time.

Discrepancy Analysis

We apply discrepancy analysis of state sequences (DA) (Studer et al. 2011), which translates the ANOVA framework into sequence analysis. The resulting pseudo-R^2 values can be interpreted as the share of the total variability between sequences explained by a specific factor of interest at each point in time (for a detailed description see Struffolino et al. 2016). The input of DA is a pairwise distance matrix based on a dissimilarity measure computed by assigning a cost to each transformation needed to align the couples of sequences (insertion—deletion and substitution costs). We used the optimal matching dissimilarity measure with an insertion—deletion cost of 1 and a substitution cost of 2. Notice that in this case, the sequences are composed of only two states, and therefore any more complex specification of costs would have been pointless (Studer and Ritschard 2016). Our explanatory factor of interest is gender. Hence, the pseudo-R^2 value for between-group differences at each time point can be interpreted as the share of the total variability in the pool of sequences that is explained by gender.

Figure 2 displays the pseudo-R^2 values for gender in East and West Germany. In the East, gender explains 12% of the variation between employment trajectories around age 20: this is most probably due to men being in military service. After age 23, gender explains very little of the variation between trajectories, because employment rates were very high for both men and women. In contrast to this, in the West gender explains an increasing share of the variability between employment trajectories starting from age 22; it reaches 15% between age 30 and 36.

The peaks in the pseudo-R^2 represent phases of the life course where gender differences reach their relative maximum: These phases are of particular substantive interest because they might represent 'turning points' along the unfolding of employment trajectories (Abbott 2009). However, within the SA framework, change is understood as lasting over a period of time, rather than as happening at specific time points (Shanahan 2000). Divergence between sequences may then result from an extended succession of critical moments and not necessarily concentrate at single peaks. Therefore, both the peaks and the phases where gender differences display sharp increase can be considered as potential turning points. We randomly select one specific time point (month) for phases of peaks and sharp increases. To be able to compare between the two contexts, we additionally selected time points in the West and applied them to the East, and vice versa. This results in five time points that constitute the starting point for our second step of analysis: 19 years, 20 years and 9 months, 23 years, 31 years, and 35 years.

3.2.2 Step 2: Qualitative Comparative Analysis

In the second research stage, we shift the focus to women, because it is their patterns that diverge from those of men, at least in the West, and investigate which configurations of factors are systematically associated with being in employment or in education at each of the five time points identified in the first research stage and within the two contexts of East and West Germany. Our explanatory conditions refer

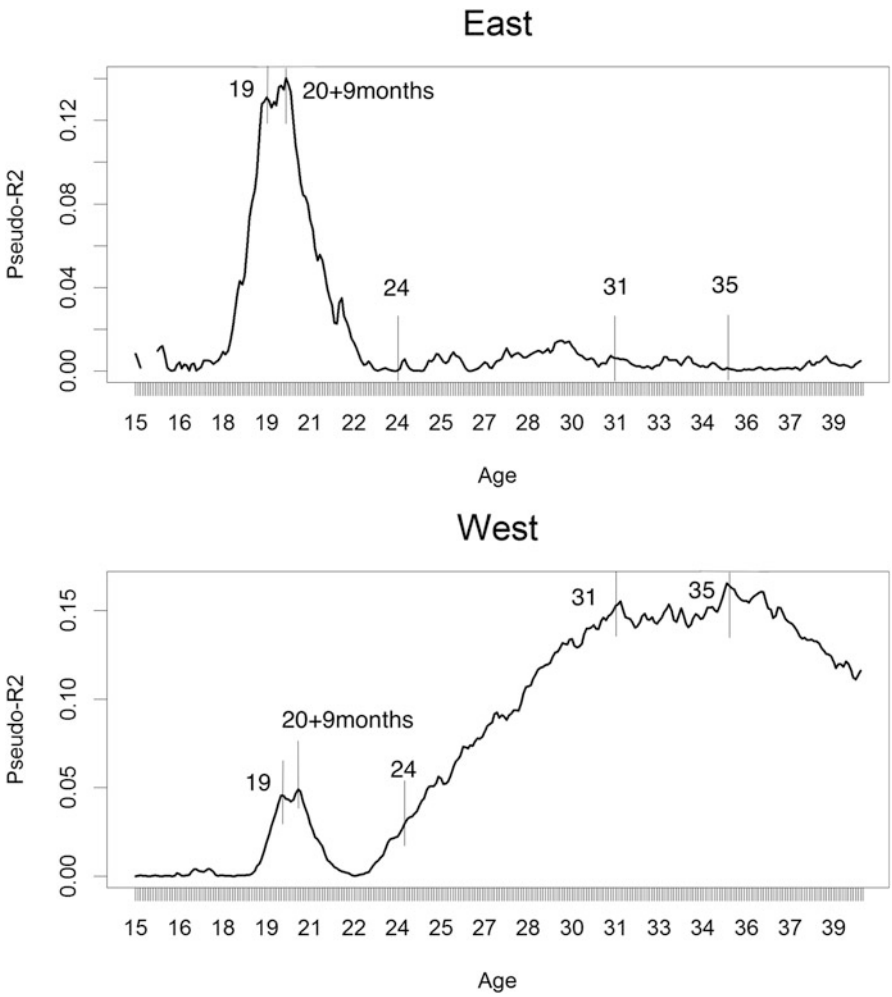

Fig. 2 Share of the variability in employment trajectories over time explained by gender in East and West Germany: Pseudo-R^2 values. (Source: authors' elaboration on NEPS data)

to women's partnership status, number of children, and family background. While many other factors could be listed as potentially explanatory, for the sake of simplicity of this empirical exercise we focus on the three identified as the most relevant according to previous literature (see Sect. 3.1). Moreover, parsimony in identifying potentially explanatory conditions is required when applying QCA, as the number of possible combinations grows exponentially with the inclusion of new conditions.

Calibration

As mentioned in Sect. 2, to apply QCA, the cases under observation must be classified as members or non-members of the sets corresponding to the outcome

and the explanatory conditions. While in principle researchers can attribute set membership based on their own substantive knowledge of the cases, set membership is often assigned following the more systematic procedure of calibration, which consists of applying external criteria to transform an interval-scale variable into a crisp or fuzzy set.[2]

Two calibration techniques can be distinguished: Following the direct method (Ragin 2008, 94–97), researchers need to specify three values of the interval scale (also known as "source variable") that will identify three qualitative breakpoints in the set membership: full inclusion, full exclusion, and maximum ambiguity or crossover. For fuzzy sets, full inclusion usually corresponds to a set membership of 0.99, full exclusion to 0.01, and the crossover to 0.5, while intermediate values are assigned according to a logistic function. For crisp sets, only the value corresponding to the crossover must be specified, as set membership can only assume the value of 0 or 1. Instead, when following the indirect method of calibration, one should use substantive knowledge to determine the correspondence between values of the source variable and set membership scores. The recommended set membership scores are: 1 for full inclusion, 0.8 for cases "mostly but not fully in the target set," 0.6 for cases "more in than out of the target set," 0.4 for cases "more out than in of the target set," 0.2 for cases "mostly but not fully out of the target set," and 0 for full exclusion (Ragin 2008, 95–96). For both the direct and indirect calibration methods, the threshold criteria need to be justifiable and transparent. In our empirical application, cases are constituted by individuals—or more specifically, women—in East and West Germany observed at each of the five time points selected in the SA analytical step. The criteria for the calibration of the sets corresponding to the outcome and to the explanatory conditions are detailed in the following.

Outcome: Crisp set of "women in employment *or* in education". In this case no calibration is necessary because the source variable, corresponding to the two states analyzed in the first research stage, is already dichotomous (see above).

Explanatory Conditions:

- PARTNERSHIP: fuzzy set of "women in a stable partnership"; source variable: self-reported marital status; method of calibration: indirect; threshold criteria and set memberships: single: 0; cohabiting but unmarried: 0.6; married: 1.
- CHILDREN: crisp set of "women with children"; source variable: number of children; method of calibration: direct; threshold criteria and set memberships: no children: 0 for one or more children: 1.
- HIGH-PAREDU: fuzzy set of "women from relatively advantaged social background"[3]; source variable: number of years corresponding to the highest educational qualification attained by either parent; method of calibration: direct; threshold criteria and set memberships: these are specific to the two contexts

[2]Remember that for crisp sets membership is dichotomous, while for fuzzy sets it can vary in a continuum between 0 and 1.

[3]For the sake of simplicity, in what follows we shall talk of "highly educated parents" vs. "non-highly educated parents". However, cases belonging to this set should be understood as women originating from households where at least one parent has at least an intermediate level of education—hence, a proxy for a relatively advantaged social background.

Table 1 Descriptive statistics for the outcome and the conditions, by context and time point

	Age, EAST Germany					Age, WEST Germany				
Variable	19	20+ 9m	24	31	35	19	20+9m	24	31	35
Employment status										
Out of employment/ not in education	1.5	6.6	9.4	9.4	6.8	6.3	10.6	22.3	39.8	37.7
Employed/in education	98.5	93.4	90.6	90.6	93.2	93.7	89.4	77.7	60.2	62.3
Partnership status										
Single	90.9	55.5	17.1	7.5	8	87.3	63.3	25.8	7.4	8.7
Cohabiting	2.4	9.2	6.4	4.2	4.6	2.1	6.5	10	8.5	5.4
Married	6.8	35.4	76.5	88.3	87.4	10.6	30.2	64.3	84.1	85.9
Number of children										
0	93.2	70.2	26	6.9	3.8	92.9	82.5	61.5	23.2	18.7
1	6.3	29	57.8	34.1	33.1	6.6	15.2	27.6	30.9	28
2	0.5	0.3	15.4	49.4	51.4	0.6	2.1	9.7	37.1	40.2
3	0	0.5	0.3	8.5	10.4	0	0.3	1.2	6.6	9.4
4+	0	0	0.5	1.1	1.4	0	0	0.1	2.2	3.7
Parental education (years)										
8	0.2					0.8				
9	1.8					9.8				
10	0					0.3				
12	73.7					64.8				
13	6.6					11.8				
15	3.6					4.8				
16	2.7					1.7				
18	11.3					6.2				
N	177	177	177	177	177	820	820	820	820	820

Source: authors' elaboration on NEPS data
m months

to account for the differences in the degrees of educational expansion and in the educational systems. East Germany: 15.5 years (university education): 0.99; 12.5 years (above the 12 threshold corresponding to upper-secondary education): 0.5; 11 years (less than upper-secondary education): 0.01. West Germany: 15.5 years (university education): 0.99; 11 years (above the 10 threshold corresponding to the intermediate school certificate): 0.5; 9.5 (no or low school certificate only): 0.01.

Table 1 reports descriptive information on the source variables for the outcome and the conditions for each time point and context.

Table 2 Truth tables for the presence of the outcome, by context and time points

EAST Germany						WEST Germany					
CHILD	HIGH-PAREDU	PARTNERSHIP	OUT	N	Cons.	CHILD	HIGH-PAREDU	PARTNERSHIP	OUT	N	Cons.
		age 19						*age 19*			
0	**0**	**0**	**1**	114	**1**	**0**	**1**	**0**	**1**	653	0.942
0	**1**	**0**	**1**	38	**1**	**0**	**0**	**0**	**1**	69	0.960
						0	**1**	**1**	**1**	53	0.904
						1	1	1	0	35	0.662
		age 20+9 months						*age 20+9 months*			
0	**0**	**0**	**1**	64	1	**0**	**1**	**0**	**1**	474	0.953
1	**0**	**1**	**1**	32	0.761	**0**	**1**	**1**	**1**	144	0.943
0	**0**	**1**	**1**	29	1	1	1	1	0	102	0.571
0	**1**	**0**	**1**	28	**1**	**0**	**0**	**0**	**1**	49	0.97
		age 24						*age 24*			
1	**0**	**1**	**1**	91	0.856	**0**	**1**	**1**	**1**	282	0.939
1	**1**	**1**	**1**	23	0.861	1	1	1	0	230	0.476
						0	**1**	**0**	**1**	206	0.973
						1	0	1	0	45	0.508
						0	**0**	**1**	**1**	24	0.935
		age 31						*age 31*			
1	**0**	**1**	**1**	118	0.901	1	1	1	0	517	0.515
1	**1**	**1**	**1**	37	0.868	**0**	**1**	**1**	**1**	157	0.930
						1	0	1	0	69	0.486
						0	**1**	**0**	**1**	36	0.867
						1	1	0	0	23	0.646
		age 35						*age 35*			
1	**0**	**1**	**1**	120	0.930	1	1	1	0	549	0.569
1	**1**	**1**	**1**	38	0.911	**0**	**1**	**1**	**1**	119	0.915
						1	0	1	0	71	0.537
						1	1	0	0	35	0.714
						0	**1**	**0**	**1**	30	0.886

Source: authors' elaboration on NEPS data. Logical remainders ($N < 20$) omitted for parsimony. Consistency threshold: 0.75. In bold: configurations associated with the presence of the outcome.

Truth-Table Construction

Before proceeding to the analysis, we explore the empirical configurations displayed by the cases and their relation with the outcome for each time point and context. To this scope, we construct a so-called "truth table," whose rows correspond to possible combinations of conditions. Based on the empirical distribution of cases, each combination is connected to either the presence or the absence of the outcome. The 10 resulting truth tables are displayed in Table 2.

The intersection of our three conditions (CHILDREN, HIGH-PAREDU, PARTNERSHIP) produces eight possible configurations. However, not all configurations display empirical instances and some rows are therefore empty (the so-called "logical remainders"). To minimize random noise, we set a frequency threshold of 20 cases, meaning that configurations with less than 20 cases are classified as logical remainders.[4] In Table 2, logical remainders are omitted for parsimony. Each row represents a combination of conditions (with 1 indicating the presence, and 0 the absence of the condition) to which a certain number of cases (N) belongs. For instance, the first row in the right part of the table indicates that, in the sample of West Germany, when they were 19 years old, 635 women did not have any children, came from highly educated families, and were not in a partnership (CHILDREN = 0, HIGH-PAREDU = 1, PARTNERSHIP = 0). Among them, the vast majority was in employment or in education (OUT = 1), as indicated by the consistency parameter (Cons. = 0.942). This parameter, roughly speaking, returns the proportion of women belonging to the 0-1-0 configuration who were in employment or in education.[5]

Table 2 reveals that in each of the five time points, for both the East and the West, multiple configurations were associated with the outcome. In East Germany, this applies to all existing configurations (i.e. displaying at least 20 empirical instances), in line with the fact that in this context the great majority of women remained in employment (or education) throughout their whole life course (see Fig. 1).

Looking more closely at Table 2, we note that, with the exception of the second time point, there are only two configurations for East Germany. Moreover, they are constant for the last three time points: women in stable partnerships and with children, with (1-1-1) and without (1-0-1) highly educated parents. This finding is consistent with previous research underlining the standardization of individual life courses in the GDR (Fasang and Raab 2014; Struffolino et al. 2016). More diversity existed in West Germany. This is visible first of all in the number of rows that exhibit empirical instances: Women are generally grouped in four or five configurations. Second, such configurations vary considerably over time, as can be seen from the comparison of the rows across time points. Finally, some configurations are associated with the presence and some with the absence of the outcome.

Truth-Table Analysis

In order to remove the redundant elements from the configurations associated with the outcome and thus identify the "outcome-enabling" conditions, we perform the fuzzy-set minimization of the 10 truth tables.[6] Different strategies exist in the literature for dealing with the issue of limited diversity, which emerges in QCA as non-existent configurations or logical remainders (Schneider and Wagemann 2012, 160–177). Given the descriptive purpose of our analysis, we refrain from

[4] Alternatively, one could set a relative threshold corresponding to a given share of cases, e.g. 5%.

[5] More precisely, the consistency parameter indicates the extent to which the fuzzy set corresponding to the 0-1-0 configuration is a subset of the fuzzy set corresponding to the outcome.

[6] The analyses were performed using the QCA package (Thiem and Duşa 2013) in R.3.2.5 using the Enhanced Quine-McCluskey Algorithm without row dominance rule.

using logical remainders, meaning that configurations with less than 20 cases are not considered in the minimization process. Therefore, our solutions are to be intended as conservative (*ivi*: 182). Additionally, before proceeding to the analysis one must set a consistency threshold in order to exclude configurations that are not sufficient subsets of the outcome. We apply a consistency threshold of 0.75, which is generally considered acceptable in the QCA literature (*ibidem*), but results are generally robust even with stricter consistency levels.[7]

The results are summarized in Table 3, whose columns represent the five time points and whose rows display the "prime implicants." The prime implicants are the essential components of a configuration identified by the truth-table analysis that, read in combination with each other, constitute the logically-minimal Boolean expression sufficient for the presence of the outcome (Ragin 1987, 95–98). For example, for 19-year-old West-German women, in order to be in employment or in education, it was sufficient not to have children *and* not to be in a partnership, *or*, as an alternative, not to have children *and* to have highly educated parents. The resulting logically-minimal Boolean expression is therefore: "child*partnership + child*HIGH-PAREDU".[8]

Table 3 thus shows the dynamic development of the combinations of factors logically sufficient for being in employment or in education. Indeed, these combinations changed over time during the phases of early- and mid-adulthood in both contexts. In both East and West Germany, not having children *and* not being in a partnership emerges a prime implicant for the first two time points considered, although in combination with different prime implicants (high-paredu*PARTNERSHIP in the East and child*HIGH-PAREDU in the West), but later disappears. As women reached the stage of their life course when most entered partnerships and had children, the

Table 3 Prime implicants for the presence of the outcome, by context and time point

	Age, EAST Germany					Age, WEST Germany				
Prime implicant	19	20 + 9 m	24	31	35	19	20 + 9 m	24	31	35
CHILDREN*PARTNERSHIP			•	•	•					
high-paredu*PARTNERSHIP		•								
children*high-paredu										
children*PARTNERSHIP								•		
children*partnership	•	•				•	•			
children*HIGH-PAREDU						•	•	•	•	•

Source: authors' elaboration on NEPS data. Conservative solutions. Consistency threshold: 0.75. Upper-case letters represent the presence and lower-case letters the absence of the condition
m = months

[7]Applying consistency thresholds between 0.76 and 0.85 produces results identical to those displayed in Table 3, with the exception of the second prime implicant, which in this case consists of "child*high-paredu*PARTNERSHIP."

[8]In Boolean language, upper-case letters represent the presence and lower-case letters the absence of the condition; the star sign represents the logical AND (set intersection) and the plus sign represents the logical OR (set union).

combinations of factors sufficient for remaining in employment or in education diverged across the two contexts: In East Germany, the only prime implicant from 24 years on is constituted by women in a partnership *and* with children. In West Germany, where the transition to fertility was generally delayed compared to the East (Kreyenfeld 2004), for 24-year-old women we still find the alternative combinations of not having children *and* having highly educated parents *or* or not having children *and* being in a partnership. However, from age 31, the only prime implicant left is that of women without children and with highly educated parents.

For both contexts, we observe a collapse over time into a single prime implicant. However, the implications are very different for East and West Germany: The prime implicant for East Germany (being in a partnership *and* with children) corresponds to the prevailing configuration of attributes for women aged 24 and older. Moreover, for these Eastern German women, the outcome is skewed to the right or, in other words, the overwhelming majority were in employment or in education (consistently with what shown by Huinink et al. 1995). Hence, the prime implicant and the outcome are in a relation of almost perfect set coincidence, which should not be interpreted as causal but rather as constitutive (Borgna 2013). In substantive terms, as noted earlier, this is driven by a strong standardization of individual life courses (Fasang and Raab 2014; Struffolino et al. 2016). In West Germany, in contrast, the lasting prime implicant corresponds to a limited portion of the population and can therefore properly be read as "outcome-enabling": Among women older than 24, only those without children *and* with highly educated parents could systematically still be found in employment or in education. The prime implicant of women in a partnership *and* without children disappears, because for ages 31 and 35 it corresponds to a group that is too small to be considered an existing configuration, because the great majority of women in a partnership eventually had children, and those who did not have any originate from families with highly educated parents (see also Table 2).

Overall, while we detected a greater heterogeneity in the life courses of women in West Germany from the mere observation of the truth tables, the truth-table analyses suggest that this greater diversity does not translate into a greater variety of pathways towards employment or education. While further research would be needed for any definitive, substantial conclusions, the results of this empirical exercise are not only in line with previous evidence on women's labor-market participation in East and West Germany but also provide some new insights into the different degree of complexity of their life courses.

4 Concluding Remarks

In this contribution, we have put forward a sequential mixed-methods design to unpack the temporal and configurational complexity of social phenomena. By applying SA, researchers can analyze the unfolding of phenomena over time in terms of sequencing of states and duration of events. More specifically, through

discrepancy analysis of state sequences, researchers can identify the crucial points in time when trajectories (start to) diverge from each other. In a second research stage, QCA can be applied to investigate whether, at such crucial time points, some particular combinations of factors explain why the cases display a given state (or outcome). By analyzing these cross-sectional turning points sequentially, we preserve the longitudinal and holistic perspective on trajectories.

We have illustrated the usefulness of our framework with an empirical analysis of employment trajectories in divided Germany. We have shown that, especially for West Germany, gender significantly explains the divergence of trajectories at what we characterized as turning points in women's life course. Moreover, while a plurality of factors initially explain women's presence in employment or education in both contexts, only one configuration of factors for each context is systematically linked to the outcome at the end of the period considered. Hence, by combining SA and QCA, we were able to unveil the dynamic of conditions sufficient for women to be in employment or education.

Notwithstanding the specificities of our empirical illustration, we believe that our proposed framework holds the promise of being useful to a wide range of researchers from a variety of disciplinary fields and who are interested in various kinds of micro-, meso-, and macro-level processes that involve a temporal and configurational dimension. We limit ourselves to three fields for which this SA-QCA design clearly appears to be valuable. First, recent developments in social stratification research have emphasized the role of cumulative advantage (DiPrete and Eirich 2006) and intersectionality (Platt 2011) but struggle to combine the two aspects in a single analytical framework. Second, a holistic perspective on both the context of implementation and the process of change is central in policy analysis and evaluation, especially for the approach of theory-based evaluation (Befani and Sager 2006). Third, the definition of "critical junctures" in the historical institutionalist tradition (Thelen 1999) appears to bear more than one similarity to our treatment of outcome-enabling configurations at turning points along the sequences. Similarly, Schneider and Rohlfing (2013) and more recently Williams and Gemperle (2017) have called for an integration of QCA and process-tracing methodologies to identify the "situational" causal mechanisms. Future research could explore the potential of combining our SA-QCA framework with more qualitative methods, like process tracing to the study of social processes and critical junctures in particular.

Acknowledgements We thank Anette E. Fasang for generously sharing the routines for data preparation and providing helpful comments and suggestions on an earlier version of this chapter. For insightful comments, we are also grateful to David Brady, Markus Siewert, Eva Thomann, as well as to the participants in the Colloquia *Quantico* at the ISW-Humboldt University of Berlin and *CO:STA* at the WZB Berlin Social Science Center. We thank Ana Santiago-Vela for excellent research assistance.

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.

Abbott, A. (1992). From causes to events: Notes on narrative positivism. *Sociological Methods & Research, 20*(4), 428–455.

Abbott, A. (2009). A propos du concept de Turning Point. In M. Bessin, C. Bidart, & M. Grossetti (Eds.), *Bifurcations: les sciences sociales face aux ruptures et à l'événement* (pp. 187–211). Paris: La Découverte.

Abbott, A. (2016). *Processual sociology*. Chicago: Chicago University Press.

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History, 16*(3), 471–494.

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.

Befani, B., & Sager, F. (2006). QCA as a tool for realistic evaluations the case of the swiss environmental impact assessment. In B. Rihoux & H. Grimm (Eds.), *Innovative comparative methods for policy analysis. Beyond the quantitative-qualitative divide* (pp. 263–284). Boston: Springer.

Berg-Schlosser, D., De Meur, G., Rihoux, B., & Ragin, C. (2009). Qualitative comparative analysis (QCA) as an approach. In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods* (pp. 1–19). London: SAGE.

Billari, F. C. (2005). Life course analysis: Two (complementary) cultures? Some reflections with examples from the analysis of the transition to adulthood. *Advances in Life Course Research, 10*, 261–281.

Blossfeld, H.-P., Rossbach, H.-G., von Maurice, J., Schneider, T., Kiesl, S. K., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., Prenzel, M. S., & others (2011). *Education as a lifelong process. The German national educational panel study (NEPS)*. Berlin: Springer.

Borgna, C. (2013). Fuzzy-set coincidence analysis: the hidden asymmetries. COMPASSS Working Papers Series 72.

Borgna, C. (2016). Multiple paths to inequality. How institutional contexts shape the educational opportunities of second generation immigrants in Europe. *European Societies, 18*(2), 180–199.

Brückner, H. (2004). *Gender inequality in the life course: Social change and stability in west Germany 1975–1995*. New York: De Gruyter.

Brzinsky-Fay, C., & Solga, H. (2016). Compressed, postponed, or disadvantaged? School-to-work-transition patterns and early occupational attainment in west Germany. *Research in Social Stratification and Mobility, 46*, 21–36.

Buton, F., Lemercier, C., & Mariot, N. (2014). A contextual analysis of electoral participation sequences. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 191–211). Berlin: Springer.

Cooke, L. P. (2011). *Gender-class equality in political economies*. New York: Routledge.

Creswell, J. W. (2009). Editorial: Mapping the field of mixed methods research. *Journal of Mixed Methods Research, 3*(2), 95–108.

Cronqvist, L., & Berg-Schlosser, D. (2009). Multi-value QCA (mvQCA). In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods* (pp. 145–166). London: SAGE.

Dietrich, J., Andersson, H. A., & Salmela-Aro, K. (2014). Developmental psychologists' perspective on pathways through school and beyond. In P. Blanchard, F. Bühlmann, & J. A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 129–150). Cham: Springer.

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.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology, 32*, 271–297.

Emmenegger, P. (2011). Job security regulations in western democracies: A fuzzy set analysis. *European Journal of Political Research, 50*(3), 336–364.

Engelhardt, H., Trappe, H., & Dronkers, J. (2002). Differences in family policies and the intergenerational transmission of divorce: A comparison between the former east and west Germany. *Demographic Research, 6*, 295–324.

Fasang, A. E., & Raab, M. (2014). Beyond transmission: Intergenerational patterns of family formation among middle-class American families. *Demography, 51*(5), 1703–1728.

Fiss, P. C. (2011). Building better causal theories: A fuzzy set approach to typologies in organization research. *Academy of Management Journal, 54*(2), 393–420.

Frank, D. J., Hironaka, A., & Schofer, E. (2000). The nation-state and the natural environment over the twentieth century. *American Sociological Review, 37*, 96–116.

Gabadinho, A., Ritschard, G., Müller, N., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 4*(40), 1–37.

Glaesser, J., & Cooper, B. (2010). Selectivity and flexibility in the German secondary school system: A configurational analysis of recent data from the German socio-economic panel. *European Sociological Review, 27*(5), 570–585.

Hollstein, B. (2014). Mixed methods social network research: An introduction. In S. Domínguez & B. Hollstein (Eds.), *Mixed Methods Social Networks Research Design and Applications* (pp. 3–34). New York: Cambridge University Press.

Huinink, J., Meyer, K. U., Solga, H., Sorensen, A., & Trappe, H. (1995). *Kollektiv und Eigensinn. Lebensläufe in der DDR und danach*. Berlin: Akademie-Verlag.

Kreyenfeld, M. (2004). Fertility decisions in the FRG and GDR: An analysis with data from the German fertility and family survey. *Demographic Research, 3*(11), 276–318.

Marx, A., Rihoux, B., & Ragin, C. (2014). The origins, development, and application of qualitative comparative analysis: The first 25 years. *European Political Science Review, 6*(01), 115–142.

Mercklé, P., & Zalc, C. (2014). Trajectories of the persecuted during the Second World War: Contribution to a microhistory of the Holocaust. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 171–190). Berlin: Springer.

Platt, L. (2011). *Understanding inequalities: Stratification and difference*. Cambridge: Polity.

Ragin, C. C. (1987). *The Comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.

Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University of Chicago Press.

Ragin, C. C., & Becker, H. S. (1992). *What is a case? Exploring the foundations of social inquiry*. Cambridge/London: Cambridge University Press.

Raitano, M., & Struffolino, E. (2013). Traiettorie lavorative e salariali a inizio carriera in Italia: un'analisi longitudinale. *Stato e Mercato, 3*(99), 389–422.

Rosenfeld, R. A., Trappe, H., & Gornick, J. C. (2004). Gender and work in Germany: Before and after reunification. *Annual Review of Sociology, 30*, 103–124.

Sager, F., & Thomann, E. (2016). Multiple streams in member state implementation: Politics, problem construction and policy paths in Swiss asylum policy. *Journal of Public Policy, 37*(3), 1–28.

Sainsbury, D. (1999). Gender, policy regimes, and politics. In D. Sainsbury (Ed.), *Gender and welfare state regimes*. Oxford: Oxford University Press.

Schneider, C. (2009). *Consolidation of democracy: Comparing Europe and Latin America*. New York: Routledge.

Schneider, C. Q., & Rohlfing, I. (2013). Combining QCA and process tracing in set-theoretic multi-method research. *Sociological Methods & Research, 42*(4), 559–597.

Schneider, C. Q., & Wagemann, C. (2012). *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis*. New York: Cambridge University Press.

Shanahan, M. J. (2000). Pathways to adulthood in changing societies: Variability and mechanisms in life course perspective. *Annual Review of Sociology, 26*, 667–692.

Struffolino, E., Studer, M., & Fasang, A. E. (2016). Gender, education, and family life courses in East and West Germany: Insights from new sequence analysis techniques. *Advances in Life Course Research, 29*, 66–79.

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. LIVES Working Papers 24, NCCR LIVES, Switzerland.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179*(2), 481–511.

Studer, M., Ritschard, G., Gabadinho, A., & Muller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research, 40*(3), 471–510.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools, 13*(1), 12–28.

Thelen, K. (1999). Historical institutionalism in comparative politics. *Annual Review of Political Science, 2*(1), 369–404.

Thiem, A., & Duşa, A. (2013). QCA: A package for qualitative comparative analysis. *R Journal, 5*(1), 87–97.

Vis, B. (2009). Governments and unpopular social policy reform: Biting the bullet or steering clear? *European Journal of Political Research, 48*(1), 31–57.

Williams, T., & Gemperle, S. M. (2017). Sequence will tell! Integrating temporality into set-theoretic multi-method research combining comparative process tracing and qualitative comparative analysis. *International Journal of Social Research Methodology, 20*(2), 121–135.

Wilson, M. C. (2014). Governance built step-by-step: Analysing sequences to explain democratization. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 213–227). Berlin: Springer.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Combining Sequence Analysis and Hidden Markov Models in the Analysis of Complex Life Sequence Data

Satu Helske, Jouni Helske, and Mervi Eerola

1 Introduction

Longitudinal data often consists of multiple parallel sequences that ought to be analyzed jointly. For example, life course data may contain sequences of employment, family formation, and residence. Such data is often referred to as multichannel or multidimensional sequence data. A multichannel approach often gives a simpler representation of the data as opposed to combining states across life domains (the extended alphabet approach); the latter approach rapidly grows the state space as the number of channels and/or states grows. If some data is only partially observed, the multichannel approach also allows for handling data as it is instead of having to make difficult decisions on how to combine observed and unobserved states (Helske and Helske 2018).

Joint analysis of complex multidimensional data poses several challenges. Multichannel sequence analysis (Gauthier et al. 2010) has been the standard tool for the analysis of multichannel sequence data (for empirical applications see, e.g., Eerola and Helske 2016; Müller et al. 2012; Spallek et al. 2014). This approach is

S. Helske (✉)
Institute for Analytical Sociology, Linköping University, Linköping, Sweden

Department of Sociology, University of Oxford, Oxford, UK

Department of Mathematics and Statistics, University of Jyvaskyla, Jyvaskyla, Finland
e-mail: satu.helske@liu.se

J. Helske
Department of Science and Technology, Linköping University, Linköping, Sweden

Department of Mathematics and Statistics, University of Jyvaskyla, Jyvaskyla, Finland

M. Eerola
Centre of Statistics, University of Turku, Turku, Finland

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_11

simple and fast in computing dissimilarities between sequences, and cluster analysis is often used for grouping similar sequences. Describing and visualizing results is, however, often challenging.

We propose an approach for compressing the information within multichannel sequences and for facilitating the interpretation of such data by finding (1) groups of similar trajectories and (2) similar phases within trajectories belonging to the same group. For the first task we use the standard multichannel sequence analysis approach and for the second task we propose using hidden Markov models (HMMs). With the help of HMMs the data can then be illustrated with a graph showing typical phases within trajectories and the transitions between them and/or shown as simplified (single-channel) trajectories consisting of these typical phases. We illustrate this approach with an empirical application to complex longitudinal life course data but such an approach, and HMMs in general, are useful in various longitudinal problems across disciplines.

Hidden Markov models have been widely used in economics, bioinformatics, and engineering (see, e.g., MacDonald and Zucchini 1997; Durbin et al. 1998; Rabiner 1989), often to study single long sequences such as time series. In social sciences, such models are commonly referred to as latent Markov (chain) models (Wiggins 1955, 1973; Van de Pol and De Leeuw 1986); typically they have been used for analysing panel data with a few measurement points. In the social science framework, Vermunt et al. (1999) extended the HMM to include individual covariates and Bartolucci et al. (2007) further developed it for multichannel observations. See also Taushanov and Berchtold (2018) in this bundle.

Hidden Markov modelling have been applied in various longitudinal settings; for accounting for measurement error and unobserved heterogeneity (e.g., Van de Pol and Langeheine 1990; Poulsen 1990; Breen and Moisio 2004; Vermunt et al. 2008; Pavlopoulos and Vermunt 2015), for finding latent sub-populations (e.g., Van de Pol and Langeheine 1990; McDonough et al. 2010; Bassi 2014), and for detecting true unobservable states (e.g., various periods of the bipolar disorder in Lopez 2008).

To the best of our knowledge, few papers apply HMMs to multichannel social sequence data and they all consider binary observations. Bartolucci et al. (2007) studied criminal trajectories of 11,400 offenders, applying HMMs to ten-channel data with six time points. Ip et al. (2015) analysed and classified 18-item profiles of food security among 248 Latino farm worker households in the USA for eight time points. Rijmen et al. (2008) studied 12 parallel trajectories of emotions at 63 time points among 32 anorectic patients. Our analysis extends this framework into multichannel data with much longer and multinomial sequences.

The rest of the paper is structured as follows. We start by giving an introduction to HMMs (we assume that the reader is familiar with sequence analysis and refer to the introduction chapter in this book for the less experienced). We then proceed to framing our goals in the context of complex life course data. We continue by describing the data and the empirical analysis and show the results. We conclude with discussing the usefulness of the method, the challenges it poses, and mention some future directions.

2 Hidden Markov Model

In hidden Markov models, observations are related to a hidden process following a Markov chain. Hidden states can only be detected through the observed sequence(s), as they generate or "emit" observations on varying probabilities.

Let us assume we have multichannel sequence data with N individuals, T timepoints, and C channels and a hidden Markov model with S hidden states. Now $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iT})$ represents the hidden state sequence for individual $i = 1, \ldots, N$ from time 1 to time t and y_{itc} denotes the observation of individual i at time $t = 1, \ldots, T$ in channel $c = 1, \ldots, C$.

Figure 1 illustrates the structure of an HMM for two-channel data. The first order Markov assumption states that the probability of transitioning to the hidden state at time t only depends on the hidden state at the previous time point $t-1$. Here we also assume the same latent structure applies to all channels, i.e., hidden state z_{it} emits observed states y_{itc} in all channels c and observations $y_{it1}, \ldots, y_{itC}$ are assumed conditionally independent given the hidden state z_{it}.

The following probabilities characterize a discrete first-order hidden Markov model for multichannel data:

- *Initial probability* vector $\pi = \{\pi_s\}$ of length S, where π_s is the probability of starting from the hidden state s:

$$\pi_s = P(z_{i1} = s); \quad s \in \{1, \ldots, S\}.$$

- *Transition probability* matrix $A = \{a_{sr}\}$ of size $S \times S$, where a_{sr} is the probability of moving from the hidden state s at time $t-1$ to the hidden state r at time t:

$$a_{sr} = P(z_{it} = r | z_{i(t-1)} = s); \quad s, r \in \{1, \ldots, S\}.$$

- *C emission probability* matrices $B_c = \{b_s(m_c)\}$ of size $S \times M_c$, where $b_s(m_c)$ is the probability of the hidden state s emitting the observed state m_c in channel c and M_c is the number of observed states in channel c:

$$b_s(m_c) = P(y_{itc} = m_c | z_{it} = s); \quad s \in \{1, \ldots, S\}, \; m_c \in \{1, \ldots, M_c\}.$$

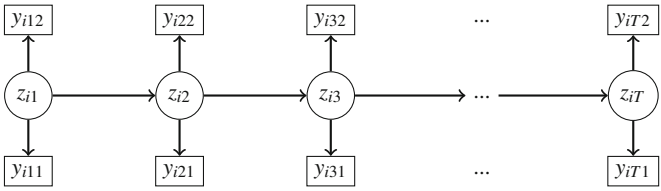

Fig. 1 Illustration of hidden and observed state sequences in a hidden Markov model for two-channel data of individual i. The hidden state at time t is illustrated with z_{it} inside a circle and the observed state at time t in channel c with y_{itc} inside a rectangle. Arrows indicate dependencies between states

Typically, the maximum likelihood estimates of these probabilities are calculated with the Baum–Welch algorithm, i.e., the expectation–maximization (EM) algorithm for HMMs (Baum and Petrie 1966; Rabiner 1989). The most probable path of hidden states for each subject given their observations and the model can be computed using the Viterbi algorithm (Viterbi 1967; Rabiner 1989). Missing observations are handled straightforwardly. When observation y_{itc} is missing, it does not contribute to the estimation of model parameters nor hidden states. See Helske and Helske (2018) for a more extensive presentation on HMMs for multichannel data.

3 Combining Sequence Analysis and Hidden Markov Models for Complex Life Sequences

For analysing complex life sequence data, we aim to compress the information into two types of components:

1. groups with similar life course patterns and
2. typical life stages within each group.

The first component corresponds to finding clusters or latent classes of individuals who have experienced similar life events in similar order and timing. The other, time-varying components should correspond to life stages during which individuals are more likely to have similar experiences, e.g., observed states within the sequences. These life stages could be either stable episodes between two transitions (e.g., employed and married without children) or characterized by transitions in some of the life domains (e.g., moving between unemployment and short-term jobs). Individuals may, and typically do, go through several different life stages during their life course.

SA followed by cluster analysis is a typical strategy for grouping life trajectories. Hidden Markov models, in turn, may be used for finding time-varying latent structures and transitions between them. At first, we use multichannel SA to compute pairwise dissimilarities and then group individuals into clusters. Separate HMMs are then fitted for each cluster. The number and nature of the hidden states are determined independently for each group.

We estimate left-to-right HMMs where transitions to previous hidden states are not allowed. We had several reasons to do this. First, left-to-right models are simpler to estimate since some of the parameters are restricted to zero. Second, due to the nature of the life trajectories, also the observed states tend to show a left-to-right behaviour and many of the HMMs would end up being estimated close to left-to-right models anyway. Third, we find that left-to-right models are often easier to interpret in the context of life course: individuals go through different life stages but even if they return to have a similar life stage compared to a previous one – say re-marriage after a divorce – this second life stage comes with a different history compared to the first time.

4 Data

We illustrate the analysis of complex life sequence data using a subsample of the German National Educational Panel Survey (NEPS) (Blossfeld et al. 2011). We restricted the analysis to the life courses of an age cohort born in 1955–1959. Only individuals who were born in Germany or moved there before the age of 14 are included.

The data consisted of monthly life statuses of 1731 individuals in three life domains (labour market participation, partnerships, and parenthood) from age 15 to age 50. For each individual, there were three parallel sequences of length 434, which made altogether 2,253,762 data points (of which 2,232,730 were observed and 21,032 were missing). Using the monthly time scale also allowed for the detection of smaller fluctuations in life courses, e.g. recurrent transitions between short-term unemployment and employment.

4.1 Sequences

The sequences in three life domains were constructed as follows:

Labour market participation with 4 states:

- Studying (in school, vocational training, or vocational preparation)
- Employed (full-time or part-time)
- Unemployed
- Out of the labour market (for other reason than studies, e.g., parental leave, taking care of children or other family members, military or non-military service, voluntary work, or other gap in the employment history)

Partnerships with 4 states:

- Single (never lived with a partner)
- Cohabiting
- Married/in a registered partnership
- Divorced/separated/widowed

Parenthood with 2 states:

- No children
- Has (had) children (biological, adopted, or foster children)

The coding for parenthood was very simple. A practical reason was that this record was available for most individuals, whereas more detailed information was often missing. On the other hand, we can argue that specifically the experience of becoming a parent is relevant as one step in the developmental process into adulthood.

For the latter two life domains, the status of each month was usually determined from the latest event. An exception was made for the rare partnerships that lasted for less than a month; there separation was coded from the following month onward. In a case of multiple records per month in the career domain, the final status was given according to assumed importance: school and vocational training came before employment, which in turn dominated over vocational preparation, unemployment, and other non-employment statuses.

Altogether 306 individuals (17.7%) had some missing information in one or two life domains. Thus, at each time point we had at least some information from each individual.

5 Analysis

We have little prior knowledge on the structure of the model; hence, how many clusters to choose and how many hidden states to include in each cluster? Since the complexity of these types of life course trajectories varies a lot (e.g., some individuals have no family-related transitions while others have many), we expected the groups to have varying numbers of hidden states.

5.1 Sequence Analysis and Clustering

We started by applying multichannel sequence analysis and computed the dissimilarities between the sequences. These were then used in cluster analysis.

The dissimilarities between sequences were determined according to the generalized Hamming distance with user-defined substitution costs (see Table 1). We set the highest cost to be the same in all life domains to give them equal weight. We gave no cost for substituting missing states since we wanted to determine dissimilarity based on the observed trajectories. Regarding the costs within different life domains, our choices were mainly based on how far the states are regarded on the pathway to adulthood and, in terms of labour market participation, also on how close the other states can be regarded to employment which is often the favourable state. The metric compares observed states time point by time point and gives a cost for mismatches. It generally works well in a multichannel problem where timing is important (Studer and Ritschard 2016) and resulted in meaningful clusters with high goodness-of-fit.

We used Ward's clustering method for the Hamming dissimilarities and chose six clustering solutions with 7–12 clusters for further examination. The choice was based on goodness-of-fit statistics, the dendrogram, and the interpretability of the clusters. Ward's method was chosen because it typically produces usable and relatively even-sized clusters compared to most of the other clustering methods (Aassve et al. 2007; Helske et al. 2015). Also, the method is hierarchical (agglomerative), so when two smaller clusters are merged, all other clusters remain the same. This

Table 1 Substitution costs for Hamming distances in three life domains: labour market participation, partnerships, and parenthood

Labour market participation

	→ ST	→ EM	→ UN	→ OU	→ *
Studying (ST) →	0	3	2	1	0
Employed (EM)→	3	0	2	2	0
Unempl. (UN) →	2	2	0	1	0
Out of LM (OU)→	1	2	1	0	0
Missing (*)→	0	0	0	0	0

Partnerships

	→ S	→ C	→ M	→ D	→ *
Single (S)→	0	2	2	3	0
Cohab. (C)→	2	0	1	2	0
Married (M)→	2	1	0	2	0
Div./sep. (D)→	3	2	2	0	0
Missing (*)→	0	0	0	0	0

Parenthood

	→ NC	→ CH	→ *
No child (NC)→	0	3	0
Has child (CH)→	3	0	0
Missing (*)→	0	0	0

means that among the $7 + 8 + 9 + 10 + 11 + 12 = 57$ clusters in the six sets of clustering results, only $7 + 2 + 2 + 2 + 2 + 2 = 17$ were unique, resulting in significant decrease in the number of models to be estimated compared to non-hierarchical clustering.

5.2 Hidden Markov Models for Clusters

At the next step, we estimated five HMMs with 4–8 or 5–9 hidden states separately for each of the 17 unique clusters—fewer hidden states for clusters with simpler observed trajectories, more for the more complex ones. Since the goal was to find general life stages between adolescence and middle age in a given group, having too few or too many hidden states was not plausible nor interpretable. When increasing the number of hidden states, at some point they lost their distinctive nature (consecutive states had very similar emission probabilities) and/or they were rarely "visited" in the most probable paths of hidden states.

A well-known problem with the HMM estimation is that most of the optimization methods are sensitive to initial estimates of the parameters. In order to reduce the risk of being trapped in a poor local optimum, we estimated the models numerous times with random starting values. We continued re-estimation until we had found the same optimum for at least 100 times (which turned out to be much more than necessary).

For each cluster, we compared the HMMs with a different number of hidden states to find the best model. Bayesian information criterion (BIC) and other information criteria are common choices for comparison of HMMs with different numbers of hidden states. Another common option for model selection is cross-validation.

We chose to use BIC as it generally selects parsimonious models. Unfortunately, here BIC kept suggesting models with more and more states. We did, however, use BIC as one source of information for choosing the number of hidden states by looking for turning points in BIC after which additional hidden states offered little improvement. In addition to BIC, the choice of the number of hidden states was based on the interpretability of the model and the prevalence of the hidden states in the individual trajectories.

5.3 Software

Analyses were conducted with the R software (R Core Team 2015) by using the packages TraMineR (Gabadinho et al. 2011) for sequence analysis, cluster (Maechler et al. 2015) for cluster analysis, and seqHMM (Helske and Helske 2018) for hidden Markov modelling. For the estimation of HMMs we used the automatic re-estimation routine for the EM algorithm provided in the model estimation function.

6 Results

The number of hidden states per cluster varied between six and eight. The model of eight clusters resulted in the smallest BIC (even the highest likelihood) and was chosen as the best model. We present a few different ways to describe the results: a table showing the most typical transitions in each cluster, a figure illustrating the structure of the HMMs, and a figure of the most probable hidden states, i.e., the trajectories of general life stages for each individual.

Table 2 describes each cluster in terms of some important transitions and states: typical labour market participation (showing the timing of completing education and the type of employment after that), partnership histories (age at first partnership, the type and number of partnerships), and parenthood (the timing of the first child). It also shows the number of individuals in each cluster and the proportion of the sample, as well as the hidden states described with the most important transitions.

Figure 2 illustrates the HMM structure for each of the eight clusters. It shows the HMMs as directed graphs where the pies represent hidden states and the slices show the emission probabilities of observed states within each hidden state (to draw attention to the most prevalent observations, we only show probabilities that are greater than 0.05). The arrows indicate transition probabilities between the hidden states—the thicker the arrow, the higher the probability.

Figure 3 illustrates the most probable hidden state paths. We have assigned similar colours to similar hidden states across clusters.

As an example of how to interpret these figures, let us look at the smallest of the clusters titled Single parents (cluster H). All individuals start from the first hidden

Table 2 Description of clusters by typical timing of the completion of education, type of employment, the timing, number, and types of typical partnerships, and the timing of parenthood. Hidden states are described with changes in the most probable observations (ordered by prevalence) (out = out of the labour market (not studying), div. = divorced or separated). For all clusters, the first two hidden states are omitted as they are approximately the same: hidden state 1 is *studies, single, no children* and hidden state 2 is *employed, single, no children*

Clusters	Education	Employment	1st partnership	Partnerships	Parenthood	N	%	Women (%)	Hidden states
Short education & early family (A)	Before 20	Mostly employed	Early 20s	1 or 2 marriages / marriage + cohab.	Early 20s	461	27	59	3. Empl./studies, married/cohab. → 4. Empl., married, child → 5. Div./cohab. → 6. Unempl./out/studies, married → 7. Empl.
Short education & later family (B)	Early 20s	Steady employment / many out of empl.	Mid-20s	1 long marriage	Around 30	403	23	46	3. Cohab./div. → 4. Married → 5. Empl./out, child → 6. Out → 7. Empl.
Long education & later family (C)	Mid-20s	Mostly empl. / some out of empl.	Varying	Long cohab., 1 long marriage	30s	266	15	32	3. Empl./studies, cohab./div. → 4. Married → 5. Out/empl./unempl./studies, child → 6. Empl., cohab./div. → 7. Empl., married
Career break & early family (D)	Before 20	Out of empl., some employed after 35	Early 20s	1 long marriage	Early 20s	159	9	96	3. Married/cohab. → 4. Empl./out, child → 5. Out, married → 6. Unempl./studies → 7. Empl. → 8. Div.
Partnership(s) & no child (E)	Varying	Varying, but mostly employed	Early 20s	1 long marriage / multiple partners	No child	177	10	49	3. Empl./studies, cohab. → 4. Empl., married → 5. Div. → 6. Cohab. → 7. Empl./out/unempl., married/div.
No or late family (F)	Varying	Mostly employed / some long unempl. after 35	Never / After 35	0 / 1 cohab. or marriage	Never / After 40	116	7	41	3. Unempl./studies/out → 4. Empl. → 5. Cohab./div. → 6. Married, no child/child
Divorced / separated parents (G)	Varying, mostly before 20	Mostly employed, some out of empl.	Varying, mostly early 20s	1 cohab. or marr., div. / sep. during 30s	Varying, typically late 20s	102	6	61	3. Empl./studies, cohab./married/div. → 4. Married, child → 5. Empl./unempl./out, div. → 6. Empl., cohab./married → 7. Div.
Single parents (H)	Before / early 20s	Mostly employed, some out of empl.	Never / After 35	0 / 1 cohabitation	Varying, typically late 20s	47	3	72	3. Out/empl./studies, child → 4. Empl. → 5. Unempl./empl./studies → 6. Empl., cohab./married/div.

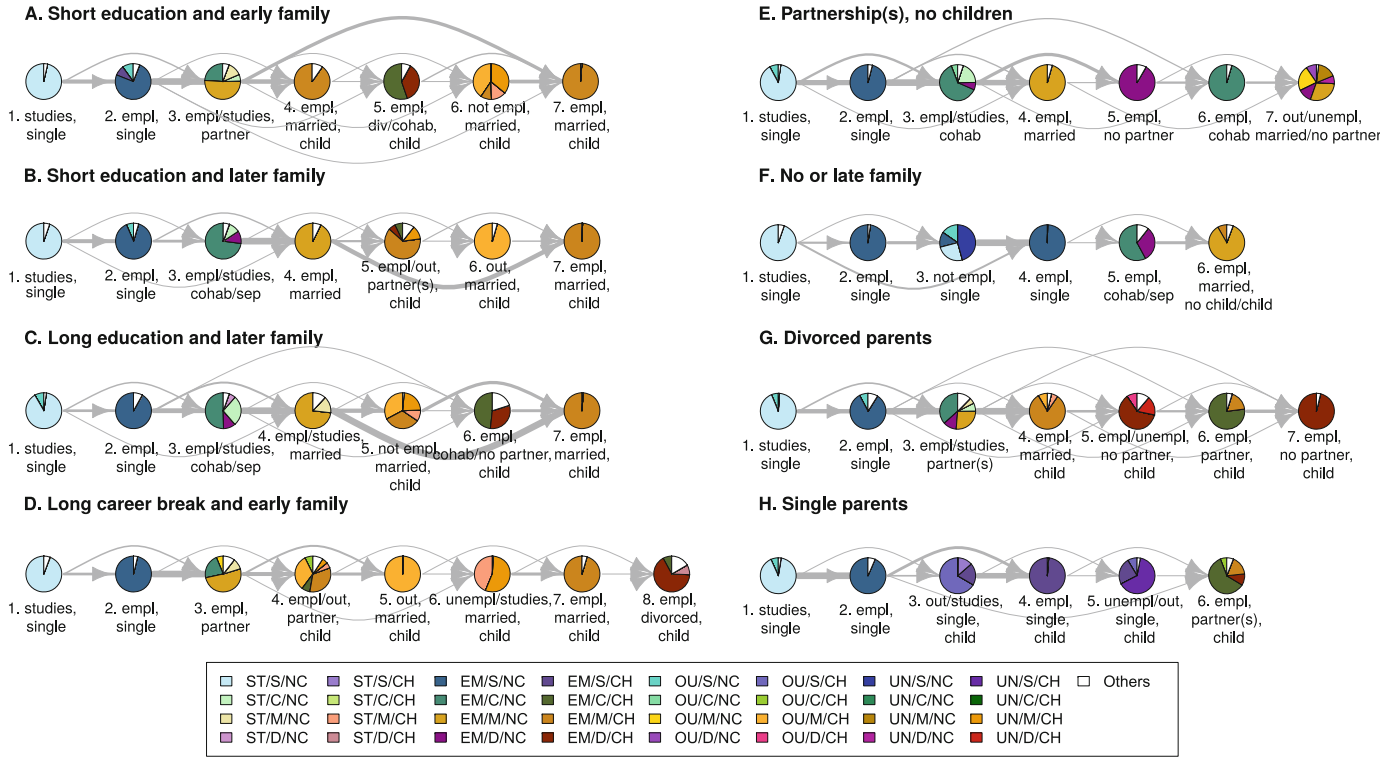

Fig. 2 HMM graphs for the eight clusters A–H. State abbreviations show labour market/partnership/parenthood statuses: ST = studying, EM = employed, OU = Out of the labour market, UN = unemployed; S = single, C = cohabiting, M = married, D = divorced/separated; NC = no children, CH = has child(ren). Hidden states are described by the most probable observations

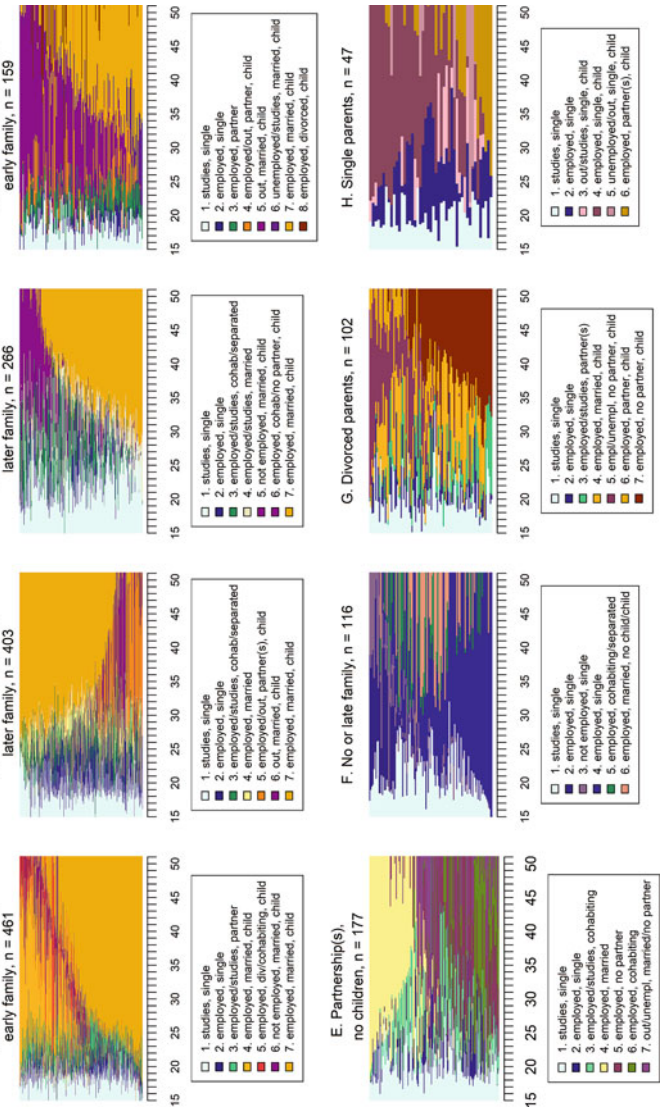

Fig. 3 Most probable hidden state paths between ages 15 and 50 for individuals in eight clusters. Hidden states are described with the most probable observed states showing labour market participation (studying/employed/unemployed/out of the labour market), partnership statuses (single/cohabiting/married/divorced/separated; also partner = cohabiting or married, no partner = divorced/separated from marriage or after cohabitation), and parenthood status (if has had children). Multiple relevant observed states within a life domain are ordered by emission probabilities. See Fig. 2 for visualizations of the hidden states in each cluster

state (State 1, indicated with light blue in the hidden state paths), a life stage where they are childless singles and mostly studying. For almost all, the next transition is to State 2 (dark blue), moving to employment. A few make a straight transition to State 3 (light pink), a life stage of becoming parents and being out of the workforce. State 4 (darker purple) describes a life stage during which individuals are singles, have children, and are employed. This is the most prevalent life stage for the members of this cluster and many stay there until the end of the follow-up. A few move out of employment, mostly to unemployment (State 5, light purple). During the last life stage, experienced by almost half of the members, individuals form their first partnerships (State 6, yellow).

In general, the clusters were well separated from each other by the timing and occurrence of labour market participation and family states. The two largest clusters composing of half of the respondents were characterized by (mostly) short education and family. The biggest difference was in the timing of partnership and parenthood transitions which occurred either earlier in life (cluster A) or later (cluster B). The third largest cluster (cluster C) mostly consisted of individuals, more often men, who had long education and later family transitions. Another cluster with early family transitions (cluster D) consisted of mostly women and was characterized with a long career break for mostly taking care of children.

Two clusters were characterized by no or very late parenthood. They differed in the timing of the partnerships; the larger cluster (cluster E) had earlier first partnerships while in the smaller cluster (cluster F) partnerships were delayed or omitted altogether. The two smallest clusters consisted of parents living divorced or separated (cluster H) or single parents (cluster G).

7 Discussion

When analysing complex sequence data with multiple channels, describing and visualizing the data can be a challenge. By combining sequence analysis and hidden Markov models the information in data can be compressed into hidden states (life stages) and clusters (general patterns in life courses). Hidden states were able to capture general life stages that included not only rather stable episodes such as being employed and married with children (e.g., State 7 for cluster A) but also life stages characterized by change, e.g., moving between unemployment and short-term employment (State 3 for cluster F).

We presented two different ways of HMM-based visualizations that give complementary information but could also be shown alone—it is up to the researcher to decide which one is more informative in each case. The HMM graphs show the structure of the hidden states and the transitions between them; also parameter estimates could be easily included in the graph. The most probable paths of hidden states show individual-level information on the approximate prevalence and timing of different life stages.

Despite its usefulness as a data reduction technique, this approach comes with some challenges. A major one is the estimation of several HMMs when the number of hidden states and clusters is unknown. For these challenges, we used a few approaches. In terms of the number of clusters, we used a hierarchical clustering method which reduced the number of models to be estimated compared to non-hierarchical clustering. We then estimated a single model numerous times with randomized starting values to find the one with the highest likelihood, using parallel computation for improved efficiency.

Another issue is that we take the SA clusters as fixed. In reality, there is, of course, a lot of uncertainty which we do not take into account. Also, we do not discuss other trajectory grouping techniques besides SA. To our knowledge, there are not many methods suitable for multichannel sequence data; we experimented with latent class analysis (Collins and Wugalter 1992) which did not lead to satisfactory results. On the other hand, regarding the parameter uncertainty, in theory it is possible to compute asymptotic standard errors from the Hessian matrix obtained from the numerical optimization algorithms, but in practice the underlying asymptions are typically not met (Zucchini and MacDonald 2009).

The mixture hidden Markov model (MHMM) offers a solution to the problem of uncertainty of clustering. In the MHMM, instead of fixing individuals to the clusters defined during the SA step, we could use all data to estimate a mixture of HMMs where each individual belongs to each cluster with some probability (preferably with a large probability for one cluster and a small probability to all others). In a complex setting, SA can be used to determine the range of potential clustering structures. It can also be of aid when setting initial values for the estimation process, which is often essential when using very large models.

Although in theory the MHMM approach allows even more flexibility to the modelling and potential for more interesting ways of inference, there are some practical computational problems in the MHMM methodology. The parameter estimation of HMMs is often very sensitive to initial values, and the computational costs increase rapidly when the number of hidden states grows. These problems are even more prominent in complex MHMMs, especially when the structure of the model (in terms of the number of hidden states and/or clusters) is not known. For this study, we were not able to find stable solutions for MHMMs despite large computational resources available—the multichannel structure, long sequences, and the relatively large number of individuals in our data was too challenging a combination for parameter estimation. Nevertheless, the MHMM can be useful in other settings. It has been successfully used for simpler problems, e.g., for accounting for measurement error and unobserved heterogeneity.

An extention not covered in this paper is the inclusion of external covariates. Personal characteristics and other relevant factors, time-constant as well as time-varying, could be used to predict transition probabilities between life stages. In MHMMs, time-constant covariates may also be used to predict cluster memberships. See, e.g., Vermunt et al. (2008) for a general presentation of such models.

We are currently studying algorithmic variations which can reduce the computational complexity of the MHMM estimation. Further research is also needed regarding model selection and the goodness-of-fit of left-to-right HMMs and MHMMs. Further theoretical and empirical studies are needed for detecting the reasons for the failure of BIC and for discovering selection criteria that are better suited for finding parsimonious HMMs.

Another topic for future research is the potential of hidden Markov models and Markovian models in general as mechanisms of generating social sequence data.

The aim of our study was to describe complex life sequence data and for that goal, the SA-HMM approach gave satisfactory results in a reasonable time. We were able to find meaningful and well-separating clusters and to visualize their complex life course information by using HMM graphs and the most probable paths of life stages for each individual.

Acknowledgements This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 6–Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:3.0.1. From 2008 to 2013, the NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research and supported by the Federal States. As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi).

Satu Helske is grateful for support for this research from the John Fell Oxford University Press (OUP) Research Fund and the Department of Mathematics and Statistics at the University of Jyväskylä, Finland, and Jouni Helske for the Emil Aaltonen Foundation and the Academy of Finland (research grant 284513).

We also wish to thank three anonymous referees for their helpful comments and suggestions.

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.

Bartolucci, F., Pennoni, F., & Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(1), 115–132.

Bassi, F. (2014). Dynamic segmentation of financial markets: A mixture latent class markov approach. In M. Carpita, E. Brentari, & E. M. Qannari (Eds.), *Advances in latent variables* (pp. 61–72). Berlin/Heidelberg: Springer.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 67*(6), 1554–1563.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.) (2011). *Education as a lifelong process-the German national educational panel study (NEPS)* (Vol. 14) [Special Issue] of *Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer.

Breen, R., & Moisio, P. (2004). Poverty dynamics corrected for measurement error. *The Journal of Economic Inequality, 2*(3), 171–191.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*(1), 131–157.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.

Eerola, M., & Helske, S. (2016). Statistical analysis of life history calendar data. *Statistical Methods in Medical Research, 25*(2), 571–597.

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.

Helske, S., & Helske, J. (2018, forthcoming). Mixture hidden Markov models for sequence data: The seqHMM package in R. *Journal of Statistical Software*.

Helske, S., Steele, F., Kokko, K., Räikkönen, E., & Eerola, M. (2015). Partnership formation and dissolution over the life course: Applying sequence analysis and event history analysis in the study of recurrent events. *Longitudinal and Life Course Studies, 6*(1), 1–25.

Ip, E. H., Saldana, S., Arcury, T. A., Grzywacz, J. G., Trejo, G., & Quandt, S. A. (2015). Profiles of food security for US farmworker households and factors related to dynamic of change. *American Journal of Public Health, 105*(10), e42–e47.

Lopez, A. (2008). *Markov models for longitudinal course of youth bipolar disorder*. Ph.D. thesis, University of Pittsburgh, Ann Arbor, MI.

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series* (Vol. 110). Boca Raton: CRC Press.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2015). *Cluster: Cluster analysis basics and extensions*. R package version 2.0.3.

McDonough, P., Worts, D., & Sacker, A. (2010). Socioeconomic inequalities in health dynamics: A comparison of Britain and the United States. *Social Science & Medicine, 70*(2), 251–260.

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. *International Journal of Social Psychiatry, 58*(3), 266–277.

Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment: Do survey or register data tell the truth? *Statistics Canada, Catalogue No. 12–001-X, 41*(1), 197–214.

Poulsen, C. S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing, 7*(1), 5–19.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika, 73*(2), 167–182.

Spallek, M., Haynes, M., & Jones, A. (2014). Holistic housing pathways for Australian families through the childbearing years. *Longitudinal and Life Course Studies, 5*(2), 205–226.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179*(2), 481–511.

Taushanov, Z., & Berchtold, A. (2018). Markovian-based clustering of internet addiction trajectories. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Van de Pol, F., & De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods & Research, 15*(1–2), 118–141.

Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology, 20*, 213–247.

Vermunt, J. K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics, 24*(2), 179–207.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 373–385). Burlington: Elsevier.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory, 13*(2), 260–269.

Wiggins, L. M. (1955). *Mathematical models for the interpretation of attitude and behavior change: The analysis of multi-wave panel*. Ph.D. thesis, Columbia University, New York.

Wiggins, L. M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes.* Oxford: Jossey-Bass.

Zucchini, W., & MacDonald, I. L. (2009). *Hidden Markov models for time series: An introduction using R* (Vol. 110). Boca Raton: CRC Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part V
Advances in Sequence Clustering

Markovian-Based Clustering of Internet Addiction Trajectories

Zhivko Taushanov and André Berchtold

1 Introduction

The clustering of trajectories has gained much interest in recent years from the scientific community, especially in the social sciences, because the number of longitudinal studies, as compared to cross-sectional ones, has been constantly increasing. As regards categorical data, the most common approach relies on the Optimal Matching (OM) to compute a distance between each pair of trajectories before clustering them, whereas the Growth Mixture Model (GMM) can be applied for continuous data. However, these two approaches suffer from some shortcomings, calling for the need to develop and apply alternative approaches. For instance, OM requires the choice of a substitution cost measure and other parameters. GMM gives a lot of importance to the shape of sequences. Therefore, there is a risk to overfit the data when nonlinear trajectories are considered on quite short sequences. The other issues of GMM include computational load, presence of local optima, missing data treatment, model selection criteria, the need for large sample size, and unclear Type I error rates (Wang and Bodner 2007).

In this paper, we study the use of a specific class of Markovian Models called the Hidden Mixture Transition Distribution (HMTD) model (Bolano and Berchtold 2016) for clustering purpose. Even if this model-based approach was developed as a tool for the analysis of continuous trajectories, it also allows for their clustering without a priori knowledge of cluster membership. Moreover, covariates can be easily included in the model.

The HMTD and GMM clustering approaches are applied and compared on a dataset of trajectories of the Internet Addiction Test (IAT). Excessive Internet use,

Z. Taushanov (✉) · A. Berchtold
Institute of Social Sciences and NCCR LIVES, University of Lausanne, Lausanne, Switzerland
e-mail: zhivko.taushanov@unil.ch; andre.berchtold@unil.ch

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_12

especially among youths, is an emerging health issue in the medical literature, with studies showing contrasting results. Surís et al. (2014) show a significant association between problematic Internet use and several somatic disorders, including back, weight, musculoskeletal, and sleep problems. Moreover, several chronic conditions are also significantly associated with problematic Internet use. In contrast, another study finds no significant Internet use effect on the development of overweight among youths (Barrense-Dias et al. 2015).

While several alternative approaches (e.g. Skarupova et al. 2015) have been introduced over the years, the tool most often used to quantify the degree of addiction to Internet is still the Internet Addiction Test (IAT), developed by Young (1998). However, since the test's scale is based on 20 items and is quite long, its psychometric properties are matters of controversy (Faraci et al. 2013) and the test is not considered suitable for the successive measurement of the same subjects (test-retest). Its use in longitudinal contexts remains problematic because of the difficulty to distinguish between the real evolution of subjects and changes due to behavior of the IAT itself.

To gain information on the behavior of the IAT in longitudinal studies, we need to compare the typical trajectories of the repeated IAT measurements with other characteristics of the subjects under study. Thus, we first cluster the IAT trajectories into a finite set of meaningful groups and then compare these groups with the known characteristics of subjects that are either time-invariant or evolve over time. Specifically, the goals of this study are (1) to separate the Internet addiction trajectories into an optimal number of meaningful categories using HMTD, (2) to explore how does the introduction of the covariates influence the previous optimal partition, and (3) to compare the HMTD clustering with an equivalent GMM clustering in order to gain information on the respective strengths of both approaches. We hypothesize that (1) the IAT scores computed for the same person can vary considerably over time, implying that the trajectories are difficult to classify; (2) a classification using covariates is easier to interpret than a classification without any additional information on the clustered variable itself; and (3) the HMTD approach can lead to more sound and easier-to-use solutions as compared to the solutions obtained using GMM. However, we must stress that it is impossible to conclude that one method is superior to another, especially using real data, without knowing the true cluster membership. So this work must be considered as a first step in the comparison of HMTD and GMM as clustering tools.

2 Data and Methods

2.1 Data

The data we considered are from *ado@Internet.ch* (Surís et al. 2012), a longitudinal study on the use of Internet among youths in the Swiss canton of Vaud (the largest

canton in the French-speaking part of Switzerland). The data were collected five times every six months from Spring 2012 (T_0, baseline) to Spring 2014 (T_4) using an online questionnaire. The data for the first time were collected from schools during the computerlab periods. Then, the students who agreed to participate in the study were contacted again by email from T_1 to T_4 to answer follow-up questionnaires on their home computer. A convenience sample of $n = 185$ adolescents who answered all five questionnaires is used for the present study (67% females; mean age at T_0: 14.1 years). For more details on the overall design of the study and data collection, (see Surís et al. 2012; Piguet et al. 2016).

The main outcome is the IAT score measured at each wave for each subject. The IAT developed by Young (1998) and validated in French by Khazaal et al. (2008) is a scale ranging from 0 to 100, based on the answers to 20 items whose possible answers range from Never (coded 0) to Always (5). Examples of items are, *How often do you find yourself staying online longer than you intended?* and *How often do you fear that life without the Internet would be boring, empty, and joyless?*

In addition to the IAT, we also considered several important characteristics of the subjects, either fixed in time—gender, age at baseline, and education track at baseline (extended requirements vs. basic requirements)—or evolving over time—emotional well-being (measured by the WHO-5 index) and Body Mass Index (BMI, computed from auto-reported measures of height and weight). Note that the WHO-5 index was not evaluated on the third wave of the study, and so for the present paper, we imputed values as the simple mean between the values of the second and fourth waves. Similarly, we imputed the BMI for the second wave of the study as the mean between the values of the first and third waves.

2.2 Clustering Using the HMTD Model

We used a specific class of Markovian Models, the HMTD model, to cluster the longitudinal sequences of continuous data. This model combines a latent and an observed level (Bolano and Berchtold 2016). The visible level is a Mixture Transition Distribution (MTD) model that was first introduced by Raftery in 1985 as an approximation of high-order Markov chains Raftery (1985) and then developed by Berchtold (2001, 2003) and Berchtold and Raftery (2002). Here, we used a Gaussian version of the MTD model, where the mean of the Gaussian distribution is a function of past observations. Because of the small size of each sequence of the observed outcome (five data points, from T_0 to T_4), long dependencies between successive observations could not be considered, and therefore we fix the dependence order for the mean of the Gaussian distributions of each component to one:

$$\mu_{g,t} = \varphi_{g,0} + \varphi_{g,1}\, x_{t-1}$$

where $\varphi_{g,0}$ is the constant for the mean for component g and $\varphi_{g,1}$ is the autoregressive parameter indicating the dependence from the previous observation x_{t-1}. Similarly the variance of each component can be written as a function of the past periods variability: $\sigma^2_{g,t} = \theta_{g,0} + \sum_{s=1}^{S} \theta_{g,s}\, x^2_{t-s}$. However given the small number of time periods in our dataset, and for the sake of simplicity, we decided to treat the variance as a constant: $\sigma^2_{g,t} = \theta_{g,0}$.

In the HMTD model, the latent level is a homogeneous Markov chain. Each state of the chain is associated with a different Gaussian component at the visible level, with the transition matrix used to determine which component best represents the current observation. To use the HMTD model as a clustering tool, we assume the hidden transition matrix to be the identity matrix. Consequently, each sequence of successive observations is associated with only one component of the model, thus generating a clustering of sequences into mutually exclusive groups. Notice that in this case, the resulting model is no more a hidden Markov model, but a mixture of Gaussian distributions. However, it is still interesting to view it as a HMTD, because it is then possible to compare the clustering model with other models, especially with semi-clustering models whose transition matrix is not the identity matrix, but a triangular matrix letting data trajectories move from one group to another in a specific order.

In addition to the clustering based on the IAT variable only, we performed a second clustering adding information from five covariates (gender, age at T_0, education track at T_0, WHO-5, and BMI). These covariates are introduced as additional terms in the specification of the mean of each visible component of the model, and the categorical variables are introduced as dummy variables. We then rewrite the mean of the g-th component as

$$\begin{aligned}\mu_{g,t} = \varphi_{g,0} + \varphi_{g,1}\, x_{t-1} + \varphi_{g,2}\,\text{Gender}(\textit{male}) + \varphi_{g,3}\,\text{Age}\\ +\varphi_{g,4}\,\text{Education}(\textit{extended}) + \varphi_{g,5}\,\text{WHO-5} + \varphi_{g,6}\,\text{BMI}\end{aligned}$$

with *female* and *basic requirements* used as reference modalities for Gender and Education, respectively.

In practice, continuous covariates are centered around the sample mean before computing the clustering model in order to allow for a better convergence of the estimation algorithm. A comparison of the two specifications of the mean, with and without covariates, illustrates whether the inclusion of covariates in the model helps to improve the clustering process. It must be mentioned that, in addition to these two HMTD models, many other specifications were tried, following a hierarchical approach (Bolano and Berchtold 2016), but none of these alternative specifications seemed to give a more useful clustering of IAT trajectories.

The HMTD model is estimated by maximizing its log-likelihood. When the variance of each component of the model is constant, the log-likelihood can be derived with respect to all parameters, but in the general case of time-varying variances (Berchtold 2003), the log-likelihood is generally not differentiable, and the solution space can be very complex. A specific heuristic is then applied to obtain

the solution (Taushanov and Berchtold 2017). Since this heuristic can accommodate to all possible specifications of the HMTD model, we used it for all computations. Regarding cluster assignation, we used the standard Viterbi algorithm which is able to find the best sequence of hidden states in function of the observed data and of the current model (Forney 1973). In the specific case of clustering, the Viterbi algorithm simply assign each observed trajectory to the most likely component.

We used a bootstrap procedure to obtain confidence intervals for each parameter, but since our goal here was to validate not the initial classification itself, but the parameters associated with the model describing each visible component of the model, we adopted the following approach: Instead of performing the bootstrap on the whole original sample, we divided the original sample into as many groups as can be retained in the final classification. We then applied a single-component version of the HMTD model to each sub-sample separately in order to estimate the coefficients using bootstrap. By applying the model on the sub-samples separately, instead of on the initial sample, we avoided the so-called label-switching problem that is very common in latent variable clustering. The inconvenient of separate bootstrapping is that since we rely on the validated clustering solution, we ignore the model uncertainty including the weights of each cluster. We computed the confidence intervals using 1000 bootstrap samples, and we used the results to evaluate the significance of the estimated parameters.

All computations were done using R, and a specific package should be released soon. In the meantime, a first version of the R syntaxes is available on https://github.com/ztau/5352.

2.3 GMM as a Gold Standard Alternative

To evaluate the HMTD approach as a tool for clustering sequences of continuous data, we need a gold standard alternative. We choose the Growth Mixture Model (GMM) approach for that purpose, since it is the only true longitudinal clustering tool used in the social sciences.

Growth modeling includes several similar frameworks aiming to model and discover the patterns of individual changes in a longitudinal data framework (Reinecke and Seddig 2011; McArdle and Epstein 1987). The basic growth model assumes that all trajectories belong to the same population and that they may be approximated by a single average growth trajectory using a single set of parameters. However, several models extend these assumptions; for example, the latent class growth analysis (LCGA) model, which assumes null variance-covariance for the growth trajectory within each class (Nagin 1999; Jung and Wickrama 2008), and the heterogeneity model (Verbeke and Lesaffre 1996), which goes a bit further but still imposes the same variance-covariance structure within each group of subjects. Therefore, we discuss the more flexible GMM in this section and use it in our analysis as gold standard.

The GMM developed by Muthén and Shedden (1999), Bauer and Curran (2003), and Wang and Bodner (2007) is designed to discover and describe unknown groups of sequences that share a similar pattern. This method may be represented as a mixture of mixed-effects models in which each of the unknown subpopulations follows a distinct linear mixed-effects model. Its main advantage over other similar models—like the heterogeneity model (Verbeke and Lesaffre 1996)—is that it allows for estimation of a specific variance-covariance structure within each class (Francis and Liu 2015). Within-class inter-individual variation is possible for latent variables via distinct intercept and slope variances, represented by a class-specific fixed-effects and random-effects distribution. In other words, the variation in an expected group-specific trajectory is distinct for each group (heterogeneity in growth trajectories). Because of these advantages, the model is a reference point in continuous longitudinal data modeling with various applications in criminology (Francis and Liu 2015; Reinecke and Seddig 2011), health and medicine (Muthén and Shedden 1999; Ram and Grimm 2009), psychology, and social science (Muthén 2001), among others.

The GMM approach uses both observed and latent variables. The observed variables consist of a p-dimensional vector of continuous dependent variables Y (often a variable with repeated measurements) and a q-dimensional vector of covariates X. The latent variables are represented as a continuous m-dimensional vector η. Finally, to indicate the group in which each subject is included, we use a dummy variable with multinomial distribution stored in a k-dimensional binary vector c (Muthén and Shedden 1999). The equation of the GMM approach for individual i then becomes

$$Y_i = \Lambda \eta_i + \epsilon_i, \tag{1}$$

where Λ is a $p \times m$ parameter matrix (or matrix with basis vectors) that can be seen as a matrix of factor loadings, η_i is a vector of latent continuous variables, and ϵ_i is an error term vector with zero mean.

In our case, the latent variable parameter matrix Λ has one column with parameters for the latent factor accounting for the intercept and another for the latent factor accounting for the slope. The general equation for every η is

$$\eta_i = Ac_i + \Gamma x_i + \zeta_i, \tag{2}$$

where A is a matrix with columns of intercept parameters for each class, Γ is an $m \times q$ parameter matrix and ζ_i is an $m \times 1$ vector of zero mean residuals (and covariance matrix Ψ).

If we assume that some time-independent covariates z could influence the group membership c_i, a multinomial logistic regression can be considered (with parameters a and b) as follows:

$$P(c_i = K|z_i) = \frac{\exp^{a_k + b_k z_i}}{\sum_{c=1}^{K} \exp^{a_c + b_c z_i}}$$

An alternative notation of the model for subject i as part of class k at time t is

$$Y_{i,t|c_i=k} = X_{1i}(t)^T\beta + X_{2i}(t)^T\gamma_k + V_i(t)^T u_{ik} + w_i(t) + \epsilon_{i,t}, \tag{3}$$

where X_{1i} is a vector of covariates with common fixed effects β, X_{2i} is a vector of covariates with class-specific fixed effects γ_k, and V_i is a set of covariates with individual class-specific random effects u_{ik}. Finally, $w_i(t)$ is an autocorrelated Gaussian process with null mean and covariance equal to $cov(w_i(t)w_i(s)) = \sigma_w^2 \exp(-\rho|t-s|)$.

The GMM is estimated by maximizing its likelihood using an ordinary EM algorithm. The continuous latent variables η and group membership variables c are considered missing data. The R package *lcmm* (Proust-Lima et al. 2017) was used to compute the GMM.

2.4 Statistical Analyses

To start with, we used the HMTD model to identify the best clustering of the IAT dataset without covariates, considering solutions from two to five groups. The best solution was selected on the basis of the Bayesian Information Criterion (BIC) (Raftery 1995). We then added covariates to this first model and analyzed the two resulting models, with and without covariates, particularly focusing on the IAT trajectories that did change group when covariates were added to the initial model. In order to isolate the impact of the covariates from any other computational issue or local optimum, we used the optimal solution obtained without covariates as a starting point for the full model. Therefore, we observe how this new model escapes the previous optimum.

We then computed the GMM models using the same dataset, and we compared the classifications obtained with the HMTD and GMM approaches. The usefulness of each covariate for discriminating between groups was evaluated using either a chi-square test for categorical covariates, or a single factor ANOVA for continuous ones. Notice that since it is not easy to compare two solutions with different number of clusters, we chose to compute a four-cluster GMM solution with all covariates instead of finding its own optimal number of clusters.

Our results are presented as figures displaying the IAT trajectories, and as tables describing the characteristics of subjects classified into groups and giving the HMTD model parameters.

3 Results

We provide here the results of the various clustering performed using the HMTD and GMM approaches, and we compare the resulting classifications. Notice however that given the iterative nature of the optimization algorithms, it is never possible to be sure that the final models are the best possible ones. Therefore, results should never be overinterpreted.

3.1 HMTD Clustering

Without covariates, the best model identified by the BIC is a four-component model (model 1). Figure 1 shows the IAT trajectories in each group. We clearly differentiate a group with average volatility and IAT level (group 1), a group with relatively low scores and variability (group 2), a group with very low variability and a low and constantly diminishing IAT score (group 3), and a group with more complex trajectories and hence variability (group 4).

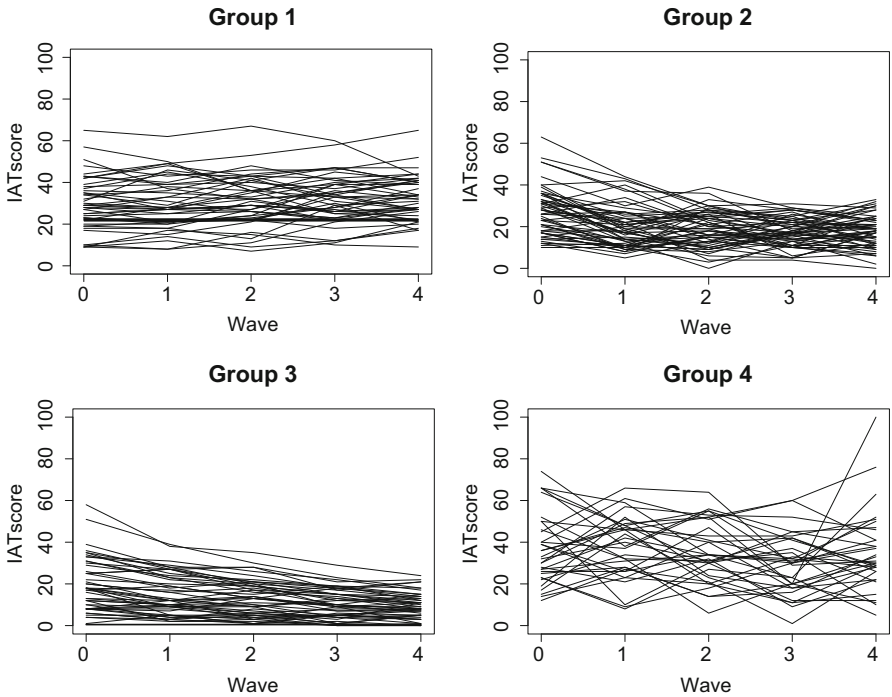

Fig. 1 IAT trajectories associated with each group in the four-group HMTD solution without covariates (model 1)

When we include the covariates in model 1 (Fig. 2) and relabel the four groups of the solution in order to match the groups of model 1, we obtain a similar four-group structure (model 2). As a comparison of the two figures might show, the most important difference is with the first two groups: group 2 of model 2 lost its higher-valued trajectories and focused more on a low IAT-level and stable trajectories. This change will be explored in more details later.

Table 1 provides the parameter estimation for both models. In addition to the point estimates, we also provide the 95% bootstrap confidence intervals.

As regards the first model without covariates, the θ_0 parameters giving the variance of each component of the model confirm the first impression given by Fig. 1: Group 4 is characterized by a much higher variability than the three other groups, and group 3 has the lowest variance, indicating less variation among the successive observations of a single individual. Parameters φ_0 corresponding to the constant in the modeling of the mean of each component also take expected values, with higher values associated with groups showing higher average IAT level. Finally, the autoregressive parameter φ_1 takes a value closer to one for the groups with trajectories showing smoother evolutions from one wave to the next, that is groups 1 and 3. All parameters of this first model are significant at the 95% level, as demonstrated by the confidence intervals.

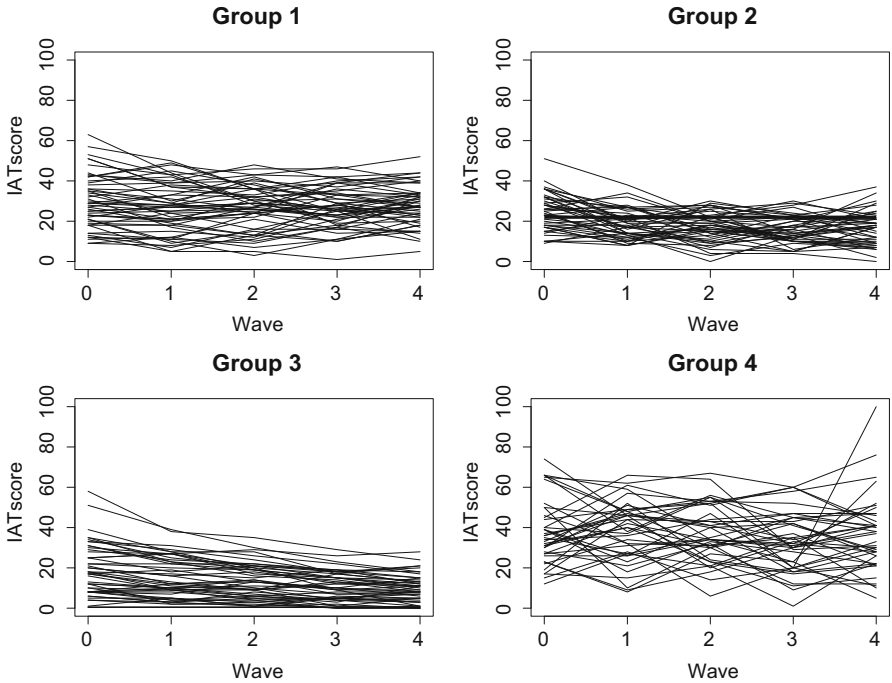

Fig. 2 IAT trajectories associated with each group in the four-group HMTD solution with five covariates (model 2)

Table 1 Estimated coefficients for the two HMTD models. For each parameter, we also provide the minimum and maximum values, and the 95% bootstrap confidence intervals obtained from 1000 bootstrap samples. Significant parameters at the 95% level are indicated with an asterisk

Model 2		θ_0	φ_0	φ_1	**WHO-5**	**BMI**	**Gender** (male)	**Age**	**Education** (extended req.)
			Model 1						
group 1	coefficient	41.098*	2.798*	0.891*					
$n = 46$	min, max	23.898, 108.133	1.881, 30.144	-0.070, 0.964					
	95% interval	(26.312; 38.317)	(3.417; 8.162)	(0.752; 0.919)					
group 2	coefficient	49.712*	7.228*	0.572*					
$n = 56$	min, max	0.001 89.683	5.721 22.126	-0.036 0.615					
	95% interval	(36.871; 54.309)	(6.507; 10.735)	(0.375; 0.563)					
group 3	coefficient	11.570*	1.072*	0.753*					
$n = 50$	min, max	7.535, 19.014	-0.102, 2.518	0.685, 0.814					
	95% interval	(8.715; 13.091)	(0.456; 1.737)	(0.715; 0.787)					
group 4	coefficient	186.582*	15.511*	0.514*					
$n = 33$	min, max	0.000, 384.934	6.951, 39.902	-0.186, 0.743					
	95% interval	(0.002; 291.178)	(14.906; 28.189)	(0.226; 0.560)					
					Model 2				
group 1	coefficient	38.587*	-1.148	0.656*	-2.156*	1.331*	-2.663*	0.941	-0.468
$n = 52$	min, max	24.369 39.084	-27.000 47.072	0.371 0.714	-3.503 -1.014	-0.079 3.269	-7.317 0.084	-2.232 2.883	-3.271 4.786
	95% interval	(27.009 37.035)	(-9.082 31.695)	(0.507 0.672)	(-2.962 -1.405)	(0.596 2.535)	(-5.433 -1.672)	(-1.116 1.675)	(-2.107 1.928)
group 2	coefficient	43.371*	-1.444	0.539*	-0.288	-1.025	0.549	0.638	-2.705
$n = 45$	min, max	0.530 86.137	-15.527 55.269	0.100 0.601	-2.788 1.896	-3.692 0.427	-5.127 4.760	-3.290 1.838	-8.183 1.675
	95% interval	(34.931 52.838)	(-1.607 18.077)	(0.217 0.540)	(-1.266 0.723)	(-1.950 0.073)	(-0.319 2.851)	(-0.545 0.888)	(-4.887 0.237)
group 3	coefficient	9.162*	0.256	0.723*	-0.707*	-1.424*	-0.757	0.142	1.060*
$n = 48$	min, max	5.987 11.771	-20.810 29.836	0.627 0.780	-1.955 0.161	-3.110 -0.036	-4.049 2.320	-1.995 1.870	-0.547 5.000
	95% interval	(7.222 10.013)	(-10.036 16.612)	(0.666 0.758)	(-1.487 -0.269)	(-2.069 -0.900)	(-2.151 0.458)	(-1.030 0.959)	(0.350 2.288)
group 4	coefficient	153.507*	22.274	0.307*	-2.771*	-4.014*	11.303*	-1.135	2.063
$n = 40$	min, max	99.659 242.775	-48.500 96.927	-0.099 0.518	-6.676 2.444	-9.000 -0.593	4.427 20.000	-6.000 4.656	-6.000 8.000
	95% interval	(108.832 198.571)	(-48.500 54.129)	(0.011 0.378)	(-5.327 -0.650)	(-7.542 -2.135)	(7.541 18.251)	(-3.493 4.095)	(-3.450 7.002)

As regards model 2, even if the first three parameters (θ_0, φ_0, and φ_1) take values different from those of model 1, θ_0 and φ_1 take values in the same range as of model 1. On the other hand, important differences are found for the constant parameter φ_0, and this parameter is no more significant in any group. Note that θ_0 and φ_1 tend to take smaller values in model 2. This can be interpreted as the first proof of interest of the covariates included in model 2: the groups are now more homogeneous (lower intra-group variance) and the explanation of a specific trajectory relies less on the immediately preceding observation. As regards the covariates, Age is never significant and could be eventually removed from the model. This could be due to the lack of a real age difference between participants (from 13 to 15 years old at baseline). However, the four other covariates remain significant for at least one of the groups.

When we consider each component of model 2 separately, the changes occurring in the trajectories associated with the first component are found related to the well-being of the concerned adolescents: a higher well-being such as measured by the WHO-5 index is significantly associated with a lower IAT-level. Males tend to have a lower IAT level than females, and a higher BMI is associated with higher IAT level. In group 3, a higher WHO-5 or BMI is associated with reduced IAT level, but being in the extended requirement school track is associated with a higher IAT level. Finally, in group 4, a higher WHO-5 or BMI is associated with reduced IAT level, and males tend to show a much higher IAT level than females.

Table 2 provides the main characteristics of the subjects classified into each group. For time-dependent variables, we considered the average value of each individual. A comparison is performed for each variable separately to test whether the groups are significantly different with regard to the variable. Considering only the two HMTD models, we observe that in addition to the expected differences in IAT level, the only other variable with significantly different values across groups is the WHO-5 measure of well-being. For both models, we observe two groups (2 and 3) with lower average IAT scores. The same two groups also display higher emotional well-being, as compared to the other groups, confirming previous results (Surís et al. 2014). No differences are observed for the other covariates, even if Gender comes close to significance in model 1. Even if not significant at the 95% level, probably because of the reduced sample size, we find a gender separation at the sample level; groups 2 and 4 contain a higher proportion of boys compared to the other two groups. The education track also shows a difference at the sample level: the first two groups contain more individuals following the highest education track as compared to groups 3 and 4. On the other hand, no notable difference is observed between the groups for Age and BMI, even if BMI, used as a covariate in model 2, is statistically significant in the modeling of the mean of each component.

Table 2 Characteristics of subjects classified into groups for different clustering. The p-value gives the result of the test comparing the different groups for each variable. The number of sequences classified into each group is provided in brackets after the group number

		IAT mean (sd)	**WHO-5** mean (sd)	**BMI** mean (sd)	**Gender** % male	**Age at T_0** mean (sd)	**Educ. at T_0** % extended req.
HMTD model 1							
Group 1	(46)	30.94 (11.7)	63.43 (15.6)	19.97 (2.35)	24	14.13 (0.50)	80.5
Group 2	(56)	20.29 (9.78)	71.01 (15.6)	20.02 (3.30)	45	14.05 (0.59)	67.9
Group 3	(50)	13.31 (9.88)	72.28 (13.6)	20.45 (2.57)	24	14.14 (0.67)	64.0
Group 4	(33)	34.69 (16.1)	63.49 (16.8)	20.06 (3.03)	39	14.27 (0.45)	60.6
p		<0.001	<0.001	0.764	0.055	0.381	0.214
HMTD model 2							
Group 1	(52)	27.43 (11.3)	67.35 (16.6)	20.12 (2.40)	31	14.19 (0.60)	71.2
Group 2	(45)	18.57 (8.41)	70.85 (15.2)	19.96 (3.53)	40	14.02 (0.45)	73.3
Group 3	(48)	13.62 (9.97)	70.64 (14.0)	20.46 (2.54)	21	14.10 (0.69)	64.6
Group 4	(40)	36.36 (15.6)	63.06 (16.4)	19.96 (2.86)	43	14.22 (0.48)	65.0
p		<0.001	0.015	0.741	0.113	0.331	0.746
GMM 2							
Group 1	(169)	22.08 (13.2)	68.79 (15.8)	20.20 (2.90)	32	14.15 (0.57)	0.68
Group 2	(16)	39.90 (14.8)	61.13 (13.9)	19.40 (2.12)	43	14.00 (0.52)	0.75
p		<0.001	0.022	0.210	0.496	0.322	0.771
GMM 4							
Group 1	(76)	13.35 (8.97)	73.32 (14.2)	20.69 (2.75)	32	14.09 (0.61)	0.63
Group 2	(31)	38.98 (11.2)	58.48 (16.2)	20.15 (2.40)	29	14.16 (0.52)	0.74
Group 3	(75)	26.46 (10.2)	67.09 (15.4)	19.62 (2.98)	33	14.17 (0.55)	0.73
Group 4	(3)	54.06 (18,3)	62.40 (9.66)	18.78 (3.30)	100	14 (0)	2/3
p		<0.001	<0.001	0.043	0.094	0.802	0.593
GMM 4 cov							
Group 1	(98)	18.79 (10.6)	69.64 (14.6)	20.06 (2.90)	29	13.91 (0.32)	77.9
Group 2	(44)	18.58 (10.5)	68.88 (17.8)	20.85 (2.95)	24	15.16 (0.55)	44.0
Group 3	(28)	39.38 (12.9)	64.95 (18.2)	20.14 (2.73)	48	14.24 (0.44)	48.3
Group 4	(15)	41.98 (14.8)	60.00 (13.7)	19.36 (2.05)	54	14.00 (0.41)	76.9
p		<0.001	0.032	0.321	0.058	<0.001	<0.001

3.2 Usefulness of the Covariates

From the results of the previous section, we find that the inclusion of covariates in the first classification obtained with the HMTD model helped us better differentiate the four groups, but without entirely changing their interpretation. We would like to better understand the changes in trajectory classification that occurred between these two models. Table 3 indicates how many subjects changed groups between the initial model without covariates and model 2 with covariates. As noted earlier, most of these changes occurred between groups 1 and 2. In particular, 19 second-group subjects of model 1 were transferred to the first group in model 2, and the steady

Table 3 Number of IAT trajectories associated with each group in HMTD models 1 (without covariates, rows) and 2 (including covariates, columns)

	Model 2			
Model 1	Group 1	Group 2	Group 3	Group 4
Group 1	31	6	2	7
Group 2	19	34	1	2
Group 3	2	3	45	0
Group 4	0	2	0	31

Table 4 The characteristics of 19 subjects moving from group 2 to group 1 (group 2→1) as compared to subjects staying in the same group (either 1 or 2) in both HMTD classifications. The means (numerical variables) or proportions (categorical variables) are provided, and differences with the subjects remaining in the same group (either 1 or 2) are assessed using t-tests and χ^2-tests with continuity correction: ns: non-significant

	IAT	**WHO-5**	**BMI**	**Sex** (% male)	**Age**	**Education**
Group 2→1	22.76	72.93	19.69	57.9	14.26	52.6
vs group 1	31.26**	62.77**	20.41 ns	9.70***	14.16 ns	80.6 ns
vs group 2	18.72 ns	71.42 ns	20.30 ns	38.2 ns	13.97 ns	73.5 ns

*p <0.05; **p < 0.01; ***p < 0.001

low Internet addiction profile of the second group became even more pronounced, with the higher Internet addiction subjects joining the first group. However, since some trajectories simultaneously left group 1 for the three other groups, the average IAT level of group 1 also decreased. Overall, the inclusion of covariates appears beneficial for the differentiation of trajectory features among groups.

The 19 individuals who switched from group 2 to group 1 represent the main difference between the two models, with all the other changes concerning at the most seven subjects. Thus, it is interesting to explore how these individuals differed from those who remained in the first or second group in both classifications. Table 4 summarizes our findings using t-tests and χ^2-tests to compare the different variables. The average IAT scores are quite different between the three considered sub-groups, and, as expected, the "moving" sub-group shows an Internet dependence level between the two "stable" sub-groups. Thus, the moving individuals were among the most Internet-problematic members of the full second group of model 1, and even if the average IAT score is not the only indicator of group affiliation, a visualization of the trajectories would confirm the ambiguous nature of these individuals. The moving subgroup is also significantly different from the group of individual staying in group 1 as regards the WHO-5 index of well-being and the gender ratio, with a higher well-being and higher proportion of males among the moving subgroup. No other significant differences are observed.

3.3 GMM Clustering

Without covariates, the best GMM solution in terms of BIC is a two-group solution, but given the high difference in number of trajectories associated to each group (169 vs 16), this solution in not really interpretable and hence less useful than the four-group solution given by the HMTD approach. Therefore, we also estimated a four-group GMM (Fig. 3).

In the two-group solution, a large majority of trajectories are associated with group 1, and only 16 sequences are associated with group 2. The average IAT level is higher in group 2, but both groups exhibit an important variability, as indicated in Table 2. Moreover, in terms of interpretation, one can only say that IAT sequences with a clear increasing trend are separated from the other sequences. In the four-group solution, even if the number of groups is the same as in the HMTD models, there is no a priori correspondence between the HMTD and GMM groups. In the four-group GMM solution, the number of subjects per group shows much more variability than that observed with the HMTD group, with the majority of individuals classified in groups 1 or 3, and only three subjects in group 4.

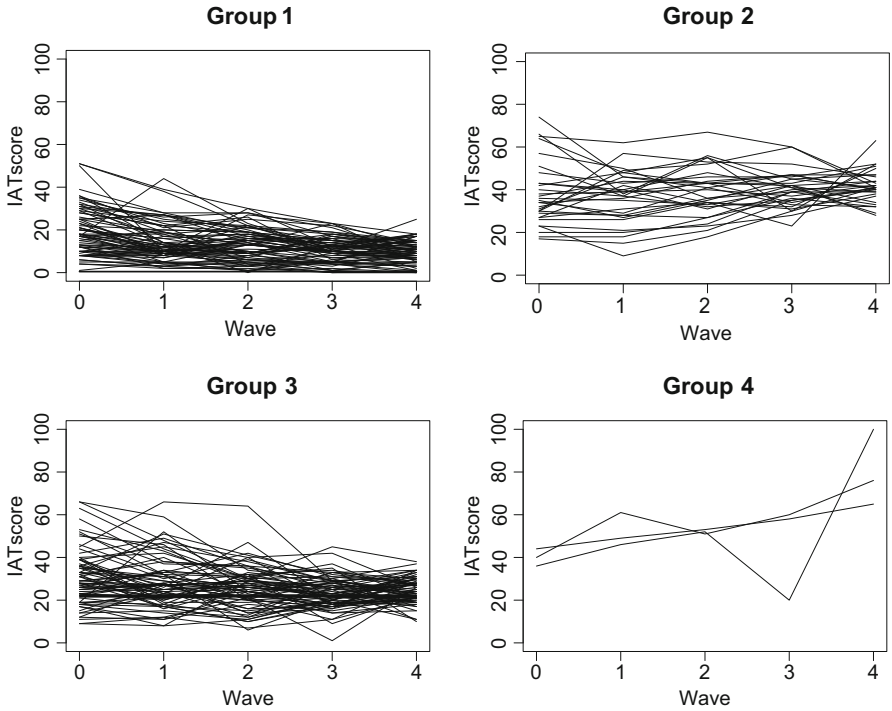

Fig. 3 IAT sequences associated with each group in the four-group GMM solution without covariates

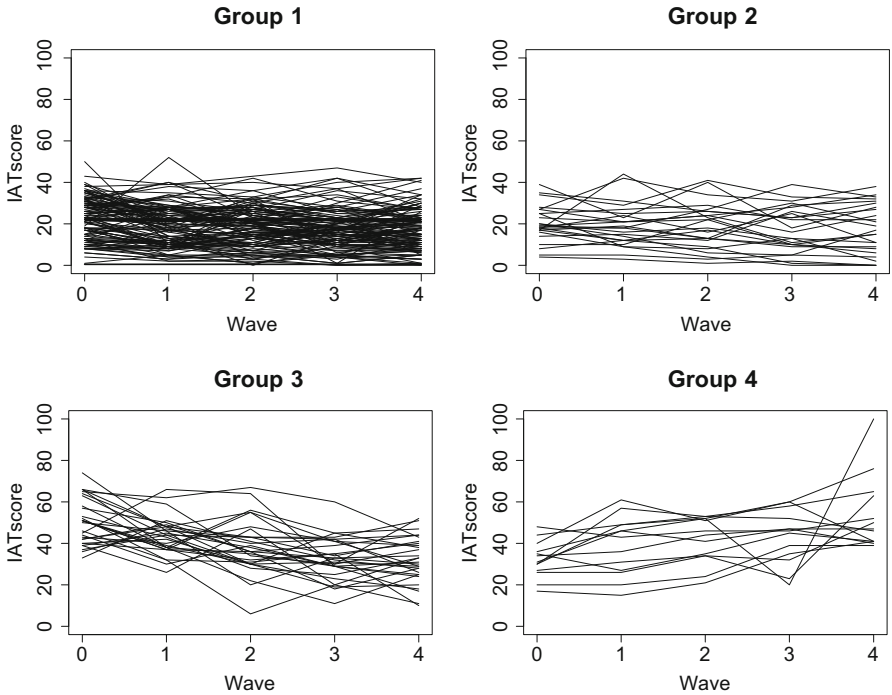

Fig. 4 IAT sequences associated with each group in the four-group GMM solution with four covariates

Finally, as with the HMTD approach, we enhanced the four-group solution by adding covariates. Four of the five covariates used in the HMTD approach appeared useful in the GMM solution as well. Figure 4 displays the resulting groups obtained after adding Gender and Education as predictors for group membership (multinomial regression on c_i), and WHO5 and BMI as fixed effect. On the other hand, Age was not included in the final model because the estimation process would then lead to a one-group solution. Another important issue with the GMM approach is the results' sensitivity to the order in which the covariates are included in the model. Various covariate combinations were tested before we chose the above-mentioned combination as the best one in terms of clustering results. For instance, *classmb = gender + education sector* does not give the same results as *classmb = education sector + gender*. This surprising result may be due to a bug in the *lcmm* R package, but in our opinion the reason could rather be the optimization procedure. It is well known that EM-type algorithms converge to the nearest local optimum, and that this optimum is not always the global one. Therefore, the solution depends on the initial values of the parameters, especially when the solution space is complex, which is the case here.

As Fig. 4 shows, the number of trajectories associated with each group is quite variable, with the large majority assigned to group 1. The first two groups are

characterized by low variability and an overall low IAT level. The trajectories in these two groups seem very similar, but since this four-group solution might be suboptimal and is computed only for the purpose of comparison with HMTD models 1 and 2, a three-group solution could merge these two groups into one group. The last two groups have a higher average IAT level, both exhibiting a general linear trend over time, decreasing in group 3 and increasing in group 4.

Table 2 gives the characteristics of individuals classified in each group of the GMM models and compares the groups for each variable. Note that given the large differences in group size, the test results for the GMM models should be interpreted with caution. As observed earlier in the HMTD case, significant differences exist between groups for both the IAT and WHO-5 variables. A significant difference exists also for BMI in the four-group GMM model without covariates. More interestingly, the Age and Education track at baseline also show significantly different values across groups, with one of the variables (Education track) being included in the model as covariable, but not the other. This difference between the HMTD and GMM clustering points to the fact that the solutions provided by both approaches are not identical or interchangeable, and that the two models used information in a different manner to provide usable data sequence clusterings.

4 Comparison of HMTD and GMM

When used for clustering purposes, the HMTD and GMM models share some characteristics: They both represent a kind of mixture model, they can include covariates of any type at the visible level, and they can also include covariates at the latent level and use them to estimate the initial probability of each cluster. However, HMTD and GMM also have several differences. First of all, since GMM is a mixture of mixed models, it is able to accept both fixed and random effects. Another difference is the possibility of HMTD to include an autoregressive specification for the variance and thus to allow for the clustering of longitudinal sequences whose variance evolves in time. For instance, sequences becoming more instable over time can more easily be grouped together. However, to exploit this feature, it is necessary to work with long data sequences, what was not the case here with the IAT example.

Another feature of HMTD that is worth stressing is the possibility of using it to perform different kind of clustering (Bolano and Berchtold 2016). The transition between components is driven by the hidden transition matrix A. In this paper, A was constrained to be a diagonal identity matrix, implying that each sequence was assigned to one and only one group, and all sequences assigned to the same group were described by the same visible model. However, there are several alternatives. For instance, different latent states may be required to alternate over time in order to find the optimal modeling of a given sequence. If A is constrained to have the following structure:

$$A = \begin{pmatrix} a_1 & 1-a_1 & 0 & 0 \\ a_2 & 1-a_2 & 0 & 0 \\ 0 & 0 & a_3 & 1-a_3 \\ 0 & 0 & a_4 & 1-a_4 \end{pmatrix}$$

where a_1, a_2, a_3 and a_4 are transition probabilities, then one performs at the same time a modeling and a clustering of the data sequences. The first two states are used to model the first cluster, and states 3 and 4 are used to model the second cluster. In other words, data sequences are clustered into two groups, but inside each group there are two different visible models allowing for a better representation of these sequences when their behavior evolves over time.

Another specification of A would allow some sequences to remain always in the same cluster, whereas other ones could transit at some point in time from the first to the second cluster:

$$A = \begin{pmatrix} a_1 & 1-a_1 & 0 & 0 \\ a_{21} & a_{22} & 1-a_{21}-a_{22} & 0 \\ 0 & 0 & a_3 & 1-a_3 \\ 0 & 0 & a_4 & 1-a_4 \end{pmatrix}$$

5 Conclusion

Hidden Markovian models are known to be valuable tools to analyze the dynamics in longitudinal continuous data and in life course data (e.g. Helske et al. 2018). The present study demonstrates that the sequences of continuous longitudinal data can also be classified into as many groups as required, and that the HMTD model can be used as a valid alternative to GMM. The inclusion of covariates has beneficial effects on clustering, because the resulting groups have lower intra-variability compared to the solution without covariates.

In a comparative study involving the use of GMM for clustering, our first finding is that the HMTD approach is a good alternative to GMM, because in terms of interpretability its results are at least as interesting as the results given by GMM. However, on the basis of just one practical example, we obviously cannot conclude that one approach is better than the other; moreover, this is not the purpose of this study. What we can conclude is that the HMTD approach is not only theoretically, but also practically useful to classify sequences of continuous data in mutually excluding groups.

In the literature, excessive Internet use has been found to be highly related to several somatic conditions, sleep disturbance in particular. However, in this paper, our main objective is not to explain IAT trajectories, but to find ways to classify such trajectories into meaningful groups. Moreover, there is still an ongoing debate on the direction of the relationship between Internet use and sleep disturbance, not

to speak of causality. Therefore, we chose not to consider sleep disturbance in this analysis, but to concentrate on other covariates that are more neutral to IAT scores. Nevertheless, even with this restriction, the results obtained with the HMTD model are highly significant and allow for a sound interpretation. The four resulting groups differ in terms of average value and variability. The relationship observed between IAT and the emotional well-being of subjects suggests that both concepts are linked and that a higher risk of Internet addiction is related to poorer well-being. Gender is also a discriminating factor between groups, with a lower proportion of males in the first and third groups, but, given the small sample size, the differences are not significant at the population level.

The main strength of this study is the demonstration of the usefulness of the HMTD approach as a valuable alternative to the GMM approach for clustering continuous data sequences. Researchers would be advised to consider both approaches to take full advantage of the information in their data. However, some weaknesses of this study are to be mentioned. At the theoretical level, we include covariates in the HMTD model only at the visible level, but it is also possible to include them at the latent level as well in order to enhance the prior probabilities of each cluster. As regards the application of the model to IAT trajectories, we used a rather small convenience sample; this is not representative of the population of adolescents living in the canton of Vaud. More analyses need to be conducted with larger databases to define a real typology of IAT trajectories.

Overall, in spite of some shortcomings, the HMTD model can be considered as a complete framework for the analysis of continuous data sequences. It is an explanatory tool as well as a clustering tool, and by adding covariates, constraints on the transition matrix, and autoregressive modeling of the mean and variance of each component, the model goes well beyond the traditional Markovian models such as homogeneous Markov chains or hidden Markov models.

Acknowledgements 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 (grant number: 51NF40-160590). The ado@internet.ch study was financed by the Public Health Department of the Vaud canton and by the Swiss National Science Foundation (grant number: 105319_140354). The authors are grateful to both institutions for their financial support. The funding bodies had no role in the design and conduct of the study; collection, analysis and interpretation of data; or preparation, review, or approval of the manuscript.

References

Barrense-Dias, Y., Berchtold, A., Akré, C., & Surís, J. C. (2015). The relation between internet use and overweight among adolescents: A longitudinal study in Switzerland. *International Journal of Obesity, 40*, 45–50.

Bauer, D., & Curran, P. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*(3), 338–363.

Berchtold, A. (2001). Estimation in the mixture transition distribution model. *Journal of Time Series Analysis, 22*, 379–397.

Berchtold, A. (2003). Mixture transition distribution (MTD) modeling of heteroscedastic time series. *Computational Statistics and Data Analysis, 41*, 399–411.

Berchtold, A., & Raftery, A. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science, 17*, 328–356.

Bolano, D., & Berchtold, A. (2016). General framework and model building in the class of hidden mixture transition distribution models. *Computational Statistics and Data Analysis, 93*, 131–145.

Faraci, P., Craparo, G., Messina, R., & Severino, S. (2013). Internet addiction test (IAT): Which is the best factorial solution? *Journal of Medical Internet Research, 15*(10), e225.

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE, 61*, 268–278.

Francis, B., & Liu, J. (2015). Modelling escalation in crime seriousness: A latent variable approach. *Metron, 73*(2), 277–297.

Helske, S., Helske, J., & Eerola, M. (2018). Analysing complex life sequence data with hidden Markov modelling. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302–317.

Khazaal, Y., Billieux, J., Thorens, G., Khan, R., Scarlatti, E., Theintz, F., Lederrey, J., Van Der Linden, M., & Zullino, D. (2008). French validation of the internet addiction test. *Cyberpsychology Behavior, 11*(6), 703–706.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*(1), 110–133.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahawa: LEA.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*(2), 463–469.

Nagin, D. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods, 4*(2), 139–157.

Piguet, C., Berchtold, A., Zimmermann, G., & Surís, J. C. (2016). Rapport final de l'étude longitudinale AdoInternet.ch. Lausanne: Raisons de santé.

Proust-Lima, C., Philipps, V., & Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *Journal of Statistical Software, 78*(2), 1–56.

Raftery, A. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society, Series B, 47*(3), 528–539.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Ram, N., & Grimm, K. J. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33*(6), 565–576.

Reinecke, J., & Seddig, D. (2011). Growth mixture models in longitudinal research. *AStA Advances in Statistical Analysis, 95*(4), 415–434.

Skarupova, K., Olafsson, K., & Blinka, L. (2015). Excessive internet use and its association with negative experiences: Quasi-validation of a short scale in 25 European countries. *Computers in Human Behavior, 53*, 118–123.

Surís, J. C., Akré, C., Berchtold, A., Fleury-Schubert, A., Michaud, P. A., & Zimmermann, G. (2012). Ado@Internet.ch: Usage d'internet chez les adolescents vaudois. Raisons de santé 208. Lausanne: Institut universitaire de médecine sociale et préventive.

Surís, J. C., Akré, C., Piguet, C., Ambresin, A. E., Zimmermann, G., & Berchtold, A. (2014). Is internet use unhealthy? A cross-sectional study of adolescent internet overuse. *Swiss Med Wkly, 144*, w14061.

Taushanov, Z., & Berchtold, A. (2017). A direct local search method and its application to a Markovian model. *Statistics, Optimization and Information Computing, 5*(1), 19–34.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association, 91*(433), 217–221.

Wang, M., & Bodner, T. E. (2007). Growth mixture modeling identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods, 10*(4), 635–656.

Young, K. S. (1998). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior, 1*(3), 237–244.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Divisive Property-Based and Fuzzy Clustering for Sequence Analysis

Matthias Studer

1 Introduction

In this paper, we introduce property-based and *fuzzy* clustering and discuss their usefulness in the context of sequence analysis. We also present some tools available in R code by which to conduct these analyses. These two clustering methods aim to overcome some of the limitations of the more "traditional" ones, such as partitioning around medoids or agglomerative clustering.

Most of the clustering methods used in sequence analysis are polythetic, with the notable exception of model-based clustering. This means that cluster memberships are defined according to a broad set of properties and by comparing a given sequence to all the other ones. For this reason, the rules that define which sequences belong to which cluster are implicit. In other words, we do not know exactly on which grounds a sequence is assigned to a given cluster. Having a typology based on implicit rules of cluster membership has two disadvantages.

First, the resulting clustering is sample-dependent, which means it cannot be compared to another typology that is created in a subsample or another sample, for instance. Even comparing two typologies that appear to be similar might be problematic, since their underlying implicit clustering rules might differ.[1] On the other hand, having explicit rules would allow one to validate a typology in other

[1] We cannot be sure that the clustering rules are sufficiently similar, because we do not know them. They are only implicit.

M. Studer (✉)
NCCR LIVES and Geneva School of Social Sciences, University of Geneva, Geneva, Switzerland
e-mail: matthias.studer@unige.ch

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_13

samples and reproduce a given (validated) typology in other studies. Implicit rules therefore hinder the reproducibility of life-course research, as well as the possibility of undertaking a large-scale literature review.

Second, implicit rules make the interpretation of clustering more difficult. The first step after creating a typology is usually to try to interpret and recover these implicit rules by using different kinds of graphics and analyses. Explicit clustering rules would make the interpretation of clustering results much easier.

Monothetic clustering—which here we call "property-based clustering"—aims to create a sequence typology defined by explicit classification rules. In this paper, we introduce a method that is based on the "DIVCLUS-T" algorithm proposed by Chavent et al. (2007). Following the work of Piccarreta and Billari (2007), we also discuss its use in sequence analysis and extend their work by proposing for consideration new sets of state sequence features. Finally, we make the analytical results broadly available in the `WeightedCluster` package.

In this paper, we also discuss the use of *fuzzy* clustering, which aims to overcome another limitation of "traditional clustering." In sequence analysis, we usually use *crisp* clustering (i.e., each sequence is assigned to only one sequence type). In *fuzzy* clustering, however, each sequence belongs to one or more clusters, with a certain degree or strength (D'Urso 2016).

The *fuzzy* approach has several advantages over the more usual *crisp* one (D'Urso 2016). First, sometimes some sequences are between two sequence types. In *crisp* clustering, these sequences would be assigned to one of the two types; in *fuzzy* clustering, however, these sequences would be considered a *hybrid-type* or a *mixture* of the two types (D'Urso 2016). From a statistical viewpoint, *fuzzy* clustering might lead to better results if some sequences are between two (or more) sequence types. This case might occur frequently, according to Warren et al. (2015), who argue that exact cluster membership should not be trusted. From a sociological perspective, this approach is of special interest when the trajectories are not strongly structured into types, and when we can think that some individuals can be influenced by several sequence types.

Second, in *fuzzy* clustering, membership is thought to be *gradual*. Some sequences are more central (typical) of a given type than are others. This is also an interesting property from a sociological viewpoint. From the Weberian *ideal–typical* perspective, nobody is the perfect incarnation of an ideal type, but some are closer to it than others. This sociological perspective is similar to the gradual membership approach inherent in *fuzzy* clustering.

For all these reasons, the use of the *fuzzy* clustering approach is promising in life-course research and sequence analysis. However, to the best of our knowledge, it has been only seldom used in sequence analysis. One of the likely reasons for this paucity is the lack of proper tools by which to analyze sequences in conjunction with a membership matrix, instead of a categorical covariate (as in *crisp* clustering). For instance, Salem et al. (2016) used *fuzzy* clustering, but they ultimately assigned each sequence to the cluster with the highest membership in all subsequent analyses, thus turning in fact back to *crisp* clustering. In this paper, we propose different tools to fill this gap and make use of the full information of the membership matrix.

This paper is organized as follows. It starts by presenting property-based clustering and the set of sequence properties whose consideration we propose. We then turn to the presentation of *fuzzy* clustering and introduce several ways of representing the results, before interpreting them. We also discuss how to properly analyze cluster membership according to some explanatory covariates. Before concluding, we briefly show how to run the proposed analysis in `R`, using our `WeightedCluster` library.

2 Sample Issue

The usefulness of the proposed methodology is illustrated by using the data and the research question from McVicar and Anyadike-Danes (2002), who studied school-to-work transition in Northern Ireland. Their analysis was undertaken in two steps. They started by identifying ideal-typical trajectories, before explaining clustering membership by using information such as qualification at the end of compulsory schooling, family background, and demographic characteristics. Their aim was to "identify the 'at-risk' young people at age 16 years and to characterize their post-school career trajectories" (p. 317). To build our clustering, we use optimal matching with constant cost.

3 Property-Based Clustering

The aim in using property-based clustering is to build a sequence typology by identifying well-defined clustering rules that are based on the most relevant properties of the analyzed object. In the literature, these clustering methods are called *monothetic divisive* clustering methods (Chavent et al. 2007), and they were first introduced in sequence analysis by Piccarreta and Billari (2007). We propose here a conceptual presentation of the "DIVCLUS-T" algorithm (a detailed presentation can be found in Chavent et al. 2007). The "DIVCLUS-T" algorithm is very similar to one proposed by Piccarreta and Billari (2007). Our choice between the two is based on availability in `R`. Here, we mainly extend the work of Piccarreta and Billari (2007), by proposing new sets of sequence features for consideration.

3.1 Principle

Property-based clustering uses two kinds of information. First, the sequence properties are used to build the rules. Second, it uses a dissimilarity matrix, which is used to measure variation in the sequence and how much of this variation can be explained by a given property; this is in line with the discrepancy analysis framework (Studer et al. 2011).

The method then works in two steps: tree building and splits ordering. In the tree building phase, all sequences are grouped in an initial node. Then, this node is split according to one of the object properties, into two subnodes or clusters. This property and the associated split are chosen in such a way that the split "explains" the biggest share of the sequence discrepancy (Studer et al. 2011; Chavent et al. 2007; Piccarreta and Billari 2007). The process is then repeated on each new node until a stopping criterion is found. First, the algorithm might stop because there are no further relevant properties by which to make a split. Second, nodes with only one observation are obviously not split. This first step is roughly equivalent to the procedure that Piccarreta and Billari (2007) propose.

As in Piccarreta and Billari (2007), our implementation of the "DIVCLUS-T" algorithm also makes it possible to specify a minimum number of observations per node. Splits that would lead to at least one node with fewer than this minimum number of observations are discarded. This is a useful extension, if we want to ensure that all clusters represent at least a given percentage (for instance, 5%) of the sequences. One might also restrict to "significant" splits by using permutation tests, as in discrepancy analysis (Studer et al. 2011). However, the usefulness of the latter approach is subject to discussion, as the concept of significance is not very well defined in cluster analysis. This first step of the procedure can be seen as a decision tree, where the splits are chosen according to the explanatory power of the considered properties. In fact, our implementation is based on the tree-structured discrepancy analysis.

Once the whole tree is built, the splits are ordered according to their overall "relevance." More precisely, this "relevance" is measured by calculating the increase in the share of the overall discrepancy that is explained by adding a split. This procedure has the advantage of maximizing a global criterion. Ultimately, the result of this procedure is a series of nested partitions ranging from one group to a number of groups, any of which depend on the stopping criteria or a maximal number of groups to consider.[2] This second step is the major difference from the procedure proposed by Piccarreta and Billari (2007), where the final clustering and the stopping criteria depend on a pruning procedure.

Figure 1 graphically represents the procedure, using our illustrative example of school-to-work transition in Northern Ireland, for the first nine splits. The order of the splits is presented on the right, with the associated number of clusters. We start at the top of the tree with a single cluster. At this stage, the most relevant feature in splitting this top node into two is to have spent more (or less) than 17 months in higher education (property "duration.HE"). Stopping at this level would result in a cluster in two groups (presented on the right of Fig. 1). The clustering in three groups is obtained by splitting the node "less than or equal to 17 months in higher education" in two groups: having spent more (or less) than 33 months in employment (criterion "duration.EM"). This last split was preferred to the solution of splitting the node "more than 17 months in higher education," because it has greater explanatory power regarding sequence variation at a global level. The procedure then continues until some stopping criteria are met.

[2]In other words, new groups are added by dividing one of the existing groups into two subgroups.

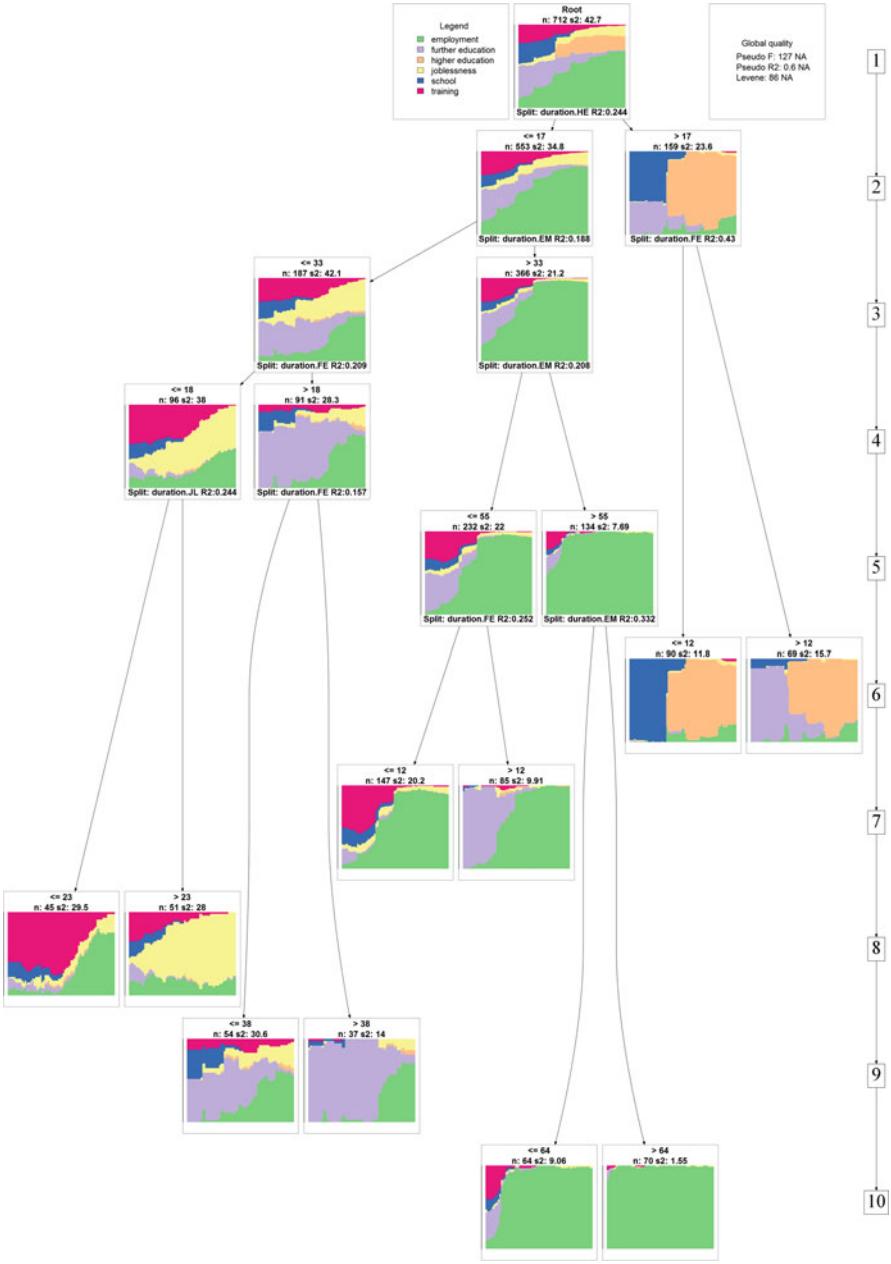

Fig. 1 Clustering tree

As with any agglomerative or divisive clustering, the choice of the number of groups can be grounded on cluster quality measures (see Studer 2013, for a review). Here, various clustering quality measures agree on choosing either the solutions in four ($ASW = 0.41$) or nine ($ASW = 0.38$) groups as the best clustering. According to the empirical evaluation of Chavent and Lechevallier (2006), for a small number of groups, the "DIVCLUS-T" algorithm tends to produce better clustering (measured on the basis of statistical criteria) than Ward clustering. However, the reverse becomes true as the number of groups increases.

3.2 Property Extraction

The results of property-based clustering are highly dependent on the properties of the object to be clustered. In the empirical evaluation made by Chavent and Lechevallier (2006), having a large number of meaningful properties was one of the key elements that led to good-quality clustering. In this section, we propose a set of properties worthy of consideration. In our implementation of the algorithm, these properties are automatically extracted.

Within the life course paradigm, three main dimensions of the trajectories are of central interest: the timing, the duration, and the sequencing of the states (Studer and Ritschard 2016). We propose the automatic extraction of various properties that measure each of these dimensions.

To measure the timing of the state, we consider the state at each time position t. If we consider the sequence of length ℓ, we therefore end with ℓ categorical covariates that measure the situation over time. Piccarreta and Billari (2007) propose another way of measuring the timing, by considering the spells that form the sequences. They generate one property $A_{s,k}$ that stores the age at the beginning of the kth spell in state s. If the property is not observed (i.e., there is no kth spell in state s), they propose setting it to $\ell + 1$, where ℓ is the length of the sequence. However, we take here a slightly different strategy: we set it to a missing value. Missing values are then treated as a special case when defining a split. Although we use the numeric information, when available, to define the split by using an inequality, the missing values are attributed to one of the two groups—whichever gives the best result.

We propose the use of two sets of properties to measure duration. First, we consider the duration in the successive spells. This is achieved by generating one property $D_{s,k}$ that stores the duration of the kth spell in state s. Second, we consider the overall time spent in each state that results in one property per state.

We use frequent subsequence mining to extract the properties that measure the state sequencing. With this method, the aim is to uncover frequent subsequences within a set of sequences (Studer et al. 2010; Agrawal and Srikant 1995; Zaki 2001). A subsequence s is defined as a subsequence of x if all the states of s occur in the same order as in x. For instance, the sequence $A - C$ is a subsequence of sequence $A - B - C$ because A and C occur in the same order. A subsequence is said to be frequent if it is found in more than a predefined percentage of sequences. Using

this framework, several sets of sequence properties can be extracted. First, we look for frequent (in our case, 1%) subsequences in the sequence of distinct successive states (DSS). This step generates one property (i.e., variable) per identified frequent subsequence, and stores the number of times that the subsequence is found in each sequence. For instance, the subsequence $A-C$ occurs twice in the sequence $A-B-C-B-C$. We also consider the age at the first occurrence of the pattern (i.e., when the pattern starts). Second, we look for frequent subsequences within the transition sequences. This is achieved by specifying a distinct state for each transition. For instance, the DSS $A-B-C$ will be recoded as $A-AB-BC$ before running the analysis. Here again, the number of occurrences and the age at the first occurrence are stored as properties.

Finally, the user can consider and add other properties. The algorithm computes different sequence complexity measures. Piccarreta and Billari (2007) suggest adding information about covariates such as education level. Application-specific sequence properties could also be of interest. For instance, one might have an interest in adding the time spent in a state of joblessness within the last 12 months of our sequences, if professional integration at the end of the sequence is of primary concern.

Although all these properties are automatically extracted, they need to be carefully chosen according to the issue under investigation. For instance, for a study that mainly concerns itself with timing differences, we suggest restricting attention to timing-related properties. For this reason, in our implementation of the algorithm, one can specify the sets of properties to be considered. Table 1 summarizes the properties considered in this study. The first column contains the name of the property used in our `R` implementation, and the second provides brief descriptions of the sets of properties.

Table 1 Sequence properties considered in the clustering algorithms

Name	Description
`state`	The state in which an individual is found, at each time position t
`spell.age`	The age at the beginning of each spell of a given type
`spell.dur`	The duration of each of the spells presented above
`duration`	The total time spent in each state
`pattern`	Count of the frequent subsequences of states in the DSS
`AFpattern`	Age at the first occurrence of the above frequent subsequence
`transition`	Count of the frequent subsequence of events in each sequence, where each transition is considered another event
`AFtransition`	Age at the first occurrence of the above frequent subsequence
`Complexity`	Complexity index, number of transitions, turbulence

3.3 Running the Analysis in R

Property-based clustering can be run using the `seqpropclust` function available in the `WeightedCluster` package. Aside from the state sequence object `myseq`, one needs to specify the distance matrix (argument `diss`), the properties (by default, all properties are computed), and the maximum number of clusters under consideration.

```
R>  ## Clustering using properties "state" and "duration"
R>  pclust <- seqpropclust(myseq, diss=diss, maxcluster=5,
       properties=c("state", "duration"))
R>  ## Displaying the resulting clustering
R>  seqtreedisplay(pclust, type="d", border=NA,  showdepth=TRUE)
R>  ## Computing clustering quality and cluster membership
R>  pclustqual <- as.clustrange(pclust, diss=diss, ncluster=5)
```

The clustering membership can be extracted by using the `as.clustrange` function, which also computes various cluster quality measures. See Studer (2013) for a detailed presentation of this procedure.

4 Fuzzy Clustering

In *crisp* clustering, each sequence is assigned to exactly one sequence type. The result is a categorical covariate that summarizes the typology. In *fuzzy* clustering, each sequence can belong to more than one cluster; this is achieved by computing the degree or strength of membership of each sequence to each identified sequence type (D'Urso 2016). This is of central interest when the sequences are not thought to be strongly structured, or when some sequences could have been influenced by more than one type.

More precisely, the result of *fuzzy* clustering is a membership matrix $\mathbf{u}$ comprising one row per individual and one column per cluster. Each value u_{iv} of this matrix measures the membership strength of an individual i to each cluster v. These membership degrees, which usually sum to 1, are also called "probabilities," and they lead to two slightly different interpretations. The concept of membership "strength" or "degree" refers to the closeness of each sequence to each type. The notion of "membership probabilities" refers to the chances that the underlying sequence was generated according to one of the types.

In this section, we start by presenting the algorithm used herein. We then propose different approaches to describing and visualizing the results of *fuzzy* clustering, using the membership matrix. Finally, we discuss possible strategies by which to analyze how explanatory covariates are linked to cluster membership; again, this is based on the membership matrix. We hope that the availability of these tools will lead to more widespread use of *fuzzy* clustering in sequence analysis.

4.1 Fanny Algorithm

We use the *Fanny* algorithm proposed by Kaufman and Rousseeuw (1990) and later adapted by Maechler et al. (2005). This algorithm aims to minimize the following function:

$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{iv}^{r} u_{jv}^{r} d(i, j)}{2 \sum_{j=1}^{n} u_{jv}^{r}} \quad , \tag{1}$$

where n is the number of observations, k a predefined number of groups, u_{iv} the membership value of individual i to cluster v, and $d(i, j)$ the distance between sequences i and j. The exponent r is a *fuzziness* parameter that needs to be set by the user. A value of 2 is often used, but values between 1.5 and 2.5 are usually recommended (D'Urso 2016). The standard procedure is to start with $r = 2$ and use a smaller value if the algorithm does not converge.

Using our sample data, we used a value of 1.5 for the fuzziness parameter. We kept a solution in seven groups, based on the interpretability of the resulting clustering and the aim of the study.

4.2 Plotting and Describing a Fuzzy Typology

Once the clustering has been computed, typically, the first step is to describe each cluster and give it a first interpretation. In this section, we propose several approaches to doing so by using the membership matrix.

4.2.1 Most Typical Members

A first way to label the clusters and interpret them is to identify typical sequences. In "traditional sequence analysis clustering," this can be done by identifying the medoid or a representative sequence based on other criteria, such as neighborhood density (Gabadinho et al. 2011). A natural way to do it with *fuzzy* clustering is to choose the sequence with the highest membership strength in each cluster. The first row of Table 2 presents this information. Using this strategy, we can have a first look at our clustering membership matrix. We found a first cluster related to full employment, then three patterns of education (i.e., training or further education) followed by employment: two patterns leading to higher education, and one pattern of training followed by joblessness.

Table 2 also presents descriptive statistics of membership strength for each cluster. As one will recall, in *fuzzy* clustering, each sequence has a measure of its membership strength in each cluster. Hence, it is not possible to compute a percentage of sequences belonging to each cluster, as we would do for *crisp*

Table 2 Descriptive statistics of the cluster membership matrix

	Mean	Min.	Max.	SD	Sum
(EM,70)	0.20	0.00	0.99	0.31	142.98
(TR,23)-(EM,47)	0.17	0.00	0.94	0.25	123.27
(FE,22)-(EM,48)	0.17	0.00	0.94	0.23	119.09
(FE,46)-(EM,24)	0.12	0.00	0.88	0.19	88.22
(FE,25)-(HE,45)	0.10	0.00	0.96	0.23	73.96
(SC,25)-(HE,45)	0.14	0.00	0.98	0.29	97.21
(TR,22)-(JL,48)	0.09	0.00	0.80	0.16	67.27

Table 3 Example of an augmented dataset

Sequence	Weight	Cluster
s_1	u_{11}	1
s_1	u_{12}	2
s_1	u_{13}	3

clustering. However, average cluster membership provides similar information: it can be interpreted as the relative frequency of each cluster, if sequences are weighted according to their membership strength.

The maximal value is also interesting, as it provides an estimation of the quality of the chosen representative. The higher the membership, the better the representative (i.e., a value of 1 would identify a sequence that fully belongs to that cluster). In some clusters, the maximum is quite low, if we consider that the maximal possible value is 1. For instance, in the training–joblessness cluster, the maximum equals 0.8. Hence, our representative is also close to other clusters—perhaps cluster 2. To describe the clusters, we therefore need to take into account more information than just the sequences with the highest membership.

4.2.2 Weight-Based Presentation

Our second proposition in analyzing the *fuzzy* cluster is to weigh the sequences according to their membership strength or probabilities. We augment the dataset by repeating the sequence s_i of individual i k times (i.e., once per cluster). We therefore have k sequences for individual i, denoted as $s_{i1} \cdots s_{ik}$. We weight these sequences according to their membership degree $u_{i1} \cdots u_{ik}$. Hence, even if the same sequence were repeated k times, its weights will sum to 1. We then create a new categorical covariate in this augmented dataset, and it specifies the cluster (ranging from 1 to k) of the associated membership degree.

Table 3 presents a small example, to make the presentation clearer. Suppose we have three clusters named 1, 2, and 3. For individual 1 with the sequence s_1, we have three observations (i.e., one per cluster). The first observation is weighted according to the strength of membership to the first cluster, and the associated cluster covariate is set to 1. We then repeat the same procedure for each observation.

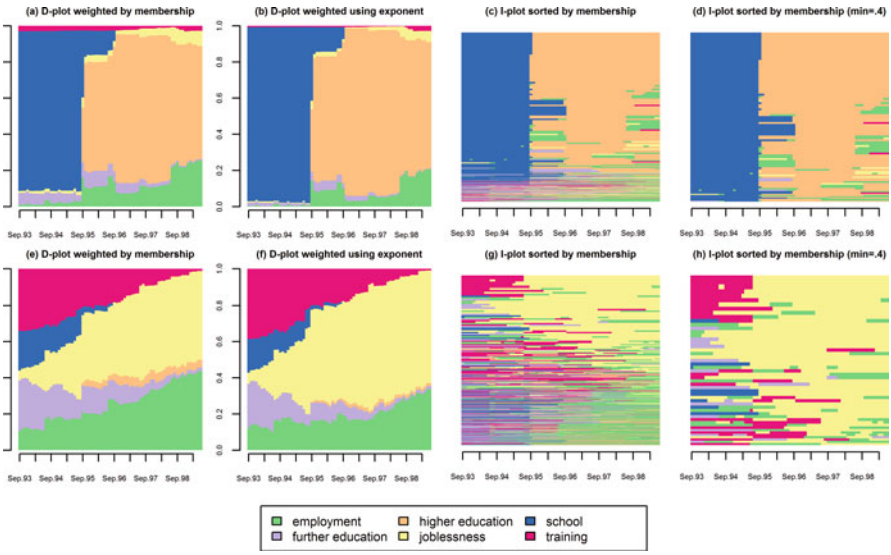

Fig. 2 Membership-weighted plots of the sequence for clusters "(SC,25)-(HE,45)" and "(TR,22)-(JL,48)"

This weighting strategy allows us to use any tools available for weighted sequence data. For instance, we can use a sequence distribution plot. Figure 2 proposes several plots of these membership-weighted sequence data for the clusters "School–Higher Education" and "Training–Joblessness." Subfigures (a) and (e) present sequence distribution plots for these clusters. If the cluster "School–Higher Education" seems to be quite well defined, the "Training–Joblessness" one shows more discrepancy, as we already noted.

This weighting strategy is also supported from a more statistical perspective. Minimizing Eq. 1 is equivalent to minimizing a residual sum of squares in a discrepancy analysis of this augmented dataset (Studer et al. 2011). More precisely, it minimizes the residual sums of squares of this augmented dataset, where each sequence is weighted $u_{i1}^r \cdots u_{ik}^r$ and the explanatory categorical covariate would be the cluster $1 \ldots k$.

Following this reasoning, we can weigh the sequences by using the exponent (here, $r = 1.5$). The result might be closer to the underlying algorithm. However, the interpretation is more difficult and, as we will see, it is also interesting in the following analysis to rely on the membership strength. For this reason, this approach should be used mostly to describe the clusters. The result of this strategy is shown using a d-plot in subfigures (b) and (f); in both cases, the cluster seems to be better defined.

By using index plots, we can take a closer look at the longitudinal patterns. In this case, we additionally suggest ordering the sequences according to membership degree. The result is shown in subfigures (c) and (g). The most typical sequence lies at the top of the subfigures, with a high membership degree; meanwhile, the bottom

shows less-characteristic patterns. Interpretation should be made with caution, as it depends on the maximal membership degree. In the cluster "School–Higher Education," this maximum is close to 1, while in the other one it reaches only 0.8.

It can be interesting to focus on the sequences with the highest membership. In subfigures (d) and (h), we discarded sequences with a membership degree below 0.4. Interestingly, among the sequences with the highest membership in the "Training–Joblessness" cluster, we find sequences starting in "Further Education" or "Employment," for instance.

We propose several methods by which to visualize and describe a *fuzzy* typology; these methods allow us to properly interpret this typology. However, most sequence analysis applications go beyond the typology description by studying the factors that influence the kinds of trajectories being followed. We now turn to this kind of analysis for *fuzzy* clustering.

4.3 Analyzing Cluster Membership Using Dirichlet Regression

In typical sequence analysis, one often relies on multinomial regression to explain cluster membership (Abbott and Tsay 2000). The aim is to identify how covariates explain the trajectory type that is followed. This cannot be done with *fuzzy* clustering, because our typology is described by a membership matrix and not by a categorical variable. Assigning each sequence to the cluster with the highest membership strength is not a solution either, for in doing so, we would lose all the added value inherent in *fuzzy* clustering.

Several models are available to analyze membership matrices, which can be seen as "compositional data" (Morais et al. 2016; Pawlowsky-Glahn and Buccianti 2011). Here, for two reasons, we suggest relying on Dirichlet regressions (Maier 2014), which are extensions of beta regression (Ferrari and Cribari-Neto 2004). First, interpretations of them are very similar to those of multinomial models, if we use the so-called alternative parametrization. Second, good performance is reported in Maier (2014) and in Morais et al. (2016), even under some violations of the statistical assumptions. In the current model, one of the categories (i.e., clusters) is chosen as the reference; we then estimate how explanatory factors influence the likelihood of being fully classified in a category, rather than in the reference.

Interpretations of the coefficients are very similar, then, to the multinomial ones, and they can be interpreted in the usual log-odds scale. Their exponents can therefore also be interpreted as "odds-ratio" values on cluster membership. In a Dirichlet regression, one can also estimate the effect of covariates on a "precision" parameter that measures the precision of estimation. (This parameter is named "precision" because it takes a high value when the residual variance of the dependent variable tends to be lower.) This can be used to take into account possible heteroscedasticity.

Table 4 presents the coefficients of the Dirichlet regression. We used the employment cluster as the reference. To simplify our presentation, we included only three covariates: gcseq5eq (the qualifications gained by the end of compulsory

Table 4 Dirichlet regression of cluster membership

	(TR,23)-(EM,47)	(FE,22)-(EM,48)	(FE,46)-(EM,24)	(FE,25)-(HE,45)	(SC,25)-(HE,45)	(TR,22)-(JL,48)
(Intercept)	-0.14^{*}	-0.14^{*}	-0.40^{***}	-0.77^{***}	-0.92^{***}	-0.53^{***}
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Grammaryes	0.09	0.03	0.05	0.18	0.82^{***}	0.16
	(0.13)	(0.13)	(0.13)	(0.13)	(0.12)	(0.13)
gcse5eqyes	0.13	0.37^{***}	0.55^{***}	0.98^{***}	1.06^{***}	0.42^{***}
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
funempyes	0.10	0.06	0.13	0.09	0.04	0.26^{*}
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)

Log likelihood = 6402.14; Num. obs. = 712; Precision= $1.22^{***}(0.02)$; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

education: five or more GCSEs at Grades A to C, or equivalent), Grammar (grammar school secondary education), and funemp (father unemployed at the time of the survey).

Let us interpret the coefficient of our covariate on the membership degree (adequacy with) or chance (if we think about probabilities) to follow the more "at-risk" pattern "(TR,22)-(JL,48)," as opposed to employment. We observe no significant difference between those having had a grammar school education and those who had not. However, individuals with unemployed fathers did tend to have a higher membership in this cluster (significant positive coefficient) than in the employment cluster—or, if we take the "probability" interpretation, they have a lower chance of following this pattern than the reference (employment). The same applies to those who had the five-grade qualification (variable `gcse5eq`)—probably because they are very unlikely to follow the reference employment trajectory.

Additionally, it is often useful to understand the distinctive features of each cluster. For *crisp* clustering, this can be achieved by running a logistic regression on a dummy variable that measures cluster membership. Here, we can make use of beta regression, which aims to model a dependent variable that lies in the [0, 1] interval (Ferrari and Cribari-Neto 2004).[3] The interpretation of the coefficient is similar to that of the Dirichlet regression. The exponent of the coefficients can be interpreted as an "odds-ratio" on cluster membership. Here, again, a "precision" parameter allows us to take into account over- or under-dispersion. The results lead to similar conclusions but further highlight that those who had the five-grade qualification (variable gcse5eq) are very unlikely to follow the employment trajectory type of sequence.

[3] Unlike logistic (binomial) regression, beta regression does not assume that the dependent variable is a proportion (i.e., the result of a count of 0 and 1). Furthermore, it can cope with over- or under-dispersion.

4.4 Running the Analysis in R

The Fanny algorithm is available in the `cluster` package, through the `fanny` function. Aside from the distance matrix `diss`, one needs to specify the number of groups (argument `k=7`) and set the argument `diss=TRUE` to specify that we provided a distance matrix and not a dataset. Finally, the value of the fuzziness parameter r can be set through the `memb.exp` argument (default value of 2). The returned object provides the membership matrix (`fclust$membership`) and additional information such as quality measures or related *crisp* clustering.

```
R>  ## Fuzzy clustering in 7 groups using r=1.5
R>  fclust <- fanny(diss, k=7, diss=TRUE, memb.exp=1.5)
R>  ## Display the resulting clustering with membership
     threshold of 0.4
R>  fuzzyseqplot(myseq, group=fclust$membership, type="I",
         membership.threshold=0.4, sortv="membership")
R>  ##Estimation of Dirichlet Regression
R>  ##Dependent variable formatting
R>  fmember <- DR_data(fclust$membership)
R>  ## Estimation
R>  bdirig <- DirichReg(fmember~var1+var2|1,
         data=mydata, model="alternative")
R>  ## Displaying results of Dirichlet regression
R>  summary(bdirig)
R>  ## Estimation of beta regression
R>  breg1 <- betareg(fclust$membership[, 1]~var1+var2, data=mydata)
R>  ## Displaying results
R>  summary(breg1)
```

All the visualizations proposed here are available in the `WeightedCluster` package, through the `fuzzyseqplot` function. The function works in the same ways as the usual `seqplot` function available in `TraMineR`, except that the `group` argument should be a membership matrix or a `fanny` object. Furthermore, one can specify a membership threshold (for instance, 0.4) and whether graphics should be weighted by membership strength. If one wants to weights the sequences using the fuzziness parameter, one should set `memb.exp` to the correct value. By default, the fuzziness parameter is not used; hence, the `memb.exp=1`. When using index plots (`type="I"`), one can additionally set `sortv="membership"` to sort the sequences in each plot according to their membership strength.

Dirichlet regression can be estimated using the `DirichReg` function in the `DirichletReg` package (Maier 2014), while the beta regression can be computed with the function `betareg` available in the `betareg` package (Cribari-Neto and Zeileis 2010). For the former, the dependent variable should first be formatted using the function `DR_data` before estimating the model using `DirichReg`. For beta regressions, a separate regression should be estimated for each cluster. One needs

to specify the cluster membership strength on the right-hand side of the `R` formula, while adding covariates on the left-hand side as usual. In both cases, one can set a data frame where covariates should be found, using the usual `data` argument.

5 Conclusion

In this paper, we introduced two alternative clustering methods, each of which has its own strengths. We believe that property-based clustering is a very promising sequence analysis tool. Having clustering membership rules allows one to reproduce and validate a typology; furthermore, it significantly simplifies the interpretation of the clustering.

Property-based clustering is also useful in understanding the underlying criteria used by a dissimilarity measure to compare trajectories. For instance, in our example application, all splits were made according to the overall time spent in different states. This prevalence of duration illustrates once again that optimal matching tends to favor duration while comparing sequences. The use of other distance measures such as those reviewed in Studer and Ritschard (2016) or those introduced in this bundle in Collas (2018) or Bison and Scalcon (2018) would lead to the selection of other properties. For instance, sequence pattern properties would probably be selected by using a distance sensitive to sequencing, such as SVRspell (Elzinga and Studer 2015).

On the other hand, *fuzzy* clustering has been seldom used in sequence analysis. Nonetheless, the method should be useful in many situations. First, in many cases, exact cluster membership is doubtful (Warren et al. 2015). *Fuzzy* clustering allows one to relax the assumption that cluster memberships have been correctly retrieved by the cluster analysis; it does so by allowing multiple cluster memberships. This is also an interesting perspective from a sociological viewpoint, as trajectories might be influenced by several trajectory types. Second, in *fuzzy* clustering, membership is thought to be gradual; this too is interesting from a social science perspective. Some trajectories might be more typical of a type than others.

The aim of this study was to develop tools by which to facilitate the use, interpretation, and analysis of both clustering methods. However, further application of these methods is still needed to fully assess their strengths and weaknesses with regards to sequence analysis. We believe that this study is a first step in that direction.

Acknowledgements The author warmly thanks the two reviewers for their insightful comments on an earlier version of the manuscript. This publication benefited from the support of the Swiss National Centre of Competence in Research LIVES Overcoming vulnerability: Life course perspectives (NCCR LIVES), which is financed by the Swiss National Science Foundation (grant number: 51NF40–160590). The author is grateful to the Swiss National Science Foundation for its financial assistance.

References

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research*, *29*(1), 3–33. (With discussion, pp. 34–76).

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In P. S. Yu & A. L. P. Chen (Eds.), *Proceedings of the International Conference on Data Engeneering (ICDE), Taiwan* (pp. 487–499). IEEE Computer Society.

Bison, I., & Scalcon, A. (2018). From 07.00 to 22.00: A dual-earner couple's typical day in Italy. Old questions and new evidence from social sequence analysis. In G. Ritschard & M. Studer (Eds.), *Sequence Analysis and Related Approaches: Innovative Methods and Applications*. Cham: Springer (this volume).

Chavent, M., & Lechevallier, Y. (2006). Empirical comparison of a monothetic divisive clustering method with the ward and the k-means clustering methods. In V. Batagelj, H.-H. Bock, A. Ferligoj, & A. Žiberna (Eds.), *Data science and classification* (pp. 83–90). Berlin/Heidelberg: Springer.

Chavent, M., Lechevallier, Y., & Briant, O. (2007). DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis, 52*(2), 687–701.

Collas, T. (2018). Multiphase sequence analysis. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software, 34*(2), 1–24.

D'Urso, P. (2016). Fuzzy clustering. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 545–573). New York: Chapman & Hall.

Elzinga, C. H., & Studer, M. (2015). Spell sequences, state proximities and distance metrics. *Sociological Methods and Research, 44*(1), 3–47.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799–815.

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2011). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, & J. Filipe (Eds.), *Knowledge discovery, knowledge engineering and knowledge management* (Communications in computer and information science (CCIS), Vol. 128, pp. 94–106). Berlin/Heidelberg: Springer.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis*. New York: Wiley.

Maechler, M., Rousseeuw, P., Struyf, A., & Hubert, M. (2005). Cluster analysis basics and extensions. Rousseeuw et al. provided the S original which has been ported to R by Kurt Hornik and has since been enhanced by Martin Maechler: Speed improvements, silhouette() functionality, bug fixes, etc. See the 'Changelog' file (in the package source).

Maier, M. J. (2014). Dirichletreg: Dirichlet regression for compositional data in R. Research Report Series/Department of Statistics and Mathematics 125. WU Vienna University of Economics and Business, Vienna.

McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A, 165*(2), 317–334.

Morais, J., Thomas-Agnan, C., & Simioni, M. (2016). A tour of regression models for explaining shares. Working Paper 16–742, Toulouse School of Economics.

Pawlowsky-Glahn, V., & Buccianti, A. (Eds.) (2011). *Compositional data analysis: Theory and applications*. Chichester: Wiley.

Piccarreta, R., & Billari, F. C. (2007). Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(4), 1061–1078.

Salem, L., Crocker, A. G., Charette, Y., Earls, C. M., Nicholls, T. L., & Seto, M. C. (2016). Housing trajectories of forensic psychiatric patients. *Behavioral Sciences & The Law, 34*(2–3), 352–365.

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. LIVES Working Papers 24, NCCR LIVES, Switzerland.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society, Series A, 179*(2), 481–511.

Studer, M., Müller, N. S., Ritschard, G., & Gabadinho, A. (2010). Classer, discriminer et visualiser des séquences d'événements. *Revue des nouvelles technologies de l'information RNTI, E-19*, 37–48.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research, 40*(3), 471–510.

Warren, J. R., Luo, L., Halpern-Manners, A., Raymo, J. M., & Palloni, A. (2015). Do different methods for modeling age-graded trajectories yield consistent and valid results? *American Journal of Sociology, 120*(6), 1809–1856.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning, 42*(1/2), 31–60.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

From 07.00 to 22.00: A Dual-Earner Couple's Typical Day in Italy

Old Questions and New Evidence from Social Sequence Analysis

Ivano Bison and Alessandro Scalcon

1 Introduction

How do dual-earner couples organize their workdays and how do they (de)synchronize their daily activities? These are the questions that we address in this paper using a multichannel sequence analysis approach. Our purpose is to consider the couples' division of work-family activities in holistic terms by setting it within the context of everyday life, that is, the overall temporal pattern of combination of His and Her multiple activities.[1]

Our multichannel approach is based on a Lexicographic Index (Bison 2011) that seeks to overcome some optimal matching limits of sequence analysis (Bison 2009). The case-study concerns Italian dual-earner couples and uses data from the Italian Time Use Survey 2008 (Istat 2011).

We know that for dual-earner couples the risk of experiencing a certain "lack of family time" is higher than for other couples (Saraceno 2012), due to the combination and the rigidity of His and Her work constraints. The spouses of these couples face various challenges: according to their working schedules, they are required to find the right amount of time for their family—i.e. housework, childcare and other non-paid work—as well as with their family—i.e. desirable

[1] We immediately point out that, in this paper, we only consider heterosexual couples, because of the limitations of the Italian Time Use Survey questionnaire. Moreover, we stress that the choice of using the male pronoun before the female one is perfectly conscious: to make the reading easier, we needed to follow a single criterion and we decided to cite the spouses following the order of records in our data files, i.e. male-female.

I. Bison (✉) · A. Scalcon
Department of Sociology and Social Research, University of Trento, Trento, Italy
e-mail: ivano.bison@unitn.it

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_14

and shared activities, such as free time—integrating collective needs with individual ones. In other words, the time scarcity of dual-earner couples obliges them to adopt a *complementary strategy* (Mansour and McKinnish 2014) in order to reconcile their multiple family needs of "production" and "consumption", while preserving their personal satisfaction with work and daily life. The two individual careers have to coexist with a third one: family life.

In this scenario, an important component of a dual-earner couple's strategy is synchronization/desynchronization. For instance, previous studies show that a certain degree of desynchronization of working schedules may be a useful solution for partners because it can promote a more equal division of housework and child care (Presser 1994; Chenu and Robinson 2002; Lesnard 2008; Naldini and Saraceno 2011). At the same time, it is recognized that a certain degree of synchronization in work commitments can encourage the partners to spend time together in other desirable activities (Hamermesh 2002; Lesnard 2008).

More generally, looking at (de)synchronizations is important because they reveal a latent behavioral pattern of different work-family specializations and suggest new explanations for the continuing persistence of gender inequalities in the division of work-family activities. The study of (de)synchronizations could enable identification of multiple equilibria (Esping-Andersen et al. 2013). According to Esping-Andersen and colleagues, such study could reveal different behavioral patterns of work-family specializations—(i) egalitarian, (ii) unstable and (iii) traditional—resulting from systematic co-action by different generative mechanisms, both symbolic-cultural (e.g. Berk 1985; West and Zimmerman 1987) and economic-material (e.g. Becker 1964; Coverman 1985; Manser and Brown 1980).

How do we measure (de)synchronizations? There are two radically different main approaches: time budgets—the dominant approach in the time use literature—and sequence analysis.

In the former approach, we measure (de)synchronizations as the amount of time in which both the spouses have or have not done an activity at the same time or have or have not spent that time together in the same place. Thus, we obtain synchronicity ratios or percentages. However, in this way nothing is known about when the activity schedules overlap. This is a crucial limitation for two main reasons.

First, time is socially structured, and so too are social rhythms and constraints. Hence, being simultaneously at work at 10 a.m. or 10 p.m. has radically different impacts on a couple's daily life. Furthermore, different timings in working schedules may have radically different impacts on daily life if combined with other time demanding features, like for instance the institutional constraints of children's schedules (e.g. school hours).

Second, take the case of a full-time shift perfectly synchronized with a part-time afternoon shift: by considering only the duration of the overlap, we will mistakenly classify it as a highly desynchronized working schedule. However, such kinds of structural desynchronization—due simply to differences in duration—should not be compared with hypothetical other kinds with the same off-scheduling amount but different and more complex organization during the day (Nock and Kingston 1984).

According to Lesnard (2008), if we know little about how family time is daily balanced with work time for both spouses, this is mainly because we are used to adopting the too simplistic approach of the dominant time-budget perspective. To date, scholars have underestimated the importance of daily scheduling, while paying more attention to total amounts of time (Lesnard 2008). They have traditionally acquired time budget information related to different daily activities, but these should be seen in a holistic perspective that makes it possible to study the couple's days as a whole, avoiding the manipulation of time as if it were clay.

An alternative to the time-budget approach is sequence analysis (Lesnard 2008). According to Hallberg (2003), "while the traditional time allocation model typically studies the total time spent in, e.g., market work, over a day or a week, it provides little or no insight into the temporal pattern of time-use and therefore, potentially, misses a vital part of the mechanisms underlying empirical observations". A sequence analysis of time-use would evidence the routine aspects of daily life, as well as the couples' projects (Hagerstrand 1982), the performance of their complementary strategy across several daily constraints and unexpected events (Hellgren 2014). Finally, the analysis of time-use temporal patterns—instead of time budgets—seems more relevant in the study of the daily strategies and behaviors of a couple (Hallberg 2003).

We have pointed out that in order to understand the complexity of work-family balance strategies, it is necessary to study the couple's daily time-use pattern as a whole and in a more holistic way by adopting a multichannel sequence analysis approach. In the following section, we introduce the Lexicographic Index used to measure the resemblance between multinomial sequences. Section 3 sets out the data and methods. Section 4 presents the main statistical and graphical results of this study. Finally, Sect. 5 is devoted to summing up the main findings.

2 The Lexicographic Index

There are three main problems with current techniques used to compute the distances among sequences. One derives from the way in which similarity between two sequences is defined (Abbott and Tsay 2000; Wu 2000; Dijkstra and Taris 1995; Elzinga 2003; Bison 2009); the second is how to handle multinomial sequences (Abbott 1990); the third is how to treat a multichannel sequence as a whole (Gauthier et al. 2010).

Here we present an alternative method for computing distances among sequences. The lexicographic index (Bison 2011) is based on the sorting order of two different modes of observing events in a binary sequence. The first order is given by duration, that is the quantity of time, and is therefore based on the total number of observed events u in the sequence x. The second order is timing, that is when this event happens, i.e. the 'places' s_k $(k = 1, \ldots, u)$ in the sequence

when 1 occurs.[2] For instance, we may have only three binary sequences of length 3 and $u = 1$. They differ according to when the event occurred; at time $t1$, $t2$ or $t3$. Hence, we may order these sequences [100], [010] and [001] according to the time order of events. Because the nature of the sorting order is double, the proposed index consists of two distinct parts.

The first part, $d'(x)$, ranging from 0 to 1, takes account of the duration and therefore the different amounts of realization u recorded in the sequence:

$$d'(x) = u/T \text{ for } u > 0 \quad \text{and} \quad 0 \text{ for } u = 0 \tag{1}$$

where T is the length of the sequence.

The second part, $d''(x)$, ranging from 0 to 1, takes account of timing and therefore the different numbers of combinations displayed by the sequences with variation in the amount of time. It is

$$d''(x) = \begin{cases} 0 & \text{for } u = 0 \\ \dfrac{\binom{T}{u}}{\binom{T}{u}+1} \dfrac{1 + \binom{T}{u} - \left[\binom{B_u}{C_u} - \sum_{k=1}^{u} \binom{B_k}{C_k} - \binom{A_k}{C_k}\right]}{\binom{T}{u}} & \text{for } 0 < u < T \\ 1 & \text{for } u = T \end{cases} \tag{2}$$

where A_k is the exact position of s_k in the sequence, B_k is the last position that s_k can occupy within the sequence, and C_k is the first position that s_k can occupy. For example, for sequence [0101], with $T = 4$ and $u = 2$, we have: for s_1 the exact place of the first 1 is $A_1 = 2$, the last position is $B_1 = 3$ and the first position is $C_1 = 1$; for s_2 the exact place of the second 1 is $A_2 = 4$, the last position is $B_2 = 4$ and the first position is $C_2 = 2$, the value of $d''([0101])$ is 0.285714 and is obtained as

$$\frac{\frac{4!}{2!(4-2)!}}{\frac{4!}{2!(4-2)!}+1} \frac{1 + \frac{4!}{2!(4-2)!} - \left(\frac{4!}{2!(4-2)!} - \left(\frac{3!}{1!(3-1)!} - \frac{2!}{1!(2-1)!} + \frac{4!}{2!(4-2)!} - \frac{4!}{2!(4-2)!}\right)\right)}{\frac{4!}{2!(4-2)!}} .$$

In turn, these two indices are the coordinates of the sequence in a bi-dimensional space and the distance between two binary sequences (x_i, x_ℓ) is the Euclidean distance between a couple of lexicographic indices

$$r(x_i, x_\ell) = \sqrt{\left(d'(x_i) - d'(x_\ell)\right)^2 + \left(d''(x_i) - d''(x_\ell)\right)^2} \ . \tag{3}$$

[2]Duration and timing are two of the three aspects identified by Studer and Ritschard (2016) as mattering in sequence comparison. Here, we do not pay attention to the third one, sequencing, which is not a concern for studying (de)synchronization.

Passing from a binary sequence to a multinomial sequence is easy. Just as a qualitative variable of m modality can be represented by m dummy variables, so a multinomial sequence of alphabet Q can be represented by $|Q|$ binary sequences x_q with values 0–1. For example, the sequence $x = [123321]$ and alphabet $Q = \{1, 2, 3\}$ can be represented by the following three binary sequences $x_1 = [100001]; x_2 = [010010]; x_3 = [001100]$. To each of these binary sequences it is possible to apply the lexicographic index and compute the coordinates $\{d'_q(x_q); d''_q(x_q)\}$. The multinomial sequence x is therefore described by a vector of real numbers. The distance between two multinomial sequences (x_i, x_ℓ) is the Euclidean distance between their transformations $\{d'(x_{iq}); d''(x_{iq})\}$ and $\{d'(x_{\ell q}); d''(x_{\ell q})\}$. Formally, it is:

$$r'(x_i, x_\ell) = \sqrt{\sum_{q=1}^{|Q|} \left(d'(x_{iq}) - d'(x_{\ell q})\right)^2 + \left(d''(x_{iq}) - d''(x_{\ell q})\right)^2} \ . \tag{4}$$

We conclude this section of the paper by briefly discussing the index just presented. Firstly, it is not a comparison between the sequences that defines their distance. The index has a known beginning and end; each point is univocal and identifies one and only one combination of states in sequence. Two sequences which differ in the position of only one element will have different positions. From every point one can retrace the exact sequence that has produced it. A second characteristic of the index concerns its output. Each value of the index, in fact, can be conceived as a coordinate in the space of the multinomial sequence. This characteristic enables the researcher to adopt different methods to calculate the distance, but also to define forms of space other than Euclidean. Furthermore, the third characteristic is the natural way in which to handle multichannel sequences.

3 The Data, Their Organization and the Coding of the Activities in a Multichannel Approach

The goal of our analysis was to discover how 873 Italian dual-earner couples organized their daily activities during a typical work day from Monday to Friday.We used data from the Italian Time Use Survey 2008 (Istat 2011). We considered time-use diaries of dual-earner couples' activities (His and Her) from 7.00 to 22.00.[3] Each daily activity was observed every 10 min, and the data files for the sequence analysis consisted of 873 pairs of sequences, one for Him and one for Her, with

[3]Excluded from the sample were: (a) couples living with other couples (parents or others); (b) couples that filled in the questionnaire on different days, or during the weekend; (c) couples with incomplete information by one or both of the spouses; and, (d) couples in which his or her age was over 65.

a total of 90 points in time. Each couples of rows of this file corresponded to a cohabitation, while each variable corresponded to 10 min of observation and each cell of the row/column intersection stated the activities of Him or Her at time t.

In order to simplify the analysis, six different groups of activity were considered: Sleep; personal care—i.e. having a shower, eating, etc. (P.Care); paid work (Work); moving—any kind (Move); unpaid work—i.e. housework, child care, repair, etc. (H.Care); free time and other activities with or without others (F.Time).

Having defined the six daily macro-activities, the next step was to establish how to codify the daily activities of Him and Her in the couple. In this case, His activities and Her activities interact in time to give rise to the couple's daily activities. Taken individually, each of these two sequences takes the form of a series of mutually exclusive episodes. The problem is therefore how to codify two interacting sequences composed of a plurality of mutually exclusive events. To date, all the solutions proposed have been based on the generation of events combinations (Pollock 2007; Gauthier et al. 2010; Aisenbrey and Fasang 2017): that is, on the construction of a single sequence that combines the states of Him and Her.

This operation has several consequences. Firstly, as Abbott pointed out, using combinations of events requires one to pay "…the price of losing all information about the temporal 'shape' of events—their duration and their intensity in terms of producing occurrence—in short their time horizon" (Abbott 1990, p. 146). Secondly, there is the risk that distinct time-use patterns will be tied together, although the order of causality may be bi-directional.

There are various reasons to believe that daily activities of Him and Her cannot be reduced to a simple combination of states. Internally, moreover, each sequence consists of states regulated by their own mechanisms which operate differently in defining the timing and duration of each individual episode. For instance, consider the mechanisms that underlie the regulation of the states of free time and housework. In the former case, it is the working time that mainly regulates the time spent on these two activities; in the latter, we should expect a stronger interaction between gender roles.

It is therefore possible to hypothesize that the sequences of Him and Her—and the states of which they are composed—have their own underlying generative mechanisms which establish the timing and duration of episodes. These generative mechanisms work independently of each other and interact in time: they stand in a coexistence relationship. Finally, the couple's daily activities are the result of a complex process of co-action between two sequences, that of Him and that of Her, regulated by different generative mechanisms resulting from the co-action between different states. Consequently, reducing everything to a combination of events means loss of a large part of information about the temporal 'shape' of events.

A couple's daily activities, or more correctly the couple sequences analyzed here, are therefore configured by the co-action of two multinomial sequences composed of mutually exclusive episodes. By extending the proposed application of the Lexicographical Index (see Sect. 2) to this case-study, 12 binary sequences can be defined, six for His states and six for Her states, each one of length $t = 90$, that is, the overall number of points of observation. The couple sequence is defined as a

point in a 24-dimension space whose coordinates are the 24 lexicographic indexes defining the respective sequences of Him and Her. The distance between two couple sequences is given by the Euclidean distance between the two points of the two sequences in the 24-dimension space.

The coordinates defined for all 873 couples were analyzed using a k-means cluster algorithm.[4] The joint exam of the scree plot (Makles 2012) of the kink in the curve generated from the within sum of squares (WSS), the η^2 coefficient (0.43) and the proportional reduction of error, suggests that seven is the optimal number of groups drawn from a set of 20 cluster solutions with random starting points.

4 From 7.00 to 22.00: A Typical Working Day of a Dual-Earner Couple in Italy

It is not news that the everyday life of a dual-earner couple is complex. It involves a long and difficult schedule of: waking up, having a shower, breakfast, taking the car-bus-train, going to work, beginning work, lunch, resuming work, coming back home, then housework and family/child care for Her, relaxation for Him, dinner, and at the end of the day, before they go to sleep, some leisure activity. Overall (Fig. 1), this was also the typical daily routine followed by our 873 Italian dual-earner couples from 7.00 to 22.00. Looking at the most frequent activity combinations in the morning, at 7.00, 75.0% of couples were involved in personal care or going to work. From 8:00am to 6:00pm all the couples were at work.[5] At 6:00pm, the couples started to be desynchronized: She was engaged in housework/children care; meanwhile He continued to work until 7.00pm. From 7.00pm to 7.30pm, He had some free-time activities, while She continued her housework activity. Finally, together, they had dinner and engaged in free-time activities.

Differences in the spouses' daily time-budgets for each activity (Table 1) also confirm a well-known finding on the unequal gender division of work-family activities (Gershuny and Robinson 1988; Raley et al. 2012; Craig et al. 2014). For instance, on average, She spends 2 and a half hours more than her partner on housework and childcare, while He spends 1 h and a half more than his partner on paid work.

However, the timing of this daily organization changes when we move from general into the seven clusters. In this case, there emerges a more composite picture of daily life, where "multiple equilibria" (Esping-Andersen et al. 2013) of time allocation during a typical workday and (de)synchronization strategies jointly explain the dual-earner couple's patterns of time use. The average time activities (Table 1) show a clear difference in the time spent on each activity by Him and Her within partners and among clusters.

[4]Our distance measure between sequences could as well be used for clustering with the property-based and fuzzy methods addressed by Studer (2018) in this bundle.

[5]The absence of a break for lunch does not mean that spouses do not eat; only that, overall, there is not a common time interval for lunch due to the different work schedules.

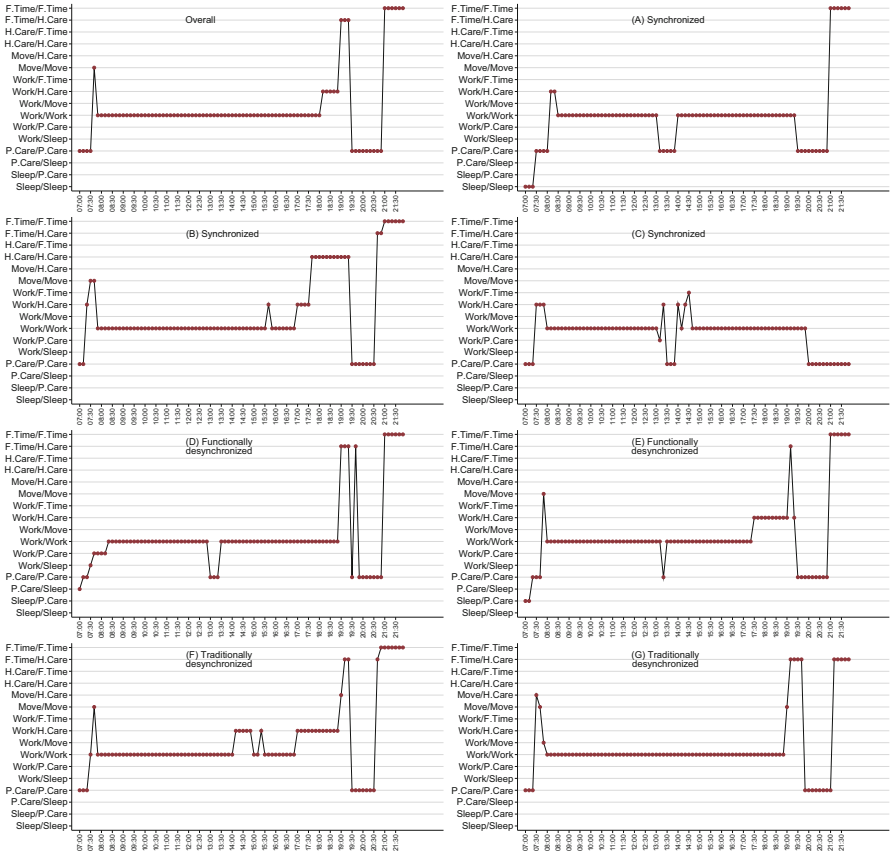

Fig. 1 Modal sequence graphs of the (de)synchronized patterns of His/Her activities (labels show Him activity/Her activity in that order)

Moreover, on shifting the focus to the schedules of each activity, the modal sequence graphs[6] (Fig. 1), give us a clearer picture of how the strategies of dual-earner time-use change over time in a typical workday. Both within the spouses and among the clusters the differences in time-use (Fig. 1) mainly occur in the second part of the workday. Until noon the couple's everyday lives are quite "synchronized". Him and Her show differences in the afternoon, when fewer women than men are at work and when the women shift their activities from paid to unpaid work (housework, child care, etc.). In other words, gender inequalities in the work-family balance are generally set in the afternoon.

[6]For each cluster and for each point-in-time, the most frequent activities combination was identified. On this criterion, only 16 of all the 36 (six for each spouse) possible combinations were found to be frequently performed by the couples, suggesting a certain routine by couples in everyday life.

Table 1 Mean time (in hours:minutes) spent on each activity, by cluster

	Sleep		Personal Care		Paid Work		Unpaid Work		Moving		Free Time		
Cluster	M	W	M	W	M	W	M	W	M	W	M	W	*N*
(A)	01:04	00:44	02:02	01:58	07:27	06:09	01:10	03:27	01:17	01:19	02:00	01:24	108
(B)	00:04	00:03	01:50	01:49	07:26	06:30	01:51	03:51	01:23	01:18	02:27	01:28	158
(C)	00:14	00:11	02:06	02:01	07:58	06:16	01:13	03:30	01:42	01:26	01:47	01:36	67
(D)	00:03	00:38	02:09	02:01	08:38	06:07	00:50	03:35	01:18	01:17	02:02	01:23	99
(E)	00:51	00:02	02:10	01:53	07:46	06:58	01:05	03:25	01:14	01:24	01:55	01:18	129
(F)	00:08	00:08	02:08	01:51	08:45	06:07	00:30	03:59	01:24	01:18	02:05	01:37	179
(G)	00:01	00:00	02:02	01:50	08:32	07:28	01:08	03:15	01:35	01:38	01:42	00:49	133
Total	00:19	00:13	02:03	01:54	08:06	06:32	01:06	03:37	01:24	01:23	02:02	01:22	873
F test	31.5**	42.5**	3.6**	1.6	12.9**	8.6**	18.9**	2.8*	3.1**	3.9**	4.3**	6.9**	

Note: *$p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Hence, the preliminary results seem to suggest that, on the one hand, gender specializations in different activities can assume different meanings when contextualized in the whole day and, on the other hand, that the partners' daily life seems to develop with socially shared, recognized, and identifiable patterns of combined time use. This insight raises two further questions. The first is how these patterns result from a complex process of adaptation to both work-social-family constraints and individual needs. The second question concerns how the daily times are combined by spouses, and how their performed combinations are random instead of being regulated by common generative mechanisms.

In order to investigate the complex process of adaptation of dual-earner couples' daily time organization, we ran a multinomial logistic regression model to verify if such patterns resulted from working-social-familial and individual constraints (Table 2). For this analysis, we used information about the couple's educational qualifications, social class, economic sector, and the presence of children.[7]

A joint reading of the modal multichannel sequence graphs (Fig. 1), the multinomial logistic regression parameters (Table 2) and the margins estimated probability (Table 3) quite clearly shows what are the (de)synchronization strategies adopted by couples and what may be the hidden generative mechanisms (Hallberg 2003). We highlight the importance of the presence of children, work sector and the educational level in explaining the cluster differences (Table 2).

Three different forms of time-use organization are highlighted by the graphs (Fig. 1). The first is characterized by a general synchronization of the spouses'

[7]Educational qualifications were classified as: (1) compulsory level—elementary school certificate (including no educational qualifications) and lower-secondary school certificate (including 2-to-3 year vocational certificates); (2) upper-secondary school diploma (including post-secondary diplomas); and, (3) university degree (including postgraduate qualifications). Social class was classified according to the EGP scale: (I+II) Service class; professionals, administrators, and managers; (IIIa) Routine non manual workers; (IVabc) Petty bourgeoisie; Farmers; (VI+VIIab) Skilled and non-skilled workers; Agricultural Labourers. The economic sector (agriculture and industry, private services and public services) of the couple was the combination of the main job sector of Him and Her. The couple's educational level (and social class) is defined as the highest educational level (social class position) between the spouses.

Table 2 Multinomial logistic regression on the seven clusters by presence of children and sector, level of education and social class of couple. (Jackknife replication). Reference cluster (A)

	B	C	D	E	F	G
Children (ref. No child)						
Children 0–14	0.72**	0.15	0.13	0.66*	0.52	0.92***
Children 14+	0.84**	−0.17	0.52	0.54	0.82**	0.93**
Sector (ref. both Industry)						
Both priv. services	−2.22***	−1.47**	−1.62**	−1.61**	−2.58***	−1.89***
Both pub. services	−0.38	−0.39	−1.60*	−0.92	−1.23	−1.45*
He industry & She pub. services	−0.84	−0.70	−1.08	−1.25	−1.12	−1.78**
He priv. services & She pub. services	−1.39*	−1.16	−0.66	−0.45	−1.10	−1.44*
Others	−1.24*	−1.23*	−0.99	−0.85	−1.58**	−0.97
Education (ref. University)						
Upper-secondary school diploma	0.73**	1.11**	0.72*	0.18	0.77**	0.88**
Compulsory	0.24	0.56	0.54	−0.04	0.75*	0.88*
Social class (ref. I+II)						
IIIa	1.29**	−0.99*	0.03	0.65	0.36	0.67
IVabc	0.90	−0.62	−0.59	0.71	0.25	−0.47
VI+VIIab	1.93**	−0.23	0.20	1.25**	0.64	−0.04
Constant	−0.73	0.37	0.43	−0.10	0.59	0.00

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$; Pseudo $R^2 = 0.05$

different activities during the day. This maximally "egalitarian" (Esping-Andersen et al. 2013) gender participation in unpaid work seems able to preserve the free time of the spouses. Couples in clusters (A) and (B) are associated with the highest synchronization levels.

These dual-earner couples are characterized by a tertiary educational level for couples in cluster (A) and a secondary educational level for those in cluster (B) (Table 3). Thus, a low educational level seems to be an obstacle to an egalitarian strategy of synchronization.

What distinguishes the two clusters is the presence of children (Table 2), which increases the probability of being a member of cluster (B), while the couples in cluster (A) are more likely to be without a child. There are also differences of occupational sector and class between the couples in clusters (A) and (B). Couples in cluster (A) work in the private services sector, while those in cluster (B) are mainly employed in the public services sector. At the same time, couples in cluster (B) are mainly employees (IIIa or VI + VIIab), while in cluster (A) they are more likely to be self-employed (I + II or IVabc).

This particular combination of characteristics—and constraints—creates synchronized couples' patterns (Fig. 1). However, there are some substantive differences. In cluster (B), the spouses seem to have breakfast together before going

Table 3 Predictive margins probability (in percentage) on the seven clusters by presence of children, sector, level of education and social class of couple. Marginal effects at reference profile*

	A	B	C	D	E	F	G
Children (ref. No child)							
Children 0–14	7.9	7.8	13.2	13.8	13.7	23.9	19.8
Children 14+	7.1	7.8	8.6	18.4	11.0	29.2	18.0
Sector (ref. both Industry)							
Both priv. services	46.3	2.4	15.4	14.1	8.4	6.4	7.0
Both pub. services	26.7	8.8	26.1	8.3	9.6	14.2	6.3
He industry & She pub. services	28.9	6.0	20.7	15.1	7.5	17	4.9
He priv. services & She pub. services	26.4	3.2	11.9	21.1	15.2	15.9	6.3
Others	30.6	4.3	12.9	17.5	11.8	11.4	11.6
Education (ref. University)							
Upper-secondary school diploma	5.9	5.8	25.8	18.7	6.4	23.1	14.3
Compulsory	7.2	4.4	18.2	19.1	6.2	27.5	17.4
Social class (ref. I + II)							
IIIa	9.0	15.6	4.8	14.2	15.5	23.4	17.5
IVabc	11.6	13.7	9.1	9.9	21.4	27.1	7.3
VI + VIIab	6.7	22.2	7.7	12.6	21.3	23.0	6.5
Reference profile[a]	12.2	5.9	17.7	18.8	11.0	22.1	12.3

[a]Reference profile: Children (No child), Sector (both Industry), Education (University), Social class (I + II)

to work and starting it synchronically. They both stop working quite early in the afternoon, favored probably by their kind of job and the economic sector in which they are employed. At 17:00 She is already at home, while He comes back at 17:40. Thereafter, both the spouses spend the rest of the day at home, doing housework and childcare before having dinner together and, finally, enjoying most of their free time synchronically. The only second part of the day in which they are not synchronized is the one immediately after dinner, when She postpones her free time for 20 min due to housework (Fig. 1).

In cluster (A) the absence of children and the type of work (self-employment in the private services sector) would seem to explain why the couples start their day in a manner differently from the others (Fig. 1). Both the spouses wake up together, and later than the couples in other clusters. They also have breakfast at the same time. Then He leaves the house while She quickly tides up before going to work. Job commitments fill equally most of their daily time. Moreover, their lunch and dinners are synchronized. Finally, the extent of their job commitments and the parallel absence of children seem to pull the spouses in cluster (A) directly to free and leisure time.[8]

[8]The absence of housework does not mean that spouses do not perform any housework. Simply, they are more likely to do it in a non-regular way, during brief and scattered moments of spare time.

Cluster (C) falls—although not completely—within the synchronized time-use patterns (Fig. 1). Couples in this cluster have some features in common with those of cluster (A). In particular, like dual-earner couples (A), those in cluster (C) are more likely to be childless. They also are mainly members of the upper class (I+II) (Table 3).

Cluster (C) has some characteristics in common with clusters (A) and (B) also in terms of daily time organization, even if its pattern ends with a longer tail of synchronized personal care: spouses may still be having dinner together at the end of the observation (22:00). However, what really makes cluster (C) unique is the time organization around lunch. While for cluster (B), there is no specific time for lunch, and for cluster (A) the time interval for lunch is well defined between two work 'segments', for cluster (C) the break from work is longer for Her. Moreover, around a certain synchronized lunch-time, there is a certain desynchronization due to His work commitments and Her housework tasks. Finally, before going back to work, She is even able to spend a short time relaxing. Here, the sequence of activity combinations around lunch is much more chaotic, fragmented and desynchronized compared with the clusters (A) and (B). However, except for this desynchronized part of the day, probably due to different work commitments, the rest of the day is mainly synchronized.

Alongside the synchronized patterns other desynchronized daily time-use patterns emerge. These strategies of desynchronization seem to be specialized into two forms, on the basis of the kind of tasks sequentially performed and combined by the two spouses during the day.

The first kind of strategy is called functionally desynchronized. Here, gender differences in activities-in-time appear to be an adaptation to structural desynchronization (Nock and Kingston 1984) of His and Her working schedules. The difference in work duration between men and women appears to produce a counterbalancing force by which—at the end of the work day—She 'compensates' the different spread of paid-work commitments of Him with unpaid work, in a quite calibrated way that preserves the free time of both the spouses. The gender division of work-family activities is "unstable" (Esping-Andersen et al. 2013), mostly due to "structural constraints" of the partners' working schedules.

Couples in clusters (D) and (E) are associated with the clearest functionally desynchronized patterns. For both clusters, Her working schedule is shifted forward in the afternoon (Fig. 1) and in most cases at least one spouse of these dual-earner couples is employed in the public sector. The most important differences between these couples is that (D) do not have a child while (E) do so (Table 2).

In cluster (D), He starts work much earlier than Her. On the other hand, She spends more time on personal care before going out to work. The probable absence of children may be helpful in this regard. At the end of the workday, these spouses come back home later and synchronically. Once at home, they desynchronize

Moreover, they may not necessarily do the housework every day, maybe postponing the chores to the weekend.

themselves again (Fig. 1) and while He takes a break to relax, She does some housework. It seems that there is some sort of compensation of daily time activities: He starts work much earlier than Her in the morning, and the gendered housework at the end of the day seems useful in establishing the balance, before dinner. Finally, they both eat and relax together (Fig. 1).

In the time-use pattern (E), She wakes up a little before Him, probably because of the young children's demands. They have breakfast together before going to work, and they start working synchronically. In the afternoon, She leaves the workplace much earlier than Him, perhaps in order to devote herself again to childcare and housework. After His return from work, they eat together, before spending synchronous free time. Again, the clear non-cooperation of Him in the household tasks may be due to the evident spread of work commitments during the whole day.

The second desynchronizations strategy is what we call traditional. Here, the couple's distribution of activities during the day does not seem to follow any compensatory mechanism. The overall desynchronization seems to be weakly linked to the "structure" of the spouses' work commitments (Nock and Kingston 1984). Conversely, it appears to be an outcome of a more "traditional" gender attitude to the work-family balance. Here, the result is a marked overload in paid/unpaid work for women (Mattingly and Blanchi 2003), with stronger evidence of the gendered leisure gap (Beblo and Robledo 2008).

Couples in clusters (F) and (G) are characterized by the presence of younger children, a low level of education, mainly compulsory level, and are mainly employed in the industrial sector. There are some differences in job features: couples in cluster (G) are mainly members of the white collar middle class (IIIa), while those in cluster (F) are mainly members of the petty bourgeoisie (IVabc). Moreover, cluster (G) shows a relatively high presence of couples where He works in the industrial sector and She in the public sector or He does so in the private sector and She in the public sector.

The time-use pattern of cluster (F) is apparently similar to that of cluster (E). In fact, She comes back home before Him and deals with domestic chores. However, compared with cluster (E) we note a greater extension of Her household commitments, from the early afternoon until the evening, when He has already finished his workhours. Thus, on one hand the desynchronization seems functional for the long time spent by Him at work; on the other hand, this couple's time-use pattern does not show any cooperative or compensatory forms of time-use organization between the spouses (Fig. 1).

Last but not least, cluster (G) is certainly the maximum expression of the traditional desynchronization. The time-use pattern (G) describes a couple in which everything is on Her shoulders. The delay of the exit from home is followed by a long journey to work. Then, she continues to work until the late afternoon. Finally, when both the spouses return home, He takes a break and rests, while She continues to do housework and child care. The only synchronized moment in the final part of this couple's pattern is when they have dinner. Among all the time-use patterns, this is certainly the one with the highest level of gender inequality in regard to the daily work-family balance challenge (Fig. 1).

5 Conclusions

In the introduction of this paper, we pointed out the importance of adopting a multi-channel sequence analysis approach to gain better understanding of the complexity of the work-family balance through holistic study of dual-earner couples' daily time use as an overall pattern. At the end of this paper, it is evident that the sequence analysis of time use diaries provides a rather clear and meaningful representation of the main patterns of the everyday organization of Italian dual-earner couples. The analysis shows the clear co-action of multiple generative mechanisms that give shape and relevance to each of seven patterns and define different forms of (de)synchronization in the everyday-life organization of both individuals and couples.

These patterns are attributable to three different strategies for organization of daily activities, and three types of equilibria (Esping-Andersen et al. 2013) within the family. In fact, these patterns describe three sets of work-family equilibrium strategies performed by dual-earner couples, with different expected levels of desirability. The first defines the synchronization strategies (clusters A, B and C). Hence, the housework division by gender is "egalitarian" because both partners participate in the housework and are able to share most of the free time available. The second defines the functional desynchronization strategies (clusters D and E). The division of housework by gender is "unstable" (Esping-Andersen et al. 2013) mostly because of "structural constraints" of the partners' work schedules (Nock and Kingston 1984). Nevertheless, the behavior of Him and Her reflects a collaborative complementarity which still tends to preserve the free time of both. The third pattern defines the traditional desynchronization strategies (clusters F and G). Partners are characterized by an unequal division of housework. They exhibit the classic features of a "traditional" equilibrium where the woman has heavy overexposure to home/child care tasks and limited free time availability (Mattingly and Blanchi 2003; Beblo and Robledo 2008).

The close relations with certain household features (the presence of children and the couple's level of education, social class and job sector) support the contention that such behaviors and patterns result, on the one hand, from the internal bargaining among each couple conditioned by the cultural-economic characteristics of the partners themselves, and, on the other, by external social constraints.[9]

The time-use patterns result from the complex co-action among individual, family and social factors whose combination defines the relevance and the shape of patterns. The time balance within His and Her activities, as well as its configuration across the day, is not random; rather, it changes according to multiple latent factors.

Dual-earner couples package their daily life mainly in accordance with their work and its schedules, and therefore mainly the type of job and the economic sector (Hamermesh 2002; Warren 2003; Lesnard 2008). Moreover, the analyses

[9]The solutions of these particular couples in daily scheduling affected the spouses' level of satisfaction as an outcome of daily life quality. For details see Bison and Scalcon (2016).

show that this time packaging changes in relation to the presence of children. We observed that the presence of children (especially young ones) introduces elements of desynchronization and specialization within the couples. The impact of young children, however, may differ according to both the couples' work schedules and their gendered attitudes to work-family activities. According to such a view, the last factor is the couples' level of education. We can assume it as a proxy for the predisposition towards egalitarian gender attitudes (Hakim 2003; Oláh et al. 2014). Not by chance, the most "egalitarian" strategies of synchronization are performed by high-educated couples, while the most "traditional" strategies of not functional desynchronization are performed by couples with a low level of education.

In conclusion, the presence of children, the level of education, and job characteristics are three dimensions that contribute to defining the patterns of couples' daily activities, already constrained by several social rhythms (i.e. school hours; lunch and dinner time; shop opening hours; etc.).

Acknowledgements The authors warmly thank the anonymous reviewers for their constructive comments.

References

Abbott, A. (1990). A primer on sequence methods. *Organization Science, 1*(4), 375–392.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research, 29*(1), 3–33.

Aisenbrey, S., & Fasang, A. (2017). The interplay of work and family trajectories over the life course: Germany and the united states in comparison. *American Journal of Sociology, 122*(5), 1448–1484.

Beblo, M., & Robledo, J. R. (2008). The wage gap and the leisure gap for double-earner couples. *Journal of Population Economics, 21*(2), 281–304.

Becker, G. S. (1964). *Human capital national bureau of economic research*. New York: National Bureau of Economic Research.

Berk, S. (1985). *The gender factory: The apportionment of work in American households*. Boston: Springer US.

Bison, I. (2009). Om matters: The interaction effects between indel and substitution costs. *Methodological Innovations Online, 4*(2), 53–67.

Bison, I. (2011). Lexicographic index: A new measurement of resemblance among sequences. In A. Bryman (Ed.), *The SAGE handbook of innovation in social research methods* (p. 422). London/Thousand Oaks: Sage.

Bison, I., & Scalcon, A. (2016). From 07.00 to 22.00: A dual-earner typical day in Italy. Old questions and new evidences from social sequence analysis. In G. Ritschard & M. Studer (Eds.), *Proceedings of the International Conference on Sequence Analysis and Related Methods*, Lausanne, June 8–10 (pp. 35–71).

Chenu, A., & Robinson, J. P. (2002). Synchonicity in the work schedules of working couples. *Monthly Labor Review, 125*(4), 55–63.

Coverman, S. (1985). Explaining husbands' participation in domestic labor. *The Sociological Quarterly, 26*(1), 81–97.

Craig, L., Powell, A., & Smyth, C. (2014). Towards intensive parenting? Changes in the composition and determinants of mothers' and fathers' time with children 1992–2006. *The British Journal of Sociology, 65*(3), 555–579.

Dijkstra, W., & Taris, T. (1995). Measuring the agreement between sequences. *Sociological Methods & Research, 24*(2), 214–231.

Elzinga, C. H. (2003). Sequence similarity. *Sociological Methods & Research, 32*(1), 3–29.

Esping-Andersen, G., Boertien, D., Bonke, J., & Gracia, P. (2013). Couple specialization in multiple equilibria. *European Sociological Review, 29*(6), 1280–1294.

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.

Gershuny, J., & Robinson, J. P. (1988). Historical changes in the household division of labor. *Demography, 25*(4), 537–552.

Hagerstrand, T. (1982). Diorama, path and project. *Tijdschrift voor economische en sociale geografie, 73*(6), 323–339.

Hakim, C. (2003). A new approach to explaining fertility patterns: Preference theory. *Population and Development Review*, *29*(3), 349–374.

Hallberg, D. (2003). Synchronous leisure, jointness and household labor supply, *Labour Economics, 10*(2), 185–203, ISSN 0927-5371 https://www.sciencedirect.com/science/article/abs/pii/S092753710300006X

Hamermesh, D. S. (2002). Timing, togetherness and time windfalls. *Journal of Population Economics, 15*(4), 601–623.

Hellgren, M. (2014). Extracting more knowledge from time diaries? *Social Indicators Research, 119*(3), 1517–1534.

Istat (2011). Indagine multiscopo sulle famiglie – Uso del tempo anno 2008–2009. manuale utente. Technical report, Istat.

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.

Makles, A. (2012). Stata tip 110: How to get the optimal k-means cluster solution. *Stata Journal: StataCorp LP, 12*(2), 347–351.

Manser, M., & Brown, M. (1980). Marriage and household decision-making: A bargaining analysis. *International Economic Review, 21*(1), 31–44.

Mansour, H., & McKinnish, T. (2014). Couples' time together: Complementarities in production versus complementarities in consumption. *Journal of Population Economics, 27*(4), 1127–1144.

Mattingly, M. J., & Blanchi, S. M. (2003). Gender differences in the quantity and quality of free time: The U.S. experience*. *Social Forces, 81*(3), 999–1030.

Naldini, M., & Saraceno, C. (2011). *Conciliare famiglia e lavoro: vecchi e nuovi patti tra sessi e generazioni*. Il Mulino, Bologna.

Nock, S. L., & Kingston, P. W. (1984). The family work day. *Journal of Marriage and Family, 46*(2), 333–343.

Oláh, L. S., Richter, R., & Kotowska, I. E. (2014). State-of-the-art report. The new roles of men and women and implications for families and societies. Families and societies. FamilesAndSocieties Working Paper Series 11, Stockholm University.

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.

Presser, H. B. (1994). Employment schedules among dual-earner spouses and the division of household labor by gender. *American Sociological Review, 59*(3), 348–364.

Raley, S., Bianchi, S. M., & Wang, W. (2012). When do fathers care? Mothers' economic contribution and fathers' involvement in child care. *American Journal of Sociology, 117*(5), 1422–1459.

Saraceno, C. (2012). *Coppie e famiglie*. Feltrinelli Editore.

Studer, M. (2018). Divisive property-based and fuzzy clustering for sequence analysis. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society, Series A, 179*(2), 481–511.

Warren, T. (2003). Classand gender-based working time? time poverty and the division of domestic labour. *Sociology, 37*(4), 733–752.

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & Society, 1*(2), 125–151.

Wu, L. L. (2000). Some comments on "sequence analysis and optimal matching methods in sociology: Review and prospect"İ. *Sociological Methods & Research, 29*(1), 41–64.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part VI
Appraising Sequence Quality

Measuring Sequence Quality

Anna Manzoni and Irma Mooi-Reci

1 Introduction: The Quality of Binary Sequences of Successes and Failures

Examination of binary sequences, that is, sequences containing only two distinct characters, each of which represents a class of separate states or events of the process observed, is of key interest in social science research. Binary sequences, which can be seen as a series of positive versus negative characters – that is, of successes versus failures – are abundant in health and applied social science research. For instance, in the study of the course of a therapy, patients may show several kinds of unfavorable reactions, the failures, as well as various kinds of positive reactions, the successes. Similarly, pupils may make various kinds of mistakes, as a result, for example, of failures of the teaching process; alternatively, they may produce correct responses, the successes of the teaching process. Yet, another example refers to individuals' labor market careers, which can be represented as the succession of favorable and unfavorable labor market states; employment and vocational training, for example, can be considered as successes, while unemployment and inactivity as failures.

The presence of successes and the absence of failures defines the quality of a sequence. Over the course of the treatment, therapy quality is higher when the patient's unfavorable reactions – the failures – gradually disappear from the gamut of observed behaviors; the teaching quality is higher when inadequate responses disappear from the pupil's repertoire; and career quality improves if

A. Manzoni (✉)
North Carolina State University, Raleigh, NC, USA
e-mail: amanzon@ncsu.edu

I. Mooi-Reci
University of Melbourne, Parkville, Australia

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_15

unemployment is overcome by finding stable employment. However, available techniques to quantify sequence quality have been lacking with existing measures focusing predominantly on the comparison of ordered sequences (Abbott 1995) or the variability of sequences (Elzinga and Liefbroer 2007). Brzinsky-Fay (2007) and Gabadinho et al. (2011) already recognize the need of such a measure and make an attempt to quantify the quality of sequences. Our study builds on and expands this knowledge by proposing a new method which distinguishes between states of differing quality, that is, states with different characteristics. We operationalize the concept of positive and negative states, which define successful and less successful sequences, and take into account the variation in the frequency, duration, and recency of successes and failures over the course of a trajectory. We interpret binary sequences as series of Successes (S) and Failures (F) and encode them as strings consisting of the characters S and F. Of course, the quality or successfulness, of a therapy may also depend on various characteristics of the therapist, the therapy, and the patient; the quality of teaching may depend on the teacher, the teaching instructions, or the pupil; the quality of a career may depend on individual characteristics of the worker, or labor market opportunities, among other factors. Hence, in many situations, it would be interesting to model the quality of a binary sequence in terms of one or more independent variables.

In Sects. 2 and 3, we briefly review existing methods to compare and describe trajectories, outline elementary requirements of our proposed measure of sequence quality, and discuss its properties. Next, we show an application of our measure to model the quality of labor market careers. In doing so, we provide a direct test of unemployment "scarring" theories that expect the number, duration, and the most recent spell of unemployment to largely determine the path of subsequent employment career quality due to human capital depreciation and signaling processes (Arulampalam 2001; Mooi-Reci and Ganzeboom 2015). Specifically, we ask to what degree unemployment is negatively associated with the quality of one's future employment career, and at what rate, if at all, previously unemployed individuals eventually recover from it. Using data from the Household, Income, and Labour Dynamics in Australia (HILDA) Survey, we investigate the evolution of employment career quality after the occurrence of an initial spell of unemployment and as a function of an individual's attributes. We conclude with a discussion of potential extensions to our measure.

2 Common Methods for Studying Sequence Trajectories

Sequence analysis has advanced rapidly in the last few decades, and several developments have been proposed, among which extensions of Optimal Matching (Abbott 1995) based on alignment techniques (e.g., Lesnard 2008, 2010; Stark and Vedres 2006; Stovel et al. 1996; Stovel and Bolan 2004), as well as non-alignment techniques (e.g., Elzinga 2003, 2010); see Studer and Ritschard (2016) for a full review. Newly developed measures, such as for example the "turbulence"

or "complexity" measures (Elzinga and Liefbroer 2007; Elzinga 2010; Gabadinho et al. 2011) have offered a quantification of the variability within rather than between sequences, but fall short in distinguishing between "good" or "bad" events over the course of a trajectory. This results in treating a series of positive events (e.g., upward job mobility) equally as a series of negative events (e.g., downward mobility), which may lead to serious substantive misinterpretations about the evolution of trajectories. Failing to attach a quality connotation to the states results in measures that ignore the variation in sequence quality.

An attempt to quantify the quality of sequences is proposed by Brzinsky-Fay (2007). The study uses data from the European Community Household Panel (ECHP) over the period 1994 to 2011 to examine sequences of school-to-work transitions of school leavers across ten European countries. It draws on explorative methods of optimal matching and cluster analysis to identify positive from negative variation of sequences where transition types with high volatility are considered negative and those with low volatility positive. While valuable, this volatility index accounts for the sequencing and variability of the transitions to infer indirectly about the nature of the transitions. However, the index suffers from the same 'old' shortcoming in which upward and negative job mobility are quantified equally because the index captures the volatility rather than the quality of the transitions. To our knowledge, one of the studies that comes closest to our measure of sequence quality is the "precarity" index proposed by Ritschard et al. (2018) in this bundle. The index draws on the complexity index as proposed by Gabadinho et al. (2011) that combines entropy with the number of transitions in the sequence and then uses a correction factor that reflects the penalizing versus rewarding quality of a transition. This correction factor is derived from the proportions of negative and positive transitions in a sequence. In doing so, the measure allows the user to determine the negative or positive transitions based on theory or data.

Different from the measure proposed by Ritschard et al. (2018) that is based on the quality of the transitions between states, our sequence quality measure is based on the quality of the states themselves and takes into account various dimensions of a particular state, such as the frequency, duration and its recency.

In the following section, we will introduce our new measure and its properties.

3 Developing a Measure of Sequence Quality: Formal Properties

Here, we introduce some notations and concepts. First, we define an alphabet, i.e., the set of states or characters that we deal with, as $A = \{F, S\}$, where F denotes a failure of some sort and S denotes a success of some kind; the specific interpretation will depend on the substantive meaning of the sequences. By concatenating the characters from the alphabet, we obtain sequences which we may denote as x, y, or z, for example. Let x denote such a sequence; then xF denotes the same

sequence, elongated with a failure F and, similarly, xS denotes the same sequence, elongated with a success S. If y is another such sequence, then xy denotes the (right-)elongation of x with y. A run is a number of consecutive successes or failures. Thus, if $x = FFSSSF$, we say that x consists of three runs: F^2, S^3, and $F^1 = F$. More generally, we say that e.g., S^n denotes a run of n successes where n is a nonnegative integer, the run-length. Nonnegative, since it is convenient to have an empty sequence $\lambda = S^0 = F^0$ that does not materially elongate any sequence: we have that $x\lambda = x = \lambda x$ for all sequences x.

These simple concepts and notations suffice to discuss some fundamental requirements that all quality measures of successfulness should adhere to. First, we require that such a measure, say $\Gamma(x)$, increases when the number of successes in a given sequence increases, independent of sequence length. Hence, we require that:

(1) For any $xy \neq \lambda$, $1 \geq \Gamma(xSy) \geq \Gamma(xy)$, equality holding precisely when $x = S^n$ and $y = S^m$ for any nonnegative n and m.

Stating that there is an upper bound of 1 is a way of saying that $\Gamma(x)$ is independent of the sequence length; however long the sequences, $\Gamma(x)$ will not exceed the value of 1. Requirement (1) also states that wherever we put an extra success into a sequence, the result will be that $\Gamma(x)$ increases, with the only exception that $\Gamma(S^n) = 1$ for all positive n. Substantively this effect can be seen as a "compensating effect" in which a positive event or state (e.g., employment) counteracts a previous negative state (e.g., unemployment or inactivity).

Our second requirement is the mirror image of the first as it pertains to the effect of failures on the value of $\Gamma(\cdot)$:

(2) For any xy, $0 \leq \Gamma(xFy) \leq \Gamma(xy)$, equality holding precisely when $x = F^n$ and $y = F^m$ for any nonnegative n and m.

Requirement (2) implies that the lower bound of $\Gamma(\cdot)$ equals zero, which is attained for any sequence that consists exclusively of failures or that contains only the empty character λ. Furthermore, requirement (2) states that the quality of a sequence diminishes when we add more failures. Hence, requirements (1) and (2) jointly imply that in a given sequence, a quality measure: (a) has a fixed range of [0, 1]; (b) increases when the number of successes increases; and (c) decreases when the number of failures increases.

Now consider the sequences $x = SSF$ and $y = FSS$. These two sequences only differ in the position of the failure F: x ends in a failure and y ends in a success. Whether x and y stood for sequences of responses of patients, pupils, or labor market states, in all cases we would consider y as the highest quality sequence because of the recency of the success. We believe this to be a general principle: the more recent the successes, the more positive the quality of the sequence. Therefore, we formulate a third requirement that precisely formalizes this principle:

(3) For any xy, $\Gamma(xFSy) > \Gamma(xSFy)$.

While frequency refers to how often states occur in a sequence, we may also be interested in "durability", which is the consecutive frequency of states or, in other words, the length of spells in a sequence. Frequency refers to the occurrence of a state, independent of whether it is part of the same or of different spells. Durability refers to state frequency within a spell. Higher durability is manifested in higher frequency of a state. However, the reverse may not be the case, as high frequency of a state may be due not only to a single durable spell, but also to several spells of short duration, in which case we will see low durability despite high frequency. For example, take the sequence $x = S^2F^2$, which includes two consecutive states of success, that is a spell of success with a duration of two units, followed by a spell of failure with duration of two units. Then take a sequence $y = S^1F^1S^1F^1$ including alternating spells of successes and failures, each of the duration of one unit, for a total frequency of two states of failures and two states of success. Sequences x and y will have the same frequency of successes. However, the durability of success differs, reaching a value of two consecutive successes in sequence x and only one in sequence y. Note that differences in duration at equal frequency will affect the recency of success. Therefore, a measure fulfilling the above requirements (1–3) and accounting for both frequency and recency will capture duration differences as well. In sequence x, consecutive success durability is higher compared to sequence y, although success is more recent in sequence y. In the next section, we will discuss an implementation of Γ that satisfies the above requirements.

4 Using *S*-Positions: Successes Weighed by Frequency and Recency

For a binary sequence x over $A = \{F, S\}$, we write $x = x_1, \ldots, x_n$ when x has n characters and x_i stands for the i-th position in x. Thus $x_i \in A$ for all $i \in \{1, \ldots, n\}$. For example, with $x = SSFSF$, $x_1 = x_2 = x_4 = S$ and $x_3 = x_5 = F$. The length of $x = x_1, \ldots, x_n$ equals n; we denote this by writing $|x| = n$. For a sequence x with $|x| = n$, the k-th prefix of x is the sequence $x^k = x_1, \ldots, x_k$ for $0 \leq k \leq n$ and $x^0 = \lambda$.

We begin by discussing a simple example. Consider the sequence: $x = x_1x_2x_3x_4x_5 = SSFSF$. First, we note that the first, second and fourth characters are successes. Then, we add the position-indices of these characters: $1 + 2 + 4 = 7$. If these three successes had occurred later in the sequence, as for example in the sequence $FFSSS$, the sum of the position indices would have been bigger: $3 + 4 + 5 = 12$. Therefore, we see that the sum of the position indices of the S-observations quantifies the quality level: the more S-observations and/or the more recent these are, the bigger the sum will be. However, we cannot judge the size of this sum independently of the length of the sequence, since longer sequences will have larger position indices. To adequately quantify quality, we divide the sum of the observed S-positions by the sum of all positions and we denote this ratio as Υ to distinguish it from the general $Gamma$. Hence, we obtain:

$$\Upsilon(1+2+4)/(1+2+3+4+5) = 7/15 = 0.47.$$

Clearly, this ratio is always in the range [0, 1] as the numerator is non-negative and at most as big as the denominator. Formally and more generally, it is convenient to first define a position variable:

$$p_i = \begin{cases} i & \text{if } x_i = S \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

and then define:

$$0 \leq \Upsilon^w(x^n) = \frac{\sum_{i=1}^n p_i^w}{\sum_{i=1}^n i^w} \leq 1 \tag{2}$$

for some nonnegative exponent w. Note that in the above example, we set $w = 1$. Clearly, Υ^1 quantifies sequence quality in the required way: the more failures and the later in the sequence, the smaller its numerator. As the numerator of Eq. 2 is nonnegative and cannot exceed the size of the denominator, Υ^w will be tightly bound by 0 and 1; the upper bound will be attained when all states are S-states, while $\Upsilon^w = 0$ when the sequence shows no successes at all. Table 1 illustrates the behavior of Υ^w. Specifically, the left part of the table illustrates the behavior of Υ^w when failures get more recent for three different values of the parameter w; the right part illustrates the combined effect of recency and number of failures on Υ^w.

Let us now turn to the meaning of the parameter w in Eq. 2. First, let us set $w = 0$. The denominator then reduces to $\sum_{i=1}^n i^0 = \sum_{i=1}^n 1 = n$ and the numerator counts the number of S-states, regardless of their position in the sequence. Hence, we see that:

$$\Upsilon^0(x^n) = \frac{f(S)}{n}$$

i.e., the measure calculates the fraction of S-states in the sequence x^n.

Table 1 Illustration of Υ^w behavior with varying w and recency of failure

	$\Upsilon(x)$				$\Upsilon(x)$		
	$w = 0.5$	$w = 1$	$w = 2$		$w = 0.5$	$w = 1$	$w = 2$
FFFSSS	0.62	0.71	0.85	SSSS	1	1	1
FFSFSS	0.59	0.67	0.77	SFSS	0.77	0.8	0.87
FSFFSS	0.56	0.62	0.71	SSFS	0.72	0.7	0.7
SFFFSS	0.53	0.57	0.68	SSSF	0.67	0.6	0.47
SFFSFS	0.5	0.52	0.58	FSSF	0.51	0.5	0.43
SFSFFS	0.48	0.48	0.51	SFSF	0.44	0.4	0.33
SSFFFS	0.45	0.43	0.45	SSFF	0.39	0.3	0.17
SSFFSF	0.43	0.38	0.33	FSFF	0.23	0.2	0.13
SSFSFF	0.41	0.33	0.23	SFFF	0.16	0.1	0.03
SSSFFF	0.38	0.29	0.15	FFFF	0	0	0

Next, we consider:

$$\Upsilon^1(x^n) = \frac{\sum_i p_i}{\sum_i i} = \frac{\sum_i p_i}{n(n+1)/2}$$

which is simply a formalization of the example above.

In Table 1, we illustrate how $\Upsilon^w(x)$ reacts to changes in the recency of failures and to changes in the number of failures. To demonstrate the effect of the parameter w, Fig. 1 illustrates the dynamic behavior of Υ^w for different values of w.

Specifically, we plot Υ^w for various values of w applied to a fixed sequence $x = S^4F^4S^6F^2S^8F^1S^{25}$ of length $|x| = 50$. To draw Fig. 1, we calculate, for each fixed value of w, $\Upsilon^w(x^n)$ for $n = 1, 2, \ldots, 50$. Hence, the plots show how Υ^w develops "over time", as the sequence develops from a single state to a sequence of 50 states. The sequence chosen has three runs of failures, each next run shorter than the previous one. The plots demonstrate that the bigger the parameter w, the more severe is the effect of failures, but recovery from the failures due to subsequent successes is also faster for bigger w. This holds independent of the order of success and failure runs and their length. Moreover, the plots suggest that, given enough time and enough subsequent successes, the sequences will "fully recover" from previous failures. Here, "full recovery" means that Υ will be arbitrarily close to 1. Formally, this can be seen from the fact that:

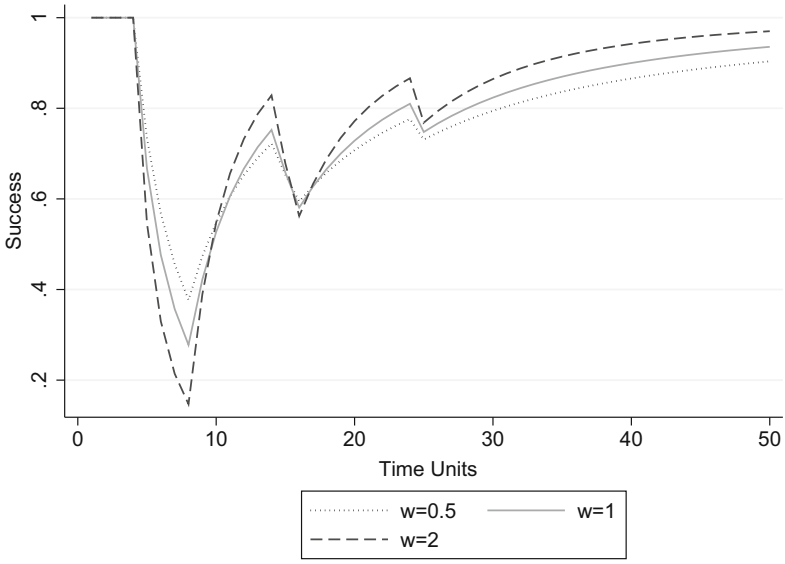

Fig. 1 Plot of $\Upsilon^w(x^n)$ (vertical axis), for various values of w and n (horizontal axis) ranging from 1 to 50, and $x = S^4F^4S^6F^2S^8F^1S^{25}$

$$\Upsilon^w(x) = \frac{\sum_i i^w - S_F}{\sum_i i^w} = 1 - \frac{S_F}{\sum_i i^w} \tag{3}$$

wherein S_F denotes the sum of (the powers of) the positions on which failures occurred: given enough subsequent successes after the last failure, the fraction will converge to 1 since the quantity $\sum_i i^w$ will increase while the quantity S_F will remain constant. Full recovery from failures is a phenomenon that happens, for example, in labor market careers of former criminals: after enough clean time ("Success") after conviction ("Failure"), their chances of getting employed are equal to those never arrested (Blumstein and Nakamura 2009).

Calculating the quantity Υ^w is straightforward; for the denominator of Υ^w, closed expressions can be derived and w does not have to be an integer (Knuth 1993). However, in applications, such closed expressions are not required since the denominator is easily calculated simultaneously with the numerator while parsing the sequence. Thus, calculating Υ^w is proportional to the length n of the sequences. The reader is aware that Υ^w as defined in Eq. 2 is just one of the many possibilities to construct a quality measure on the basis of the position-numbers p_i: in fact, for any function f that satisfies $f(a) > f(b)$ precisely when $a > b$, the quantity

$$0 \leq \Gamma(x^n) = \frac{\sum_{i=1}^n f(p_i)}{\sum_{i=1}^n f(p_i)} \leq 1$$

is a quality measure, i.e., satisfies the requirements (1) to (3) as formulated in the previous section. We believe that, by picking $f(p_i) = p_i^w$, we have covered sufficient potentially useful possibilities.

Theoretical reasons may lead a user to define a w of 0 and get an index that does not correct for the recency of failures or successes. Alternatively, a w of 1 quantifies sequence quality in the required way: the more failures and the later in the sequence, the smaller its numerator. However, should theory assign deeper penalties to failures and faster recoveries to most recent successes, then a $w > 1$ would be preferable.

5 An Application: The Quality of Labor Market Careers Among the Unemployed

In this section, we apply Υ^w to labor market careers in order to test whether unemployment has long-term negative effects on people's career outcomes, also termed in the literature as unemployment "scarring" (Arulampalam 2001; Mooi-Reci and Ganzeboom 2015). Two prominent theoretical mechanisms have been suggested to drive unemployment scarring. First, from a human capital perspective, loss and depreciation of firm and occupation specific skills over a spell of unemployment is thought to make workers less productive. This translates into job offers of a poor quality (e.g., non-standard or temporary type of jobs) that are more likely

to be discontinued and result in recurrent unemployment spells in the future. This implies that the longer the elapsed time in unemployment the lower the chances of reemployment will be.

Signalling has also been suggested as an alternative mechanism driving scarring (Spence 1973). When employers have little information about one's qualifications or abilities they use a range of observable characteristics to infer an applicant's productive capabilities. Characteristics that are subject to individuals' own actions, such as previous unemployment spells, are likely to raise red flags about a worker's unobserved quality and the tasks that they are expected to perform. The views that employers hold of applicants with previous unemployment spells are most likely to translate into lower reemployment rates and wages for some age groups (Mooi-Reci and Muñoz-Comet 2016; Mooi-Reci and Wooden 2017; Pedulla 2016). However, not all "red flags" related to previous unemployment are of equal weight. When the reason for unemployment is believed to relate directly to a worker's poor performance with the previous employer, chances of reemployment will be much lower for reasons related to unobserved worker quality. Additionally, various studies provide considerable support that unemployment spells experienced far in the past – and thus no longer relevant to the current application – or during an economic downturn tend to carry less of a negative weight compared to spells experienced directly prior to the job application (Mooi-Reci and Ganzeboom 2015; Omori 1997). This means that more recent spells of unemployment will influence more negatively one's employment prospects.

These two potential scarring mechanisms raise the following two-fold question: if unemployment is negatively related to one's future career outcomes, do previously unemployed individuals eventually recover from it, and if so, at what rate? This question will guide our analyses in the next sections.

5.1 Data

We use the Household, Income, and Labour Dynamics in Australia (HILDA) Survey, one of the richest and longest running household longitudinal survey in Australia (Summerfield et al. 2015). The HILDA survey commenced in 2001 and data is collected every year among household members 15 years and over. In 2001, the initial sample counted 13,969 persons across 7,682 households, with the main sample remaining about the same in the subsequent data collection waves. The wave-by-wave response rates are exceptionally high in the HILDA survey, with retention rates varying between 86.9% in wave 2 to 96.4% in 2013. The retention rate of the first sample was 67.4% in 2013 (i.e. this is the number of people responding in both the initial wave and the 2013 wave), which is noticeably high. Every year respondents are asked to identify their labor market status at the time of interview, that we recoded into: employment (E), unemployment (U), non-participation/inactivity (N) and retirement (R) (see Summerfield et al. 2015 for a complete list of the questionnaires). Since the focus of this study is on career

sequences, we have restricted the sample to those men and women who are not in full-time education and for whom information on labor force status is available at any interview date over the period 2001–2013. This selection leaves us with a sample of 22,081 person-year observations. Of these 22,081 observations, by far the largest concentration (i.e. 44.42%) is among the group for which we have complete labor force status information for the entire observation period between 2001 and 2013. All the analyses are performed using Stata 14.[1]

Measures: Our dependent variable, *binary sequence quality*, is a time-varying measure defined as discussed in Sects. 3 and 4 and captures the quality of career sequences starting from a first unemployment experience up until t. To investigate the evolution of employment careers after the occurrence of an initial spell of unemployment, we define a variable counting the years elapsed since the first unemployment experience (i.e. our reference category). As we observe respondents up to 12 years following a first unemployment experience, we construct 12 dummies capturing each of the years in which a respondent was observed in any of the labor market statuses. These time-specific dummies allow us to trace the rate of recovery since the initial episode of unemployment.

We also control for socio-demographic characteristics, including variables for age and age squared, and account for human capital with a variable capturing the highest attained educational level at the time of interview. Education is specified as a categorical variable with Year 11 and below (i.e., early childhood education and primary school) as the reference category and six additional categories for Year 12 (i.e., lower secondary school), Cert III or IV (upper secondary school), Advanced diploma (i.e., post-secondary non-tertiary education), Bachelor or honors, Graduate diploma, Postgraduate education. We also include a variable for Gross Domestic Product per capita as proxy of economic growth in a specific year (GDP). Finally, to guard against the possibility that observed career outcomes are driven by career fluctuations that existed prior to the first unemployment (e.g., due to periods of inactivity), we include a variable specifying the career quality before the first unemployment experience in our models. Table 2 shows descriptive statistics for our person-year sample.

5.2 Method

A key statistical challenge for our analysis is non-random selection into the initial employment state that is correlated with unobservable traits. To solve this issue, we estimate a model with correlated random effects, also known as the 'hybrid' model (Allison 2009). The hybrid model allows time-varying covariates to be decomposed into individual specific means and deviations from these individual-specific means.

[1]The `sqsuccess` package in Stata implements our quality measure as described in this paper. We thank Ulrich Kohler for developing it.

Table 2 Summary descriptive statistics

	Men		Women	
	Mean	SD	Mean	SD
Career quality				
$w = 1$	0.44	0.39	0.40	0.38
$w = 0.5$	0.41	0.37	0.38	0.36
$w = 2$	0.47	0.43	0.43	0.42
Career quality before unemployment				
$w = 1$	0.35	0.45	0.31	0.42
$w = 0.5$	0.62	0.43	0.47	0.43
$w = 2$	0.62	0.45	0.47	0.45
Age	35	14.74	36	13.99
GDP	1.5	0.67	1.5	0.68
	Percent		Percent	
Education				
Year 11 or less	35%		36%	
Year 12	21%		20%	
Certificate III or IV	20%		19%	
Advanced diploma	7%		7%	
Bachelor or honors	9%		11%	
Graduate diploma	2%		3%	
Postgraduate	3%		3%	
N observations	11,106		10,975	
n workers	2,137		2,026	

The advantage of the hybrid model is that it corrects for unobserved heterogeneity across the time-varying variables and allows for inclusion of time-constant variables which otherwise would have been dropped from a fixed effect specification (Allison 2009). The model takes the form:

$$\Upsilon_{it}^{w} = \alpha + \beta \Upsilon_{\text{BEFORE } i}^{w} + D(d_{it} - \bar{d}_i) + \gamma \bar{d}_i + \delta(x_{it} - \bar{x}_i) + \eta \bar{x}_i + \mu_i + \varepsilon_{it} \quad (4)$$

where Υ_{it}^{w} is the measure for sequence quality of worker i in year t (which covers the period 2001 to 2013); $\Upsilon_{\text{BEFORE } i}^{w}$ controls for the career quality of individual i before the time of the first unemployment occurrence and β captures its effect. D is a vector of coefficients associated with the deviations of the specific time dummies from the overall time cluster mean denoted in the bracket $(d_{it} - \bar{d}_i)$, which gives us the within-effect estimates. In addition to these time deviations, the model adds a vector of coefficients (γ) associated with the time cluster mean $(\bar{d}_i)$ to control for dependency of the repeated observations. Next, within-effect estimates for individual characteristics that are supposed to influence the overall sequence quality are captured through δ, while $(x_{it} - \bar{x}_i)$ refers to the difference between time-varying variables expected to be associated with one's sequence quality (including

age, education, and GDP) and their individual specific cluster means. Next, a vector of coefficients (η) associated with the individual specific cluster means $\bar{x}_i$ is added to the model. As described by Allison (2009) and later by Schunck (2013) this addition is necessary to estimate the evolution of the variations around the mean and to control for correlations between level 1 (i.e., respondents) and level 2 errors (i.e., occasions). Note that interpretation of cluster means is not of interest because we aim at predicting career quality based on the variations of individual characteristics around the mean, which in essence, mimics the logic behind fixed effect models. Finally, μ_i refers to the individual specific error (i.e., level 2 error) with ε_{it} referring to the level 1 error. To capture sex-specific unemployment effects, we estimate separate models for men and women.

5.3 Findings

Table 3 shows, separately for men and women, the estimated coefficients from three hybrid models that are based on three weight specifications of our career quality measure, that is: $w = 1$, $w = 0.5$ and $w = 2$. In each model, the year coefficients show how the rate of career quality of previously unemployed workers evolves in each year following unemployment. To ease the interpretation, Fig. 2 plots the key results from such models. As we clarified above, the weight parameter determines the extent to which the quality measure is affected by a failure, with stronger penalties for failures, but faster recovery for successes the bigger the parameter w.

Although coefficient estimates from both individual-specific means and their deviations are shown in Table 3, only coefficients from deviation specific means are used for interpretation, because coefficients pertaining to individual specific means have no substantive interpretation (see above and for a review, see Allison 2009).

The large positive coefficient estimates for the year dummies in all three models suggest that, for both men and women, there is a trend of recovery. Specifically, using a w of 1 among the male sub-sample in Model 1, coefficient estimates indicate an improvement in the career quality with 0.36 points in the first year following the initial unemployment spell. This improvement in the career quality continues in the second and third year with 0.49 and 0.55 points, respectively, before flattening nine years after the first unemployment spell. We find a similar progressively improving trend among women. Interestingly, in the longer run women's patterns of recovery exceed men's. However, from the confidence intervals shown in Fig. 2 (and from tests of interactions, not shown but available upon request) we can establish that gender differences are not statistically significant. Other model specifications with different weights follow expected trends of recovery: using a weight of 0.5 (Model 2) coefficient estimates are lower, indicating less penalizing failures but also slower recovery, while weights of 2 show faster recovery in the years following first unemployment. As mentioned earlier, the choice of the w estimator is determined by theory and varies depending on the purpose of research. All other parameters included in the model act in the expected direction. Age, which can be interpreted as a proxy for experience, improves the career quality, which is also positively

Table 3 Coefficients from hybrid models predicting career quality (Y_{it}) over time since first unemployment, by sex

	Men			Women		
	Model 1 ($w = 1$)	Model 2 ($w = 0.5$)	Model 3 ($w = 2$)	Model 1 ($w = 1$)	Model 2 ($w = 0.5$)	Model 3 ($w = 2$)
Deviations						
Years since first unemployment						
1	0.364**	0.319**	0.437**	0.349**	0.307**	0.416**
2	0.489**	0.444**	0.550**	0.466**	0.424**	0.520**
3	0.549**	0.508**	0.596**	0.531**	0.492**	0.573**
4	0.587**	0.550**	0.624**	0.569**	0.535**	0.601**
5	0.616**	0.583**	0.646**	0.602**	0.571**	0.628**
6	0.634**	0.604**	0.659**	0.623**	0.595**	0.643**
7	0.653**	0.626**	0.673**	0.645**	0.620**	0.662**
8	0.669**	0.646**	0.685**	0.672**	0.648**	0.688**
9	0.678**	0.658**	0.687**	0.699**	0.675**	0.715**
10	0.688**	0.670**	0.694**	0.715**	0.692**	0.728**
11	0.688**	0.674**	0.689**	0.734**	0.711**	0.744**
12	0.687**	0.678**	0.681**	0.728**	0.711**	0.728**
Education (ref: year 11 or less)						
Year 12	0.047**	0.034*	0.069**	0.093**	0.074**	0.124**
Certificate III or IV	0.019	0.012	0.031	0.118**	0.101**	0.144**
Advanced diploma	0.072*	0.056*	0.099**	0.172**	0.151**	0.200**
Bachelor or honors	0.094**	0.075**	0.126**	0.208**	0.181**	0.246**
Graduate diploma	0.067*	0.056	0.087*	0.192**	0.166**	0.231**
Postgraduate	0.130*	0.108*	0.163*	0.241**	0.212**	0.286**
Age	0.024**	0.023**	0.025**	−0.006	−0.006	−0.005
Age squared	−0.000**	−0.000**	−0.000**	−0.000*	−0.000*	−0.000*
GDP	−0.001	−0.001	0.001	0.003	0.002	0.005
Means						
Years since first unemployment						
1	0.320**	0.280**	0.386**	0.324**	0.283**	0.390**
2	0.529**	0.479**	0.598**	0.419**	0.382**	0.468**
3	0.661**	0.612**	0.716**	0.573**	0.531**	0.621**
4	0.433**	0.406**	0.460**	0.53**	0.492**	0.572**
5	0.811**	0.759**	0.862**	0.432*	0.405*	0.458*
6	0.377	0.366	0.380	0.757**	0.734**	0.768**
7	0.631*	0.618*	0.634*	0.381	0.376	0.373
8	1.211**	1.126**	1.311**	0.898**	0.844**	0.956**
9	0.360**	0.352	0.355	0.773*	0.722*	0.831*
10	0.645**	0.607	0.690	0.170	0.193	0.127
11	0.745**	0.726	0.749	0.192	0.167	0.236
12	0.760**	0.728	0.787	0.390	0.404	0.353

(continued)

Table 3 (continued)

	Men			Women		
	Model 1 ($w = 1$)	Model 2 ($w = 0.5$)	Model 3 ($w = 2$)	Model 1 ($w = 1$)	Model 2 ($w = 0.5$)	Model 3 ($w = 2$)
Education (ref: year 11 or less)						
Year 12	0.061**	0.056**	0.068**	0.037	0.033	0.041
Certificate III or IV	0.059**	0.053**	0.067**	0.063**	0.057**	0.073**
Advanced Diploma	0.106**	0.098**	0.117**	0.125**	0.112**	0.143**
Bachelor or honors	0.118**	0.109**	0.130**	0.111**	0.101**	0.125**
Graduate dipl	0.070	0.066	0.076	0.144**	0.134**	0.158**
Postgraduate	0.067*	0.063*	0.073*	0.089*	0.084*	0.096*
Age	0.011**	0.010**	0.012**	0.011**	0.010**	0.013**
Age squared	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
GDP	0.037*	0.033*	0.043*	0.034*	0.031*	0.039*
Other control variables						
Career quality before unemployment						
$w = 1$	0.138**			0.135**		
$w = .5$		0.130**			0.128**	
$w = 2$			0.146**			0.140**
Intercept	−0.271**	−0.247**	−0.304**	−0.285**	−0.255**	−0.326**
R-squared-between	0.541	0.543	0.533	0.424	0.427	0.415
R-squared within	0.583	0.616	0.526	0.509	0.545	0.451
Sigma u	0.166	0.159	0.171	0.188	0.179	0.196
Sigma e	0.183	0.162	0.218	0.191	0.168	0.228
Rho	0.450	0.490	0.381	0.492	0.532	0.425
Chi-squared	6283.534	6153.035	6443.005	4654.766	4643.767	4657.001
N	11,106	11,106	11,106	10,975	10,975	10,975

$** = p < 0.01; * = p < 0.05$

associated with higher levels of education and economic growths; that is, the higher the education level and economic prosperity the higher the probability for previously unemployed men and women to find re-employment. Finally, a high level of career quality before the first observed unemployment is related to higher career quality later in careers.

6 Conclusion and Discussion

In this study, we proposed a novel measure of sequence quality, which differs from existing methods of sequence analysis in that it is the first to quantify sequence quality by accounting for the variation in the frequency, duration and recency of failures and successes over the course of a trajectory. The possibility to differently weigh such measure also allows researchers the flexibility to adjust it based on their theoretical considerations. While quality may encapsulate more than positive or negative states, our measure captures a major dimension of quality.

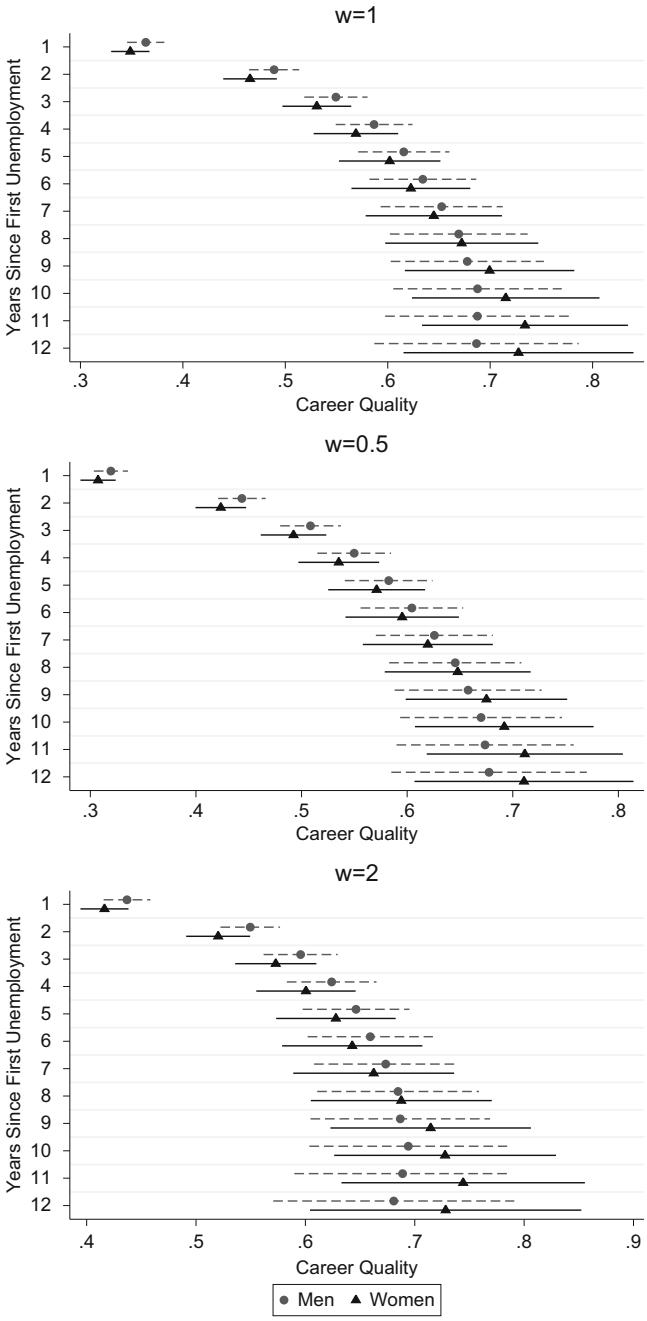

Fig. 2 Post unemployment career quality since first unemployment. Coefficient estimates from hybrid models with different weights, by sex

Drawing on theories about unemployment scarring, we defined states of unemployment and inactivity as failures and those of employment as successes to predict whether prior unemployment lead into descending spirals into inactivity and joblessness or whether patterns of full career recovery exist. We used data from the HILDA survey in Australia over the period 2001–2013 to illustrate the usefulness of our measure and investigate whether unemployment has adverse effects on future careers. Our results reveal no full career recovery among previously unemployed men and women, even over an extended period of 12 years. These results can be explained in the light of human capital depreciation and job mismatching, but signaling-related factors may be equally important in driving labor market disparities.

Some limitations with regard to the application of our measure should be mentioned. First, we applied our measure to the labor force status reported at the time of interview. Though labor force information at the time of interview is less likely influenced by inaccuracy and recall errors than retrospective types of data, it misses important information about labor force changes that occurred in-between the interview dates. Therefore, the measure of failure may be underestimated if workers experienced unemployment or inactivity spells in-between interviews. Further, the data used here are illustrative of binary sequences in which we distinguish between four possible crude labor force states. With more detailed data on people's careers it is possible to study more fine-grained labor force outcomes such as job-to-job changes in the same hierarchical position or moves to better or worse positions. Finally, an interesting issue for future research is to move beyond binary sequences to include more categories and variation in the sequences. Not all outcomes are binary. People can have neutral outcomes or outcomes that are more continuous. Therefore, quantifying the quality of categorical or continuous types of outcomes remains an important avenue for future research.

Acknowledgements Professor Cees Elzinga has highly contributed to writing the methods part of the manuscript and to developing the algorithms that we use in this. Ulrich Kohler developed the ado file to implement our success measure in Stata. We would also like to thank Professor Tim Liao for reading and commenting on an earlier version of this manuscript, as well as the participants to the LaCOSA II conference and two anonymous reviewers for valuable comments.

This work was supported (partially) by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP160101063). It uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Survey Project was initiated and is funded by the Australian Government Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the authors and should not be attributed to either DSS or the Melbourne Institute.

References

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21*, 93–113.

Allison, P. D. (2009). *Fixed effects regression models* (Vol. 160). London: Sage Publications, Inc.

Arulampalam, W. (2001). Is unemployment really scarring? Effects of unemployment experiences on wages. *The Economic Journal, 111*(475), 585–606.

Blumstein, A., & Nakamura, K. (2009). Redemption in the presence of widespread criminal background checks. *Criminology, 47*(2), 327–359.

Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review, 23*(4), 409–422.

Elzinga, C. H. (2003). Sequence similarity. *Sociological Methods & Research, 32*(1), 3–29.

Elzinga, C. H. (2010). Complexity of categorical time series. *Sociological Methods & Research, 38*(3), 463–481.

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, 23*(3–4), 225–250.

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.

Knuth, Donald. (1993). Johann Faulhaber and sums of powers. *Mathematics of Computation* 61(203): 277–294, Reprinted as Chapter 4 in D. Knuth. 2003 *Selected Papers on Discrete Mathematics*, CSLI, Leland Stanford Junior University.

Lesnard, L. (2008). Off-scheduling within dua-earner couples: An unequal and negative externality for family time. *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.

Mooi-Reci, I., & Ganzeboom, H. B. (2015). Unemployment scarring by gender: Human capital depreciation or stigmatization? Longitudinal evidence from the Netherlands, 1980–2000. *Social Science Research, 52*, 642–658.

Mooi-Reci, I., & Muñoz-Comet, J. (2016). The great recession and the immigrant-native gap in job loss in the Spanish labour market. *European Sociological Review, 32*(6), 730–751.

Mooi-Reci, I., & Wooden, M. (2017). Casual employment and long-term wage outcomes. *Human Relations, 70*(9), 1064–1090.

Omori, Y. (1997). Stigma effects of nonemployment. *Economic Inquiry, 35*(2), 394–416.

Pedulla, D. S. (2016). Penalized or protected? Gender and the consequences of nonstandard and mismatched employment histories. *American Sociological Review, 81*(2), 262–289.

Ritschard, G., Bussi, M., Reilly, J. O. (2018). An index of precarity for measuring early employment insecurity. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *Stata Journal, 13*(1), 65–76.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics, 87*(3), 355.

Stark, D., & Vedres, B. (2006). Social times of network spaces: Network sequences and foreign investment in Hungary. *American Journal of Sociology, 111*(5), 1367–1411.

Stovel, K., & Bolan, M. (2004). Residential trajectories: Using optimal alignment to reveal the structure of residential mobility. *Sociological Methods & Research, 32*(4), 559–598.

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. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society, Series A, 179*(2), 481–511.

Summerfield, M., Dunn, R., Freidin, S., Hahn, M., Ittak, P., Kecmanovic, M., Li, N., Macalalad, N., Watson, N., & Wilkins, R. (2015). HILDA User Manual – Release 14. *HILDA Use Manual – Release, 15*, 1–167.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

An Index of Precarity for Measuring Early Employment Insecurity

Gilbert Ritschard, Margherita Bussi, and Jacqueline O'Reilly

1 Introduction

Young people have found it increasingly difficult to access work or have cycled between precarious employment, inactivity or unemployment since the stark increases in youth unemployment following on from the financial crisis of 2008 (O'Reilly et al. 2015). Therefore policy approaches have sought to address the consequences of social exclusion and "scarring effects" caused by precarious trajectories (Bell and Blanchflower 2011; Gregg and Tominey 2005; Tumino 2015). In the light of this concern, we aim at developing an index of precarity that allows us to assess the quality of early employment trajectories. At the same time, this index allows us to investigate to what extent the quality of early employment trajectories is linked with future youth employment outcomes.

A vast body of literature acknowledges the increased precarity experienced by young people in the labour market; however, there is a lack of quantitative tools able to grasp the complexity of trajectories and, at the same time, assess their effects in the long run. Filling this gap implies developing a tool that evaluates quantitatively the desirability of employment trajectories and adapts to the longitudinal feature of life-course analysis. We suggest using a modified version of the original "index of complexity" elaborated by Gabadinho et al. (2010), by applying weights to differentiate between the value of labour market transitions.

G. Ritschard (✉)
NCCR LIVES and Geneva School of Social Sciences, University of Geneva, Geneva, Switzerland
e-mail: gilbert.ritschard@unige.ch

M. Bussi
University of Louvain, Louvain-la-Neuve, Belgium

J. O'Reilly
University of Sussex, Brighton, UK

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_16

The added value of the index of precarity is to assess the degree of insecurity of employment trajectories experienced by individuals. Furthermore, it can be used to predict, together with other covariates, to what extent the precarity embedded in complex employment trajectories has an impact on the type of labour market positions in the future. Another contribution of our index is its transferability: it can be used across ages, social groups and countries and for any transition across relevant states whose succession leads to a cumulative (dis-)advantage.

This contribution shortly presents the social challenges faced by young people in their first labour market experiences. Then it presents how the index is constructed and justifies this choice. We finally calculate the index using a dataset that comes from a study by McVicar and Anyadike-Danes (2002) on transition from school to work using the Status Zero Survey 2000. The dataset includes a time series sequence of 72 monthly labour market activities for each of 712 young people in a cohort survey. These young people, living in Northern Ireland were followed up from July 1993 to June 1999. We find that young people with precarious trajectories in the first three years after leaving school have dramatically higher risk of experiencing negative labour market positions two years later.

2 Rising Precarity Among Young People

Changes in the skills structure and the education provision, new forms of employment and the recent economic crisis have contributed to making youth first transitions in the labour market more unstable, complex and individualised (Mills and Blossfeld 2005; Kalleberg 2009; Smithson and Lewis 2000; Gardiner 2016). These changes, crystallised in the multifaceted phenomenon called globalisation, affect young people's first transitions directly and indirectly. Directly these have increased competition among workers and imposing a fast-changing technological environment. Indirectly the effects have affected institutions that shape young people's first employment transitions such as education and welfare systems, employment relations and family structures.

These changes have turned out to be more detrimental for those groups in the labour market with low skills and educational attainment and more likely to be discriminated, such as young people, women, ethnic minorities and migrants. Moreover young people's low negotiating power is worsened by the lack of seniority or work experience (Mills and Blossfeld 2005). According to Standing (1999) these groups have been increasingly experiencing a process of churning between various forms of employment precarity and inactivity.

Protracted experiences of precarious employment—i.e. "employment that is uncertain, unpredictable, and risky from the point of view of the worker" (Kalleberg 2009) and often of poor quality—and joblessness have become more common (Gebel 2010; Ortiz 2010; Worth 2005; Scherer 2001). These contribute to add further precarity when individuals are expected to make critical decisions that shape their life-course (Mills and Blossfeld 2005).

Volatility of employment trajectories can result in stigmatisation, negative signalling to employers. In the long run, this is likely to entail a "scarring" effect. This means that past negative and precarious experiences in the labour market can have long-term negative consequences in terms of repeated periods of unemployment, lower wages and lower levels of human capital attainment over the life cycle (Ayllón 2013; Manzoni and Mooi-Reci 2011; Cable et al. 2008; Schmelzer 2011; Weich and Lewis 1998), but also in terms of well-being (Daly and Delaney 2013).

Rising levels of precarious forms of employment as typified in the UK by zero hours contracts where working hours are not guaranteed (Bussi and O'Reilly 2016) or mini-jobs in Germany (Palier and Thelen 2010) are a reflection of this increasingly fragmented and fragile labour market. In these circumstances Schmid and Schömann (2004) have argued that it is more useful to think of (early-) employment insecurity rather than of (early-) job insecurity. Employment insecurity is better at capturing how precarity affects labour market integration and how individual negotiated choices are embedded in multiple changes and transitions over time. Employment security can involve changing employer and job but maintaining an employment relationship and stable income (Chung and Van Oorschot 2011). Schmid (2015) argues that the welfare state needs to actively contribute to secure employment over the life course rather than a focus on job security. Employment security requires an investment in skills to reinforce employability and access to work (Muffels and Luijkx 2008).

On the basis of this body of literature, we propose to develop an index of precarity to capture a range of labour market precarity comprehensively and test the hypothesis that young persons will be unemployed, inactive or in temporary employment in latter periods if they have a trajectory dominated by non-employment or precarious employment.

3 Conceptualising Precarity

Measuring precarity and its impact is not straightforward. Barbier (2005) explains that the experience of employment precarity does not translate in the same way across institutional and cultural contexts. Most of the existing literature has focused on different aspects of precarious experiences in the labour market: among others the quality of jobs (Leschke and Watt 2008); the capacity of precarious work contracts acting as bridges towards more stable jobs (Booth et al. 2002; Scherer 2004; de Graaf-Zijl et al. 2011; Cockx and Picchio 2012); and the impact of precarious work on health and well-being. Furthermore, studies on the impact of spells of inactivity and unemployment have mostly looked at single spells, or cumulative spells of unemployment/inactivity. They have rarely assessed the disadvantage derived by a succession of negative spells and downward changes in the labour market, i.e. of a protracted precarity.

Our index aims to provide a quantitative tool accounting for the cumulative process of precarity and its impact. Different summaries of a sequence more or less related to employment precarity can be found in the literature. The diversity

of the states visited, i.e. here of the labour market positions experienced, is often considered as an indicator of the uncertainty—in the sense of unpredictability—of the trajectory. It can be measured, for example, by the entropy (lack of predictability) of the distribution of the states within the sequence or by the inverse of the variance of the time spent in the successive distinct states. However, such measures do not account for the sequencing, i.e. order of the states. For example, the sequences $FFUU$ and $FUFU$ have same entropy but the order of states is different. At least two composite measures that combine a diversity indicator with something sensitive to this sequencing have been proposed in the literature. The turbulence index of Elzinga and Liefbroer (2007) combines the inverse of the variance of the durations in the distinct successive states with the number of subsequences that can be extracted from the sequence of distinct successive states. The complexity index of Gabadinho et al. (2010) combines entropy with the number of transitions in the sequence.

These indexes are intended to measure the unpredictability or instability of sequences, but this is done without taking the nature of the states into account. For example, letting E, W, and U stand for education, work, and unemployment, the sequences $EEWWW$ and $EEUUU$ would get the same entropy, turbulence and complexity values, while the second is evidently a more precarious trajectory.

Although sequence instability as measured by the turbulence or complexity index certainly contributes to understanding the precarity of a sequence, a precarity indicator has to account for the nature of the states that constitute the trajectory. A first simple solution is to distinguish between positive (e.g. employed or in education) and negative (e.g. unemployed) states like in the volatility indicator of Brzinsky-Fay (2007). This indicator is defined as the ratio between the number of positive and negative positions in the sequence. Based on this same distinction between positive (success) and negative (failure) states, Manzoni and Mooi-Reci (2018) propose a refined solution where precarity increases with the recency of failures. However, these two solutions do not explicitly account for the instability of the sequence.

A general approach to get a precarity index accounting for both the instability of the sequence and the nature of the states is to apply a correction factor based on the nature of the states to any measure of sequence instability. For example, we would get such an indicator by multiplying either the turbulence or the complexity index by Brinzky-Fay's volatility indicator.

Instead of just distinguishing between positive and negative states, we could assign degrees of precarity to the different states. A temporary job would, for instance, get a higher degree of precarity than a full time job and a lower degree compared to unemployment. Moreover, the precarity of a trajectory depends on the evolution within the sequence and we should, therefore, not only account for the nature of the states but also for the type of the state transitions—changes of states—in the sequence.

Here, we consider that the precarity of a sequence

1. increases with the sequence instability due to the lack of predictability of the different states experienced;

2. increases with the proportion of downward transitions in the sequence, i.e. proportion of transitions to a less favourable state, and decreases with the proportion of upward transitions;
3. increases with the precarity degree of the starting state.

In addition, we consider that the transitions may have different advantageous or disadvantageous degrees that should be taken into account when computing the proportion of negative and positive transitions. For instance, we could consider that a transition from full time employment to unemployment is more damaging than a transition from a full time to a part time job. We would also expect that a transition from full time to unemployment generates a higher precarity weight than a transition from a temporary job to unemployment when the former is less likely to occur than the latter.

4 The Precarity Index

We use the complexity index (Gabadinho et al. 2010, 2011) as a measure of the sequence instability and propose to qualify—penalize/reward—it by means of a correction factor derived from the proportions of negative and positive transitions in the sequence.

4.1 Defining the Index

The index of complexity of a sequence is a composite index defined as the geometric mean between the normalized entropy of the sequence and the normalized number of transitions in the sequence. The entropy is normalized by dividing it by the logarithm of the size of the alphabet, i.e. the logarithm of the number of all possible states that the person can experience, which is the maximal possible entropy given the alphabet. The number of transitions is normalized by $\ell - 1$, i.e. the length of the sequence minus one. Formally, the complexity $c(s)$ of a sequence s reads

$$c(s) = \sqrt{\frac{h(s)}{\log(n_a)} \frac{nt_s}{(\ell_s - 1)}}$$

where $h(s)$ is the entropy, n_a the size of the alphabet, nt_s the number of transitions, and ℓ_s the length of the sequence.

Now, assuming the states—the labour market positions—can be ordered from the best to the worst state (see Sect. 4.4 for how to relax this strict order requirement), we say that a state transition $A \rightarrow B$ in a sequence is negative when there is a deterioration, i.e. when the difference $\text{rank}(A) - \text{rank}(B)$ between the ranks of the origin and destination states is negative. Likewise, the transition is said positive

in case of improvement, i.e. when $\text{rank}(A) - \text{rank}(B) > 0$. Letting $q^-(s)$ be the (weighted) proportion of negative transitions in the sequence s and $q^+(s)$ the (weighted) proportion of positive transitions, we define the correction factor in terms of the difference between the two:

$$q(s) = q^-(s) - q^+(s).$$

Since $q^-(s)$ and $q^+(s)$ are proportions, we have $-1 \leq q(s) \leq 1$. The correction factors will be $1 + q(s)$ and the proposed qualified complexity index reads

$$\text{prec}(s) = \lambda\, a(s_1) + (1 - \lambda)\, c(s)^\alpha (1 + q(s))^\beta \tag{1}$$

where $c(s)$ is the complexity index of the sequence and $a(s_1) \in [0, 1]$ the starting cost, i.e. the degree of precarity associated to the starting state s_1 in the sequence. The correction factor $(1 + q(s))$ is non negative. It is greater than 1 when $q(s)$ is positive, i.e. when there are more negative than positive transitions. Thus the greater $q(s)$ the stronger the penalization of the original complexity index. The parameter λ serves to control the trade-off between the starting cost and the corrected complexity while the exponents α and β control the respective importance of the complexity and the correction. The choice of the values of these parameters is addressed in Sect. 4.2.

Different variants of $q(s)$ may result depending on whether we take into account the transition costs, and, if so, on how these costs are determined.

Let $w(s_t, s_{t+1})$ be the cost of a transition from state s_t to state s_{t+1} over two successive time positions t and $t + 1$. To get $q(s)$, we first compute the total cost $nw(s)$ of the successive transitions in the sequence s, the total cost $nw^-(s)$ of the negative transitions, and the total cost $nw^+(s)$ of the positive transitions:

$$nw(s) = \sum_{t=1}^{\ell-1} w(s_t, s_{t+1})$$

$$nw^-(s) = \sum_{t=1}^{\ell-1} I^-(s_t, s_{t+1})\, w(s_t, s_{t+1})$$

$$nw^+(s) = \sum_{t=1}^{\ell-1} I^+(s_t, s_{t+1})\, w(s_t, s_{t+1})$$

where $I^-(s_t, s_{t+1})$ is a deterioration indicator taking value 1 for state deterioration and 0 otherwise, and $I^+(s_t, s_{t+1})$ a similar function for state improvement. The proportions of cost-weighted negative and positive transitions are then:

$$q^-(s) = \frac{nw^-(s)}{nw(s)} \qquad q^+(s) = \frac{nw^+(s)}{nw(s)}.$$

4.2 Tuning the Index

The formula (1) has three tuning parameters λ, α, and β that allow the formula to encompass a whole family of indexes. We should chose these parameters such that $0 \leq \lambda \leq 1, \alpha \geq 0$ and $\beta \geq 0$.

The parameter λ determines the trade-off between the precarity of the starting state and the corrected complexity. With $\lambda = 0$, the precarity level of the starting state in the sequence would simply be ignored while with $\lambda = 1$ the precarity of the whole sequence would just be that of its starting state. It seems reasonable to give less importance to the starting state than to the corrected complexity and we suggest, therefore, a value $\lambda = 0.2$. With this value the precarity degree of the starting state receives a 20% weight while the corrected complexity counts for 80% of the index.

Parameters α and β are exponential weights that allow for some control on the respective importance of the complexity and the correction factor in the corrected complexity term. For instance, we would get an index that does not account at all for the complexity by setting $\alpha = 0$. Setting in addition $\lambda = 0$, the index would reduce to the mere correction factor. Likewise, $\beta = 0$ suppresses the correction. We get unweighted effects with $\alpha = \beta = 1$. With $\alpha > 1$ we increase the importance of the complexity. Likewise, a value $\beta > 1$ strengthens the correction, which may prove useful in case we feel the correction is insufficient. We got good results in our experiments with $\alpha = 1$ and $\beta = 1.2$.

Alongside the values of the tuning parameters, the analyst has also to make a choice regarding the precarity degree $a(s_1)$ of the starting state—the offset in formula (1)—and the weights $w(s_t, s_{t+1})$ of the transitions that impact the correction factor $1 - q(s)$. Different strategies can be envisaged for these choices including defining them on theoretical grounds, on the hypothesized rank order of the states, or deriving them from the data.

Regarding the precarity degree of the states to be used as $a(s_1)$ value, a solution using the hypothesized state order is to assign the n_a equally spaced values between 0 and 1 as precarity degree to the sorted states. The ith state would in that case get a precarity degree of $(i - 1)/(n_a - 1)$. As a data-driven approach, we could, for example, set the precarity degree of each state as the probability to visit at least k bad states during the next m periods.

Likewise, referring to the rank order of the states, we could set the costs $w(s_t, s_{t+1})$ of the transitions as the difference between the ranks of the origin and destination states. Data driven approaches for the transition costs could be for instance

1. Give higher cost weights to rare transitions by defining the weight of each transition as a decreasing function (e.g. $1 - p$) of its estimated transition probability $p(s_{t+1}|s_t)$. Here, we could either compute this probability on the original sequences or ignore the durations of the successive states in the sequences and compute the probabilities of transition on the sequences of distinct successive states (DSS).

2. Define the cost of each negative transition as the estimated probability to be in a bad state say k periods after the transition, and for each positive transition as the estimated probability to be in a good state k periods after the transition.

For the illustrative example in Sect. 4.3 and the application in Sect. 5, we retain the values $\lambda = 0.2$, $\alpha = 1$, and $\beta = 1.2$, set the precarity degree $a(s_1)$ as the n_a equally spaced values between 0 and 1, and we use the complement to 1 of the transition probabilities in the DSS sequences as transition costs.

4.3 Behavior of the Precarity Index

To illustrate the behaviour of the index, we consider the set of fictitious sequences shown in Table 1 where the states are F, Full time permanent employment, P, Part time, T, Temporary employment, U, unemployment, and I, inactivity. We assume that there is a decreasing hierarchical order of the first four states, namely F, P, T, U. We also assume that I is neither better nor worse than the other states and can therefore not be classified.

So far we have assumed the states are strictly ordered. Therefore, we consider for now only the first 15 sequences where the non-comparable state I does not appear. In Table 1, the first two sequences are the most complex with three transitions each. The first one is slightly more complex with four different states as opposed to three states in the second sequence. Nevertheless, the second looks more precarious with two downward transitions compared to only one in the first sequence. Sequences 14 and 15 are composed of only one state. They have a zero complexity degree. However, sequence 14 in which the stable state is unemployment

Table 1 Example dataset

	Sequence
1	U/2-P/3-F/2-T/2
2	F/4-P/2-U/2-F/1
3	F/2-T/2-U/5
4	F/2-P/5-T/2
5	U/2-F/5-T/2
6	U/2-P/5-T/2
7	U/2-T/2-F/5
8	T/4-U/5
9	F/4-U/5
10	F/4-P/5
11	T/4-F/5
12	T/4-P/5
13	F/7-U/2
14	U/9
15	F/9

	Sequence
16	F/3-T/3-I/3
17	T/4-I/2-P/3
18	T/3-P/3-I/3
19	F/5-P/2-I/2
20	I/3-P/3-I/3
21	F/4-I/5

Fig. 1 Precarity and its components under the strict state order assumption

(U) looks more precarious than sequence 15 composed of one favourable state (F, full-time employment). We expect the index of precarity to reflect these differences in quality and complexity.

The values of the precarity index (computed with $\lambda = 0.2$ and $\beta = 1.2$), as well as of the complexity index are shown in the right panel of Fig. 1. In the middle panel the figure shows for each sequence s the starting cost $a(s_1)$, i.e. the precarity degree of the first state in the sequence, and the correction factor $(1 + q(s))^{\beta}$.

The index behaves as expected regarding the first two sequences: The second sequence gets a slightly higher precarity value than the first despite its lower complexity. Sequence 14 with zero complexity—i.e. no transitions—gets a non-zero precarity value because the cost of its starting state—U, being unemployed—is high in terms of precarity. This is not the case for sequence 15 the other sequence with zero complexity, which starts off with employment, i.e. zero cost in terms of precarity. We observe that the sequences with highest precarity are not the more complex ones but those with only negative transitions. The index of precarity is highest for sequences 3 and 4. It is slightly lower for sequences 8, 9, 10 and 13 that also have only negative transitions, but lower complexity due to only one transition. In addition, as expected sequences with only positive transitions (7, 11, 12) get low values of precarity.

4.4 Relaxing the Strict State Ordering Requirement

A strict order of the states is hardly compatible with the complexity of reality, especially when some states could not be clearly ranked with respect to the others (e.g. should education be considered as worse or better than employment?). We consider two situations that depart from the strict state order assumption: equivalent classes of states and non-comparable states.

A *class of equivalent states* is a subset of states that are considered equivalent, i.e. states with no ordering between them. In our example, F and P could be considered as equivalent if we assume that the choice between a full-time and a part-time job is a pure employee's choice.

A *non-comparable state* is a state that cannot be ordered, i.e. a state that is neither better nor worse than any other state. This is the case of the inactivity state I in our example.

We need solutions to handle these cases at two levels: for the computation of the proportions of negative and positive transitions, and for determining the state costs in the case where we want to derive them from the state order as we do in our example.

For equivalent classes, we do not penalize or reward any transition between states of a same equivalent class. In other words, all transitions between elements of an equivalent class get a zero weight in the weighted proportion of negative and positive transitions. As for transitions from or to any non-comparable state, we also chose to neither penalize nor reward them by giving them zero weight. In addition, however, subsequences such as $PIIIU$, where non-comparable states—the Is in the example—occur in-between two regularly ranked states will be counted as a transition from its first element to the last, e.g. $PIIIU$ is counted as a transition $P \rightarrow U$, i.e. as a negative transition.

Regarding starting costs based on the state order, we assign to each state (labour market position) in an equivalence class the mean cost of the states in the class, and the overall mean starting cost (i.e. 0.5) to each non-comparable state. A consequence is that in case the highest ranked state belongs to an equivalence class, its cost would be the non-zero mean value of the class and there would be no zero starting cost.

To illustrate we have computed the precarity index for the full set of sequences shown in Table 1, i.e. including those with the non-comparable state I and assuming in addition that full-time, F, and part-time working, P, form an equivalence class. We used again a trade-off value $\lambda = .2$ and an exponent weight $\beta = 1.2$. Figure 2 shows the obtained precarity values and their components the complexity index $c(s)$, the weighted correction factor $(1 + q)^{\beta}$, and the starting cost $a(s)$.

Looking at the first 15 sequences, we see a few differences with what we found in Fig. 1 using the strict order assumption. We first observe that there is now no sequence with a zero precarity value. Sequence 15 that had a zero value under the strict order assumption gets now a small positive value. This is because of the equivalence class between the two best ranked states in the state order. The starting state F gets here the non-zero mean value between F and P as precarity degree.

Fig. 2 Precarity and its components in presence of the non-comparable state I and the equivalence class $\{F, P\}$

Considering F and P as equivalent also has consequences on the ranking of the sequences. Sequence 1 gets here a higher precarity value than sequence 2. Because of the equivalence class, the two sequences have the same number of positive and negative transitions, which is reflected by correction factors close to 1. The main difference between the two sequences is the worse starting state in sequence 1.

Among the sequences with the non-comparable state I, sequence 16 appears to be the most precarious. It is made of a single downward transition and has maximal complexity for a sequence with 3 out of 5 states. Sequences 17 and 18 have only an upward transition and get therefore low precarity values. Finally, sequences 19, 20, and 21 have only zero weighted transitions—hence a neutral correction factor of 1—and get mid precarity values.

5 Application to the School to Work Transition

We now show how the index can be used on a real world dataset using the Status Zero Survey data of McVicar and Anyadike-Danes (2002) on the school to work transition of young Northern Irish.[1] This cohort survey was used to establish a link between individual, family and school characteristics and types of trajectories. The aim was to identify those young people who are more likely to experience unsuccessful trajectories in the adult labour market. The survey provides monthly information on the labour market activities of 712 young people for 72 months (6 years) after they left compulsory schooling. Despite the fact that these data refer to the period between July 1993 and June 1999, they represent a good testing ground for our index of precarity as they focus on early employment trajectories of young people who just left education. We complete the study from McVicar and Anyadike-Danes (2002) by assessing the quality of trajectories of young people and testing whether this contributes, beyond static individual school and family characteristics, to predicting future labour market positions. This is particularly relevant in the current economic and labour market situation, where young people have been hit hard by the crisis and are often overrepresented in temporary and precarious employment.

Due to its collection structure, all individuals are aged 16 at the start of the trajectories. Here we shall ignore the first two holiday months and retain the sequences from September 1993 to June 1999, i.e. sequences of length 70. The data distinguishes between six labour market activities: school (SC), training (TR), further education (FE), higher education (HE), employment (EM), and joblessness (JL).

We use the dataset to study how the degree of precarity of the trajectory during the first 36 months—from September 1993 to August 1996—impacts the situation of the young person two years later, i.e. the 6th year after the end of compulsory school. Hence, we aim to measure the scarring effect of early precarious trajectories in mid-term labour market outcomes. More specifically, we examine the chances to be at least one month in one of the states JL, TR, or SC during this 6th year, i.e. between September 1998 and June 1999.

In order to study the precarity of the trajectories during the first 36 months, we consider the three equivalence classes:

$$C_1 = \{\text{FE, HE, EM}\},\ C_2 = \{\text{SC, TR}\},\ C_3 = \{\text{JL}\}.$$

In addition, we assume the decreasing order $C_1 > C_2 > C_3$ of the equivalent classes. Thus, for example, changing from employment (EM) to training (TR) will be considered a downward transition, a change from school (SC) to further education (FE) as an upward transition, and a change from further education (FE) to employment (EM) as neutral.

[1] The data ship with the R package TraMineR (Gabadinho et al. 2011).

Table 2 Weights based on transition probabilities in the DSS

	→EM	→FE	→HE	→JL	→SC	→TR
EM→	0.00	0.00	0.00	0.67	0.94	0.75
FE→	0.00	0.00	0.00	0.81	0.97	0.90
HE→	0.00	0.00	0.00	0.80	1.00	1.00
JL→	−0.44	−0.80	−0.97	0.00	−0.99	−0.80
SC→	−0.68	−0.83	−0.68	0.87	0.00	0.00
TR→	−0.30	−0.93	−1.00	0.78	0.00	0.00

Table 3 Ten sequences with highest precarity value

Id	Sequence	Prec
705	FE/10-TR/12-EM/11-TR/3	0.38
98	SC/8-JL/4-FE/10-JL/2-FE/10-EM/2	0.39
70	FE/22-EM/2-TR/10-JL/2	0.39
110	TR/5-JL/2-EM/6-JL/6-EM/17	0.39
634	SC/10-JL/1-EM/5-JL/2-EM/18	0.39
305	FE/5-JL/3-EM/2-JL/2-FE/5-JL/1-EM/11-JL/2-EM/5	0.39
377	JL/10-EM/2-JL/10-EM/2-FE/10-EM/2	0.40
520	SC/9-JL/3-FE/5-TR/19	0.42
405	SC/6-JL/6-FE/3-JL/7-TR/7-JL/7	0.45
408	SC/1-JL/2-EM/4-JL/5-FE/18-JL/4-EM/2	0.45

We compute the index using $\lambda = .2$, $\beta = 1.2$, and transition weights based on the transition probabilities in the DSS. The weights are shown in Table 2. Weights of upward transitions are displayed with a negative sign to recall their reducing effect on the correction factor. Table 3 lists the ten sequences with the highest values for the precarity index. All these most precarious sequences have at least as many upward than downward transitions except sequence 377 that starts with the worst state JL. Sequence number 377 appears to be precarious because it starts with the worst state JL, i.e. a high starting cost. We can also observe that among the ten most precarious trajectories the upward transitions are typically transitions with lower weights than the downward transitions. E.g., the transition JL→EM that occurs frequently in the precarious sequences has a 0.44 weight, while SC→JL present in half of the ten sequences has almost twice that weight.

In order to predict the impact of the quality of the trajectory during the initial 36 months on the future situation we build two variables from the 10 last months of the observed sequences, i.e. months 61 to 70—September 1998 to June 1999—of our sequences. The first, *bad.dur*, is the total time spent in one of the states JL, TR, or SC during the last 10 months (6th year), and the second is a binary variable, *is.bad*, taking value 1 when *bad.dur*> 0, i.e. when the individual has spells of negative states during the 6th year. There are 19.4% of the followed individuals who spent at least a month in a negative state, and Fig. 3 shows the distribution of the total number of months for those 121 cases. Clearly most of those who had bad spells spent all ten months in undesirable states. Figure 4 shows that the distribution of

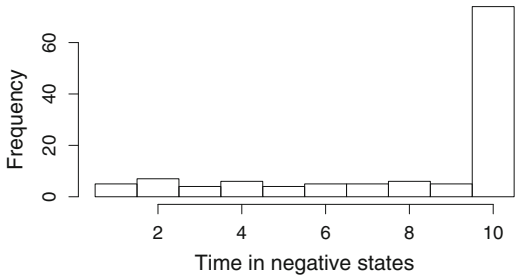

Fig. 3 Distribution of the time spent in negative states during last 10 months among those who experienced bad spells

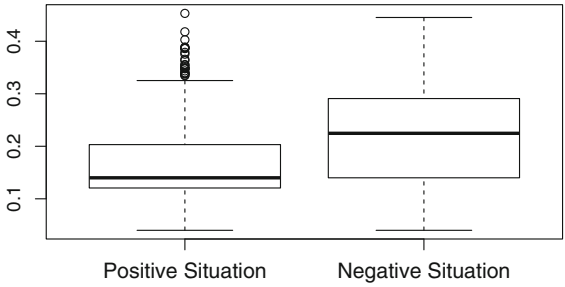

Fig. 4 Precarity degree during the first 36 months for those in a negative situation two years later versus those who are not

Table 4 Linear regression for time in bad states during last 10 months and logistic regression for 'More than 0 months in bad states during the last 10 months'

	Linear		Logistic	
	Estimate	Sig.	Odd Ratio	Sig.
(Intercept)	0.64	0.05	0.10	0.00
Precarity	8.10	0.00	632.77	0.00
Good End CS Qualification	−0.85	0.00	0.40	0.00
Male	−0.66	0.01	0.62	0.02

the precarity degree during the first 36 months greatly differs between those in bad situation two years later and the others.

The impact of the precarity degree can be measured through a logistic regression using *is.bad* as the dependent variable and the precarity as predictor. Table 4 shows the effect of the precarity degree when controlled for two other covariates, namely whether students gained good qualification at the end of compulsory school (*gcse5eq*) and whether they are males (*male*). The results evidently demonstrate the strength of the precarity degree of the early trajectory as predictor of the future situation. An increase of the precarity index by 0.1 unit for example multiplies the odd of experiencing a negative situation a few years later by about 60. From the linear regression, an increase of the precarity degree during the first 3 years of 0.1 unit leads on average to an increase of the time in negative states during the 6th year by almost one month.

6 Conclusion

In this study, we set out to develop an index to quantify the degree of precarity of individual sequences and to predict future insecurity in professional careers. Starting from the assumption that the complexity of the sequence contributes to the precarity of the trajectory, we have defined the index as a corrected complexity index. There is a multiplicative correction based on the difference between the proportions of downward and upward transitions in the sequence and an offset correction to account for the degree of precarity at the start of the sequence. Despite its relative simplicity, the index proved to be able to effectively grasp precarity. Using the data from McVicar and Anyadike-Danes (2002) on the school to work transition of school leavers in Northern Ireland, we also demonstrated the usefulness of the index for studying how precarity during the first years after compulsory schooling impacts future outcomes. There is certainly room for further improvements, for instance, by accounting for the time elapsed between transitions and/or the timing of the transitions. The concept of recency used by Manzoni and Mooi-Reci (2018) for their own quality index is also an interesting dimension to consider.

The precarity index is very flexible and can be tuned by choosing the transition weights, the degree of precarity of the starting costs, the trade-off parameter λ, and the exponent weights α and β that determine the respective importance of the complexity and the correction factor. These weights and parameters offer the analyst the possibility to adapt the index to specific contexts. However, making choices is not indispensable. The index provides most often sensible results with default parameter values and automatic methods for setting transition weights and starting precarity degrees. The only necessary information that the user has to specify is the rank order of the states with possible equivalence classes and non comparable states.

Although the indicator was specifically developed for measuring the precarity of sequences of labour market activities in order to predict future insecurity in professional careers, the index could as well be used for sequences of other domains of the life course such as family or health trajectories. The only requirement is the existence of some (partial) order between the states of the alphabet adopted, i.e. at least some states should be preferable to some others. Moreover, it could also be of interest to use the index as the dependent variable to study how precarity depends on personal characteristics such as sex, social origin or previous educational trajectory.

The index has been implemented as a beta version in `TraMineRextras` and should be made available in a next release of `TraMineR` (Gabadinho et al. 2011).

Acknowledgements Margherita Bussi and Jacqueline O'Reilly acknowledge financial support from the Horizons2020-funded NEGOTIATE project, grant agreement No 649395. Gilbert Ritschard acknowledges 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 (grant number: 51NF40-160590). The authors warmly thank the anonymous reviewers for their constructive comments.

References

Ayllón, S. (2013). Unemployment persistence: Not only stigma but discouragement too. *Applied Economics Letters, 20*(1), 67–71.

Barbier, J.-C. (2005). La précarité, une catégorie française à l'épreuve de la comparaison internationale. *Revue française de sociologie, 46*(2), 351–371.

Bell, D., & Blanchflower, D. (2011). Young people and the great recession. *Oxford Review of Economic Policy, 27*(2), 241–267.

Booth, A. L., Francesconi, M., Frank, J. (2002). Temporary jobs: Stepping stones or dead ends? *The Economic Journal, 112*(480), F189–F213.

Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review, 23*(4), 409–422.

Bussi, M., & O'Reilly, J. (2016). Institutional determinants of early job insecurity in the UK. NEGOTIATE Working Papers WP 3.4.

Cable, N., Sacker, A., Bartley, M. (2008). The effect of employment on psychological health in mid-adulthood: Findings from the 1970 British cohort study. *Journal of Epidemiology and Community Health, 62*(5), e10.

Chung, H., & Van Oorschot, W. (2011). Institutions versus market forces: Explaining the employment insecurity of European individuals during (the beginning of) the financial crisis. *Journal of European Social Policy, 21*(4), 287–301.

Cockx, B., & Picchio, M. (2012). Are short-lived jobs stepping stones to long-lasting jobs? *Oxford Bulletin of Economics and Statistics, 74*(5), 646–675.

Daly, M., & Delaney, L. (2013). The scarring effect of unemployment throughout adulthood on psychological distress at age 50: Estimates controlling for early adulthood distress and childhood psychological factors. *Social Science & Medicine, 80*, 19–23.

de Graaf-Zijl, M., Van den Berg, G. J., Heyma, A. (2011). Stepping stones for the unemployed: The effect of temporary jobs on the duration until (regular) work. *Journal of Population Economics, 24*(1), 107–139.

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, 23*, 225–250.

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.

Gabadinho, A., Ritschard, G., Studer, M., Müller, N. S. (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des nouvelles technologies de l'information RNTI, E-19*, 61–66.

Gardiner, L. (2016). Stagnation generation: The case for renewing the intergenerational contract. Report for the intergenerational commission (intergencommission.org), London Resolution Foundation. Last accessed 30 July 2016.

Gebel, M. (2010). Early career consequences of temporary employment in Germany and the UK. *Work, Employment and Society, 24*(4), 641–660.

Gregg, P., & Tominey, E. (2005). The wage scar from male youth unemployment. *Labour Economics, 12*(4), 487–509.

Kalleberg, A. L. (2009). Precarious work, insecure workers: Employment relations in transition. *American Sociological Review, 74*(1), 1–22.

Leschke, J., & Watt, A. (2008). *Job quality in Europe*. Brussels: ETUI.

Manzoni, A., & Mooi-Reci, I. (2011). Early unemployment and subsequent career complexity: A sequence-based perspective. *Schmollers Jahrbuch: Journal of Applied Social Science Studies, 131*(2), 339–348.

Manzoni, A., & Mooi-Reci, I. (2018). Measuring sequence quality. In G. Ritschard & M. Studer (Eds.), *Sequence Analysis and Related Approaches: Innovative Methods and Applications*. Cham: Springer (this volume).

McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A, 165*(2), 317–334.

Mills, M., & Blossfeld, H.-P. (2005). Globalization, uncertainty and the early life course: A theoretical framework. In H.-P. Blossfeld, E. Klijzing, M. Mills, & K. Kurz (Eds.), *Globalization, uncertainty and youth in society* (Advances in sociology series, pp. 1–24). Cheltenham: Routledge.

Muffels, R., & Luijkx, R. (2008). Labour market mobility and employment security of male employees in Europe: Trade-off or flexicurity? *Work, Employment and Society, 22*(2), 221–242.

O'Reilly, J., Eichhorst, W., Gábos, A., Hadjivassiliou, K., Lain, D., Leschke, J., McGuinness, S., Kureková, L. M., Nazio, T., Ortlieb, R., Russell, H., & Villa, P. (2015). Five characteristics of youth unemployment in Europe; flexibility, education, migration, family legacies, and EU policy. *SAGE Open, 5*(1), 1–19.

Ortiz, L. (2010). Not the right job, but a secure one over-education and temporary employment in France, Italy and Spain. *Work, Employment and Society, 24*(1), 47–64.

Palier, B., & Thelen, K. (2010). Institutionalizing dualism: Complementarities and change in France and Germany. *Politics & Society, 38*(1), 119–148.

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European Sociological Review, 17*(2), 119–144.

Scherer, S. (2004). Stepping-stones or traps? The consequences of labour market entry positions on future careers in West Germany, Great Britain and Italy. *Work, Employment and Society, 18*(2), 369–394.

Schmelzer, P. (2011). Unemployment in early career in the UK: A trap or a stepping stone? *Acta Sociologica, 54*(3), 251–265.

Schmid, G. (2015). Sharing risks of labour market transitions: Towards a system of employment insurance. *British Journal of Industrial Relations, 53*(1), 70–93.

Schmid, G., & Schömann, K. (2004). Managing social risks through transitional labour markets: Towards a European social model. TLM.NET Working Papers 2004–01, SISWO/Institute for the Social Sciences, Amsterdam.

Smithson, J., & Lewis, S. (2000). Is job insecurity changing the psychological contract? *Personnel Review, 29*(6), 680–702.

Standing, G. (1999). *Global labour flexibility: Seeking distributive justice*. London: Palgrave Macmillan.

Tumino, A. (2015). The scarring effect of unemployment from the early '90s to the Great Recession. ISER Working Paper Series 2015–05, Institute for Social and Economic Research.

Weich, S., & Lewis, G. (1998). Poverty, unemployment, and common mental disorders: Population based cohort study. *BMJ, 317*(115), 115–119.

Worth, S. (2005). Beating the 'churning' trap in the youth labour market. *Work, Employment and Society, 19*(2), 403–414.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Correction to: Unpacking Configurational Dynamics: Sequence Analysis and Qualitative Comparative Analysis as a Mixed-Method Design

Camilla Borgna and Emanuela Struffolino

Correction to:
Chapter 10 in: G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*, Life Course Research and Social Policies 10, https://doi.org/10.1007/978-3-319-95420-2_10

In the original version of this book, the second author Emanuela Struffolino was missed to be added as the corresponding author and the affiliation of the author Camilla Borgna was incorrect. The second author Emanuela Struffolino has now also been included as the corresponding author and the affiliation of Camilla Borgna is corrected as Collegio Carlo Alberto, Turin, Italy.

The updated online version of this chapter can be found at
https://doi.org/10.1007/978-3-319-95420-2_10

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_17

Index

B
Binary sequences, 261

C
Career recovery, 262
Career trajectory, 52
Childhood co-residence structures, 84
Cluster analysis, 71, 84, 131, 155, 185, 203, 223, 241
Compressing sequence data, 185
Configurations, 168
Continuous data, 203
Cox regression, 70

D
Departure from the parental home, 84
Disability, 70
Discrete-time event history analysis, 56
Dissimilarity measure between sequences, 149, 241
Divisive clustering, 225
Dual-earner couples, 241

E
Employment insecurity, 279, 281
Event history analysis, 16, 56, 64, 84

F
Failure, 261
Fuzzy clustering, 223

G
Gender, 122
Gender inequality, 50
Glass ceiling, 50
Glass escalator, 50
Growth mixture model, 203

H
Hidden Markov model, 185, 203
HMTD model, 203

I
Index of complexity, 279
Index of precarity, 279
Internet addiction test, 203

K
Kinship, 122

L
Latent Markov model, 186
Lexicographic index, 241
Life course analysis, 35
Life history calendar, 36
Life trajectories, 69
Longitudinal data, 1, 204

M
Migration, 122
Mixed-methods, 167
Monothetic, 224

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10, DOI10.1007/978-3-319-95420-2

Mortality, 69
Multichannel sequences, 35, 185, 241
Multilevel analysis, 17
Multistate model, 36

N
Network analysis, 103, 121

O
Occupational mobility, 49
Optimal matching, 1

P
Phase, 149
Polythetic, 223
Prediction probability, 36
Probabilistic model, 36
Professional competitions, 149

Q
QCA, 168
Quality measure, 261, 279

S
Scarring, 262, 279
Sequence analysis, 1, 19, 69, 84, 104, 121, 167, 223
Sequence history analysis, 83
Sequence networks, 103, 121
Sequence quality, 261, 274
Sequence structure, 103
Social network analysis, 16, 121
Social sequence analysis, 185
Success, 261

T
Temporal structures, 149
Time use, 113, 241
Trajectories, 168, 279

U
Unemployment scarring, 261, 279

V
Vertical sex segregation, 50
Visualizing sequence data, 149, 186