CHAPTER 1

INTRODUCTION

1.1 Sequence Analysis in the Social Sciences

Studying sequences of events or (temporally) ordered social processes is a primary concern of social sciences. In psychology, sociology, and political science, sequences figure prominently in various kinds of stage theories. Similarly, sequences are central to scholars engaged in life course sociology or life span psychology. Moreover, sequences are scrutinized in other fields of human and social sciences such as anthropology, archaeology, geography, and economics (for a more detailed discussion, see Abbott, 1995; Blanchard, 2019; and Cornwell, 2015). Given this broad interest coming from different branches of social sciences, it is not surprising that we encounter a wide variety of definitions and analytical tools in the largely disconnected literature on this topic. As a result, a wide range of quantitative methods as diverse as mixed latent Markov models, event history analysis, and panel regressions have been discussed as tools for analyzing sequence data. All of these methods are representatives of the so-called stochastic modeling culture and thus are based on stochastic assumptions regarding the data-generating process (Aisenbrey & Fasang, 2010; Breiman, 2001; Piccarreta & Studer, 2018). In addition, most of them share the focus on studying single or repeated transitions rather than the process as a whole. These methods are well-established, widely used, and partly also covered in previous volumes of this book series (Allison, 2009, 2014; Preacher et al., 2008).

The present volume is dedicated to a different set of analytical tools, sometimes referred to as *whole sequence methods* for categorical data or simply *sequence analysis* (SA). These tools were first introduced to social sciences in the form of optimal matching (OM) analysis by Andrew Abbott (Abbott & Forrest, 1986; Abbott & Hrycak, 1990), who imported this approach from biology and computer science (Levenshtein, 1966; Sankoff & Kruskal, 1983). OM analysis quantifies the degree of dissimilarity between sequences. In contrast to the methods just mentioned, OM and several of the SA tools have their roots in the data mining or algorithmic modeling culture and thus do not make any assumptions on the data-generating process. Accordingly, SA lends itself particularly well to studying exploratory and descriptive research questions. More specifically, SA tools are usually applied to (a) describe and (b) visualize sequences, (c) identify typical patterns among a set of sequences, and (d) examine the antecedents and consequences of these patterns (Abbott, 1990; Vanhoutte et al., 2018).

SA conceptualizes social sequences as a series of—usually temporally—ordered categorical elements. It has been applied for studying processes as diverse as labor market entry sequences (distinguishing states such as "education," "employed," "unemployed," "inactive"; e.g., Brzinsky-Fay, 2007); partnership biographies (e.g., "single," "living apart together," "cohabiting," "married"; e.g., Raab & Struffolino, 2019); pathways toward democratization ("parliamentary democracy," "presidential democracy," "military dictatorship," "monarchy"; Wilson, 2014); time use of couples ("nobody is working," "both partners are working," "only one partner is working"; Lesnard, 2008); actual and idealized relationship scripts ("meet partner's parents," "go out alone," "hold hands," "kiss," "sexual intercourse"; Soller, 2014); or basic types of figures in ritual dances ("once-to-yourself," "footing," "partners," "rounds," "hey"; Abbott & Forrest, 1986).

SA treats sequences as units of analysis. Instead of studying specific transitions between states or events in isolation, SA puts them into context by simultaneously considering the timing, ordering, and duration of all the states that make up a sequence. It thus acknowledges that the meaning and the consequences of social facts usually can be fully understood only by considering "their larger sequential context" (Cornwell, 2015). The various pathways leading to an identical labor market status at age 45, for instance, can differ considerably, which in turn can have severe consequences for the accumulation of wealth, marital status, or health outcomes.

In the first 10 to 15 years after its introduction to the social sciences, SA was predominantly used for identifying patterns in sequences by means of OM analysis (Abbott & Tsay, 2000; Aisenbrey & Fasang, 2010; MacIndoe & Abbott, 2004). This procedure involves three analytical steps: defining a coding scheme for the elements constituting the sequences (see previous examples for different processes), computing pairwise dissimilarities between sequences, and identifying patterns applying multidimensional scaling or clustering to the dissimilarity matrix obtained in the second step. The resulting clusters are used as either an independent or a dependent variable in a regression analysis to examine the determinants or consequences of specific sequence patterns. This standard procedure of analytical steps still prevails in SA today, although the methodological tools have been continuously refined and expanded since then. Most of these advances were developed in the aftermath of the "2000 controversy" (Aisenbrey, 2017) in the journal *Sociological Methods & Research,* in which SA's algorithmic modeling culture and the data handling of optimal matching analysis have been heavily criticized (Levine, 2000; Wu, 2000). Ten years later, the same journal published a special issue on SA in which Aisenbrey and Fasang (2010) provided an excellent overview of the initial critique of classical optimal matching applications and the following methodological advancements aptly referred to as the “second wave of sequence analysis,” which mainly extended OM and introduced other techniques to compare sequences. In the present volume, we will introduce both the classical optimal matching approach as well as analytical tools of the so-called second wave. In addition, we briefly highlight some more recent advances we consider to be the third wave of SA. This wave is largely characterized by the effort of bringing together the stochastic and the algorithmic modeling cultures by jointly applying SA with more established methods such as analysis of variance, event history, network analysis, or causal analysis in general (Barban et al., 2017; Cornwell, 2015; Piccarreta & Studer, 2018; Ritschard & Studer, 2018; Studer et al., 2011).

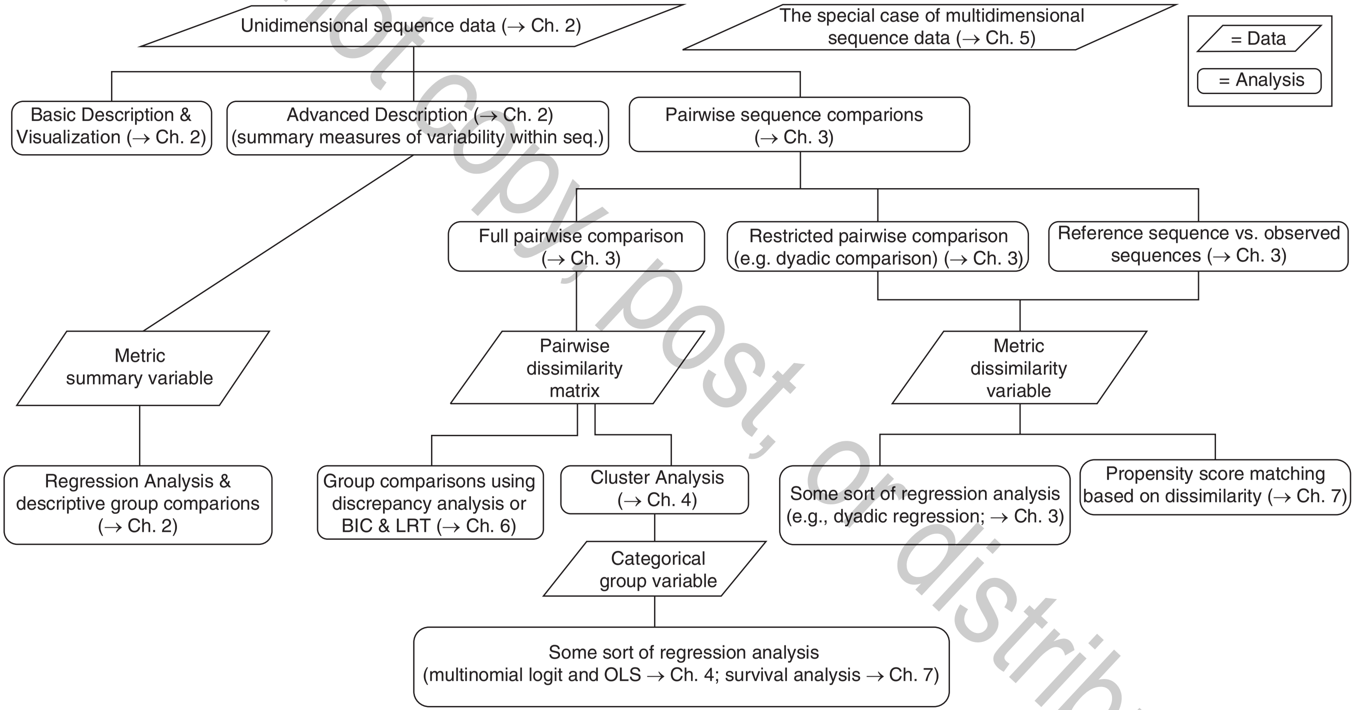
In sum, the field of SA is undergoing a burgeoning development since the early contributions in the 1980s that is clearly favored by the increasing availability of suitable (longitudinal) data and software packages. Particularly since the 2000s and within the disciplines of sociology and demography, this development is accompanied by an increasing number of publications in renowned journals that contribute to the overall visibility of the method beyond the small but steadily growing community of SA experts.

It is beyond the scope of this book to provide a comprehensive review of all the new and diverse SA tools that have been developed in recent years, nor do we strive for giving a full account of the history of sequence analysis in the social sciences. Instead, this book aims to equip researchers with the knowledge necessary to conduct their own SA. We introduce the basic concepts and techniques of SA but also consider many of the more recent developments, which to date have only been described scattered across several journal articles, book chapters, and software documentation. The book discusses the implication of the typical analytical choices involved in conducting a sequence analysis and provides several guidelines and recommendations. By covering the foundations as well as up-to-date summaries of current discussions on the key analytical steps of SA in one resource, this book addresses both readers who are new to the method and those who already used it in their research. The following section provides an overview clarifying which chapters might be of interest to these different groups of readers.

1.2 Organization of the Book

Figure 1.1 displays the structure and content of the following chapters. In Chapter 2, we first introduce the basic concepts and terminology of SA

Figure 1.1 The workflow of sequence analysis


Unidimensional sequence data (→ Ch. 2)
The special case of multidimensional sequence data (→ Ch. 5)
= Data
= Analysis
Basic Description & Visualization (→ Ch. 2)
Advanced Description (→ Ch. 2) (summary measures of variability within seq.)
Pairwise sequence comparions (→ Ch. 3)
Full pairwise comparison (→ Ch. 3)
Restricted pairwise comparison (e.g. dyadic comparison) (→ Ch. 3)
Reference sequence vs. observed sequences (→ Ch. 3)
Metric summary variable
Pairwise dissimilarity matrix
Metric dissimilarity variable
Regression Analysis & descriptive group comparions (→ Ch. 2)
Group comparisons using discrepancy analysis or BIC & LRT (→ Ch. 6)
Cluster Analysis (→ Ch. 4)
Some sort of regression analysis (e.g., dyadic regression; → Ch. 3)
Propensity score matching based on dissimilarity (→ Ch. 7)
Categorical group variable
Some sort of regression analysis (multinomial logit and OLS → Ch. 4; survival analysis → Ch. 7)

before we proceed to the critical task of defining meaningful sequences and touch upon the issue of handling missing data. Once these conceptual foundations are laid, we introduce tools to explore sequences by tabulating their properties and visualizing them with insightful graphs. Descriptive exploration includes simple counts of sequence frequencies, time spent in different states, transition rates between different states, and state patterns. Although these descriptive analyses provide a useful first overview, they might be somewhat overwhelming because they summarize the sequence data by presenting a lot of different numbers. Therefore, we introduce more advanced descriptors that either just summarize the diversity of individual sequences by calculating composite indices such as sequence complexity and turbulence, or try to assess the quality of status changes within sequences by discriminating between positive and negative transitions (e.g., from unemployment to employment or vice versa). Conveying much information in a single indicator, these composite indices are interesting by themselves, but we will also show that they are well-suited for being used as dependent or independent variables in regression analyses.

Chapter 3 is dedicated to the "core business" (Gauthier et al., 2014) and arguably also to one of the most controversial areas of SA, whole-sequence comparison methods. Much of the SA literature in social sciences is devoted to introducing, evaluating, and discussing different approaches to comparing sequences by using either OM and its extensions (also called alignment-based methods) or alternative (non-alignment-based) methods. All of these approaches aim to measure the degree of dissimilarity between sequences. Calculating dissimilarities is a crucial step in identifying patterns among a set of sequences and thus is central to most SA applications. We start this chapter by introducing the classic optimal matching technique, a computational alignment procedure that quantifies the degree of dissimilarity of two sequences by evaluating how many data transformations are required to change one sequence into another. The transformation operations are associated with specific costs, and the dissimilarity between two sequences is given as the sum of the costs for aligning them. The specification of these costs is a very controversial issue that has been addressed in many methodological contributions since optimal matching was first introduced to social sciences. Reflecting this literature, the chapter covers several techniques of quantifying the pairwise dissimilarities between sequences using either optimal matching or other approaches. We discuss the substantial implications of the cost specification (e.g., focus on timing versus ordering of states); present different strategies for setting them (e.g., data-driven, based on state attributes, theory-driven); and compare the resulting outcomes using our example data. The computation of dissimilarities offers much room for analytical decisions on the part of the researcher, which often has been blamed as being a fundamental problem of SA. We agree that the analytical choices involved in the quantification of sequence resemblance can indeed be a critical step of SA, but we consider this a strength rather than a weakness of the method. The freedom of choice allows the researchers to make theoretically informed decisions that stress those attributes of the sequences—such as order, timing, or duration—they consider most important. In addition to the computation of full pairwise dissimilarity matrices that compare every sequence with every other sequence in the data set, Chapter 3 introduces techniques that compute pairwise dissimilarities among observed sequences and some sort of reference sequence. These reference sequences could correspond to a theoretically motivated sequence, the most prevalent or representative sequence, or the sequence of a significant other person (e.g., a sibling, parent, or partner). Contrary to the full comparison approach, this procedure usually yields only one dissimilarity value for every observation that can be further analyzed as an independent or dependent variable in some sort of regression analysis. For instance, in the context of a dyadic analysis, one could examine if the stability of a newly formed partnership is affected by the resemblance of the partners' previous partnership biographies, or figure out which dyadic characteristics decrease or increase the dissimilarity between two persons' previous partnership biographies.

Chapter 4 provides an overview and a thorough discussion of the most common analytical procedure succeeding the computation of a full pairwise sequence dissimilarity matrix, that is, cluster analysis. This method is employed to identify patterns among a set of sequences and assigns every sequence to a specific cluster. The chapter introduces and compares different clustering algorithms, presents techniques for evaluating the substantive quality of the clustering, and discusses heuristics for cluster validation. Finally, we will illustrate how regression analysis can be used to either predict the assignment to a specific cluster based on covariates measured prior to the beginning of the sequences or estimate the effects of cluster membership on an outcome measured at or after the end the sequences.

Chapter 5 is devoted to the analysis of multidimensional sequence data comprising information on multiple trajectories, such as employment, residential, and family biographies. With multichannel sequence analysis (MCSA), the SA toolkit offers a dedicated method for simultaneously analyzing multiple sequences that has steadily gained in popularity since its introduction by Gauthier and colleagues (Gauthier et al., 2010). Before turning to this approach, however, we discuss the potential and pitfalls of analyzing multiple trajectories simultaneously, introduce alternative approaches to handling multidimensional sequence data, and illustrate how to examine the statistical correlation of different sequence channels.

Chapter 6 introduces two procedures for studying the relationship between covariates and sequences that circumvent the step of cluster analysis: the discrepancy analysis framework and a procedure utilizing an adjusted version of the Bayesian information criterion. Drawing on well-established methods from the stochastic modeling culture, these two approaches avoid deterministic cluster assignment and allow for testing the relevance and statistical significance of group differences. The discrepancy framework proposed by Studer et al. (Studer et al., 2011) generalizes the principle of analysis of variance, while the approach of Liao and Fasang is introducing an adjusted version of the Bayesian information criterion and the likelihood ratio test (Liao & Fasang, 2021). We introduce and compare both approaches; show how they can be used to examine the association between sequences and multiple variables simultaneously; and demonstrate how the obtained results can be complemented by the implicative statistic (Studer, 2015), a simple test statistic that allows for visualizing how sequences differ across groups.

Chapter 7 delineates a selection of other, more recent approaches that bring together the stochastic and the algorithmic modeling cultures by combining either event history analysis or propensity score matching with SA (Barban et al., 2017; Studer, Liefbroer, & Mooyaart, 2018; Studer, Struffolino, & Fasang, 2018). Although it would go beyond this introductory book's scope to present these methods in detail, we consider it useful to point out these recent developments and refer the readers to the most relevant resources to further study these methods.

Chapter 8 concludes the book by summarizing important practical recommendations covering all analytical steps involved in typical sequence analysis projects.

1.3 Software, Data, and Companion Webpage

This book is targeted at persons who want to apply SA in their own research. Instead of lengthy discussions of theoretical and statistical foundations of SA,[1] it provides clear advice on how to apply various SA tools and raises awareness of the pitfalls related to the many analytical decisions involved in such an analysis. Against this background, we considered it useful to illustrate the procedures with a real-world data set comprising sequences on family formation as well as labor market participation. In a sense, these sequences are typical representatives for SA because much of

[1] For such discussions, the reader is referred to Abbott (2017), Abbott and Tsay (2000), Blanchard (2019), Cornwell (2015), Fasang and Mayer (2020), and Halpin (2014).

the applications and methodological contributions are related to the field of life course research. The data come from the German Family Panel (pairfam), release 10.0 (Brüderl et al., 2019). A description of the study can be found in Huinink et al. (2011). The data were collected between 2008 and 2018 and include detailed accounts on family and employment biographies. For the examples presented in the following chapters, we use data from a baseline sample of more than 4,000 respondents born between 1971 and 1973 for which we reconstructed sequences covering the age span of 18–40 years. The final analytical samples comprise 1,866 cases for the family biographies and 1,032 respondents for the labor market trajectories. Although we discuss the issue of missing data in Chapter 2, our example data include only sequences without any missing data or gaps along the sequences.

The companion webpage (https://sa-book.github.io) of this book hosts the data and code required for reproducing the results presented throughout the book. Next to plain code files, the companion page provides instructions and some bonus material introducing additional examples and analytical tools. The webpage also includes colored versions of the grayscale figures presented in this volume. Together with the book, these materials provide a comprehensive resource for self-study or use in methods courses on SA. In terms of software, the companion website provides material for the free software R. In R, the key packages for SA are TraMineR, TraMineRExtras (Gabadinho, Ritschard, Müller, & Studer, 2011), and WeightedCluster (Studer, 2013). Together they represent a very powerful and versatile environment for conducting SA.

Although the companion page will focus on R, we want to mention briefly the notable set of tools for conducting SA in Stata. Like in R, these tools are available in the form of user-written packages: SQ (Brzinsky-Fay et al., 2006), MICT (Halpin, 2016b, 2019), and SADI (Halpin, 2017). In most standard applications, both software environments are well suited for conducting SA and it is up to the user which program is preferred. That said, the material on the companion website uses R for several reasons: (a) Contrary to the commercial statistical software Stata, R is freely available; (b) the SA toolkit in R is much more encompassing, comprising many methods and functions that are not available in Stata; and (c) the development of new methods is taking place mainly within the R environment, making it the more future-proof choice.