Sequence analysis: Its past, present, and future

Liao, Tim F.; Bolano, Danilo; Brzinsky-Fay, Christian; Cornwell, Benjamin; Fasang, Anette Eva; Helske, Satu; Piccarreta, Raffaella; Raab, Marcel; Ritschard, Gilbert; Struffolino, Emanuela; Studer, Matthias

How to cite

LIAO, Tim F. et al. Sequence analysis: Its past, present, and future. In: Social science research, 2022, vol. 107, p. 102772. doi: 10.1016/j.ssresearch.2022.102772

This publication URL: https://archive-ouverte.unige.ch/unige:163710
Publication DOI: 10.1016/j.ssresearch.2022.102772

Social Science Research 107 (2022) 102772

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Sequence analysis: Its past, present, and future

Tim F. Liao [a,*], Danilo Bolano [b], Christian Brzinsky-Fay [c], Benjamin Cornwell [d], Anette Eva Fasang [e], Satu Helske [f], Raffaella Piccarreta [b], Marcel Raab [g], Gilbert Ritschard [h], Emanuela Struffolino [i], Matthias Studer [h]

[a] University of Illinois at Urbana-Champaign, IL, USA
[b] Bocconi University, Italy
[c] WZB Berlin Social Science Center, Germany
[d] Cornell University, USA
[e] Humboldt University Berlin, Germany
[f] University of Turku, Finland
[g] State Institute for Family Research at the University of Bamberg, Germany
[h] University of Geneva, Switzerland
[i] University of Milan, Italy

ARTICLE INFO

Keywords:
Sequence analysis
Methodology
Life course research
Methodological review
Quantitative methodology

ABSTRACT

This article marks the occasion of *Social Science Research*'s 50th anniversary by reflecting on the progress of sequence analysis (SA) since its introduction into the social sciences four decades ago, with focuses on the developments of SA thus far in the social sciences and on its potential future directions.

The application of SA in the social sciences, especially in life course research, has mushroomed in the last decade and a half. Using a life course analogy, we examined the birth of SA in the social sciences and its childhood (the first wave), its adolescence and young adulthood (the second wave), and its future mature adulthood in the paper.

The paper provides a summary of (1) the important SA research and the historical contexts in which SA was developed by Andrew Abbott, (2) a thorough review of the many methodological developments in visualization, complexity measures, dissimilarity measures, group analysis of dissimilarities, cluster analysis of dissimilarities, multidomain/multichannel SA, dyadic/polyadic SA, Markov chain SA, sequence life course analysis, sequence network analysis, SA in other social science research, and software for SA, and (3) reflections on some future directions of SA including how SA can benefit and inform theory-making in the social sciences, the methods currently being developed, and some remaining challenges facing SA for which we do not yet have any solutions. It is our hope that the reader will take up the challenges and help us improve and grow SA into maturity.

1. Introduction

On the occasion of *Social Science Research*'s 50th anniversary, we reflect in this paper on the progress of sequence analysis (SA) since its introduction into the social sciences four decades ago by focusing on (1) the origin and the early growth of SA in the social sciences,

* Corresponding author.
E-mail address: tfliao@illinois.edu (T.F. Liao).

https://doi.org/10.1016/j.ssresearch.2022.102772
Received 25 March 2022; Received in revised form 30 June 2022; Accepted 5 July 2022
Available online 26 August 2022

(2) the major developments, especially those of the second-wave SA in the twenty-first century, by paying attention to the strengths and limitations of the developments, and (3) possible future directions for SA.

In the last two decades, SA has seen many successful applications in the social sciences, especially in life course research. According to Google Scholar, the search terms of "sequence analysis" and "life course" returned the following exponential increase in results containing the two terms: 10 for the decade of 1980–1989, 65 for the decade of 1990–1999, 408 for the decade of 2000–2009, 2320 for the decade of 2010–2019, and 1080 for the under 18 months' time from January 2020 to June 30, 2022. Fig. 1 displays a detailed yearly growth trend of SA from 1990 to December 2021 based on Web of Science journal article data. Using a life course analogy, we examine in this paper the birth of SA in the social sciences and its childhood (the first wave), its adolescence and young adulthood (the second wave), and its mature adulthood (the third wave and beyond).

A typical application of SA in the early days, according to Abbott and Tsay (2000), involves three steps: coding narratives or processes as sequences, measuring pairwise dissimilarities between sequences, and some form of data reduction such as cluster analysis, although today's applications of SA involve more, as the reader will find out later in the article. Let us illustrate the first two steps of this process using the five toy sequences in Fig. 2, in which we present five hypothetical individuals' employment statuses (employed or unemployed) over 20 intervals (e.g., months). We first code the five individuals' processes over time as sequences, as presented in the figure. Next, we compare them in terms of their dissimilarities. To compare sequences, one should rely on a criterion relevant for the social sciences. Based on a review of the life course literature, Studer and Ritschard (2016) identified three kinds of regularities for consideration. The *sequencing* of the states shows the dynamic and the path taken by individuals. For instance, observing unemployment before or after employment reveals opposite professional integration dynamics. The *timing* of the states (i.e., when an individual is in each state) or transitions (i.e., when an individual experiences a change of state) is also an important regularity in trajectories. Being unemployed at age 20 or 50 has different causes and consequences. Finally, the *duration*, i.e., the time spent in each state, is as well of interest in many applications. For instance, the overall time spent in unemployment is a key determinant for later professional outcomes.

Sequences 1 and 2 are very similar in timing and duration as they are in the same state at the same time for most of the trajectory. However, these sequences record opposite sequencing, thus different dynamics. On the contrary, sequences 4 and 5 show similar sequencing and duration, but different timing. Depending on the chosen criterion, one may reach different conclusions regarding sequence similarity. One therefore needs to make a choice about the criterion to use for comparing sequences, and this choice should be grounded in one's substantive research question. We postpone a detailed discussion of such criteria for comparing sequence similarities to a later section. In the typical SA project described by Abbott and Tsay (2000), dissimilarities are used for data reduction through clustering, that is with the objective of classifying similar sequences together into classes or clusters. However, there are several other dissimilarity-based analytical tools such as identifying representative sequences, measuring the discrepancy of sequences, and ANOVA-like analysis of sequences that have been explored in later developments as will be shown in our review of SA methodological developments.

In Section 2, we begin from a discussion of the birth of SA in the 1980s, with a summary of some important SA research and the historical contexts in which SA was developed by Andrew Abbott. Section 3 contains a thorough review of the many methodological developments in visualization, complexity measures, dissimilarity measures, group analysis of dissimilarities, cluster analysis of dissimilarities, multidomain/multichannel SA, dyadic/polyadic SA, Markov chain SA, sequence life course analysis, sequence network analysis, SA in other social science research, and software for SA. In Section 4, we discuss and reflect on future directions of SA including how SA can benefit theory-making in the social sciences as well as what theoretical approaches may be necessary for making sense of the temporal dynamics uncovered by SA. In this section, we also present the methods currently being developed which are going beyond, and solving some major problems posed by the set of SA methods developed in the last two decades. In the same section, we finally consider remaining challenges facing SA for which we do not yet have any solutions, before concluding the paper with some additional thoughts and calls for additional research.

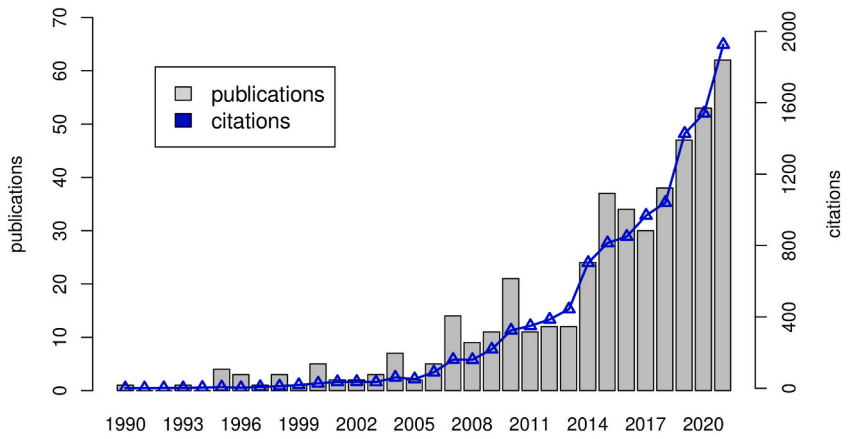

Fig. 1. Journal publication and citation trends of SA applications in life course research (1990–2021, web of science).

2. Sequence analysis: its birth and childhood

2.1. Sequence research in the 1980s and 1990s

The period from 1983 to 2000 witnessed the development of SA for addressing a variety of social scientific questions with processual or sequential relevance—such as questions about occupational or organizational careers. At the micro level, these questions may deal with life course transitions in a certain order; at the macro level, such questions may pertain to organizational developments or modernization processes. To answer questions like these, input data for analysis do not take the usual form of individual data *points* as "cases." Rather, a common property of the type of input data for answering these questions takes the form of data *sequences* as individual "cases." Thus, such a "case" in terms of its outcome measure is no longer represented by a single value but by multiple values or categorical states in SA. This sequence "case" in the first phase of SA development was typically unidimensional, and multidimensional SA did not become popular until later stages of development, to be discussed in Section 3.

During this phase of development, we find two types of SA publications: Those that proposed, expounded, discussed, or reviewed the concepts, framework, and principles of SA and those that applied SA in substantive research analyzing empirical data. Chicago sociologist Andrew Abbott, who adapted and pioneered SA for the benefits of the discipline of sociology and the broader historical and social sciences, published a series of papers adapting and applying SA in sociological and historical research (1983, 1988, 1990a, 1990b, 1995; Abbott and Forrest, 1986; Forrest and Abbott, 1990). Other disciplines took notice as well. For example, Abbott was invited to present a didactic seminar on SA at an annual conference of the Population Association of America in the mid-1990s.

There were numerous SA applications that emerged during this phase of SA development to shed light on a variety of interesting research questions. Good examples abound: Abbott and Hrycak (1990) used Optimal Matching (OM) for analyzing musicians' careers in 18th century Germany; Abbott and DeViney (1992) applied SA via OM to study welfare adoption sequences; Abbott and Barman (1997) analyzed decades of articles published in the *American Journal of Sociology* via optimal alignment enhanced by Gibbs sampling, a member of the broad class of Markov Chain Monte Carlo, to establish the emergence of the standard sociological article structure; Chan (1995) explored how OM could be applied to career data to identify typical career paths and to investigate whether social mobility paths are selective to certain people; Halpin and Chan (1998) applied OM to careers from age 15 to 35 to study work-life mobility; Han and Moen (1999) used OM for understanding the temporal patterning of retirement life in the US; in Stovel et al.'s (1996) study, OM was also applied to model the transformation of career systems at Lloyds Bank in Britain from 1890 to 1970. Note that although a typical SA application in this phase of SA development used OM, other types of distances between sequences were proposed, such as the number of moves needed to turn one sequence into another (Dijkstra and Taris, 1995). For a complete review of SA in this period of development, the reader is referred to Abbott and Tsay (2000).

2.2. A departure from the dominant analytic paradigm in social science?

The earlier section summarized some important social scientific SA publications in the first phase of SA development. It is obvious that one author with centrality stands out—Andrew Abbott—because without his contributions, we would not have SA as we know it today in the social sciences. Perhaps more relevant for our readers is that without an understanding of the historical context in which Abbott became exposed to SA in other disciplines and his own professional interests and trainings, it would not be possible for us to understand his contributions to SA developments. We describe below this historical context, which will answer the question of how it was possible for him to do what he did in those years, and we rely on Abbott's narratives for shedding light on the context in which he pioneered SA developments in the social sciences.

According to Abbott (2022), "What was necessary for [SA] to be 'discovered' for sociology was a sociologist who had several basic qualities:

1. who was profoundly interested in social sequences,
2. who had enough mathematical skills to pursue abstract theory directly through formal reasoning and to talk on some semi-professional basis with people who were much more skilled in mathematics than he and much more broadly acquainted with general developments in formal thinking,
3. who was not committed to any one of the four or five very exciting quantitative paradigms taking off in sociology under the leadership of a remarkable generation of quantitative leading scholars: James Coleman, OD Duncan, Harrison White, Leo Goodman, Nancy Tuma, Robert Hauser, etc.,
4. who had the interdisciplinary connections to encounter people from various other disciplines and hear about new developments in them, and
5. who had the familiarity and experience with programming not to be afraid of dealing with new methods that required direct computation and further coding."

We now further explain Abbott's qualifications in these five areas that together define the context for his pioneering SA developments. The first quality is evidenced by Abbott's substantive interest in histories and processes, beyond what could be represented by the so-called "General Linear Reality" (GLR). This GLR is a way of thinking about how society works, typically embodied by regression-type of linear models that are regarded as representations of the real social world (Abbott, 1988, 2001). A basic assumption of GLR is that the order of things in processes does not affect how things turn out. Abbott has been a strong proponent for going beyond GLR.

At high school, Abbott was trained in the "New Math" tradition, a response to the Sputnik crisis facing the US at the time, to boost students' science and mathematics education. In addition, when Abbott was in graduate school, canned programs like those of today were not yet available, and he did multiple regression with hand calculation in graduate school (Abbott, 2022).

The third quality is equally important as the other qualities and is closely related to the first. The influential figures at the time when Abbott was a junior scholar, James Coleman, Otis Dudley Duncan, Leo Goodman, Robert Hauser, Nancy Tuma, and Harrison White, all had their own followings and represented major paradigms of American quantitative sociological research. According to Abbott (2022), however, he did not recruit himself into any of the above quantification paradigms, but instead pursued various quantification projects of his own, related to intellectual questions that arose in his work such as his early work at a mental hospital, work that required an understanding of order and processes and served as another important aspect of the historical context for his role in SA developments.

The fourth quality follows naturally from Abbott's interest in processes, which took him beyond the confines of sociology at the time to venture into other disciplines such as computer science and history. Some of his contributions to SA, in fact, were published in history journals. He has also been active in the Social Science History Association. His major source of inspiration, however, came from reading the Sankoff and Kruskal (1983) book, a multidisciplinary book on sequence comparison where developments of methods for analyzing sequences in diverse disciplines were discussed such as the chapter on genetic sequence by Bruce Erickson and Peter Sellers (also see Erickson et al., 1980). Fifth and finally, with the early 1980s came the personal computer, and he did all his own computer programming for his OM software in the 1980s and 1990s (Abbott, 2022).

Abbott, as the sociologist with all the five qualities that define the historical context for SA development, stumbled on the Needleman-Wunsch (1970) algorithm when he obtained a preprint of the book by Sankoff and Kruskal (1983), the first book length publication on work using the Needleman-Wunsch algorithm and its variations and drawing on research on DNA, string editing, cryptography, and so on. With the knowledge of the algorithm for optimal alignment, the rest, as we all know, is history, a history of SA developments into the next stage to this day.

3. Sequence analysis: its adolescence and young adulthood

Many of the key issues and major developments of SA since its childhood have been well dealt with and summarized by various special journal issues, symposia, and books, such as *Sociological Methods & Research* vol. 29, 2000; *Sociological Methods & Research* vol. 38, 2010; Sociological Methodology vol. 45, 2015; Blanchard et al. (2014), and Ritschard and Studer (2018a). We cite many of the papers in these special collections devoted to SA throughout the current article, in various sections focusing on a range of SA methodology. We begin from the methods of visualization below.

3.1. Visualization

In the early stage of SA adolescent development, sequence visualization occupied a prominent position. Within the explorative method of SA, visualization is of high importance because it makes use of our brain's visual capacity to find patterns in seemingly chaotic data. The visualization of sequences requires the display of three categorical dimensions: observational units, time points, and states. Altogether, they sum up to a huge amount of information that must be displayed, a challenging task. One can distinguish sequence graphs according to different levels of aggregation: those graphs with the higher amount of information displayed with little aggregation often show a lower level of readability and vice versa. When SA researchers visualize sequences, they typically want to highlight the three kinds of regularities we introduced in the introduction, sequencing, timing, and duration of states that may show the dynamics and the paths taken by individuals. Different types of visualization may show one or more of these regularities. Showing the sequences of many individual observations to find similarities or regularities leads to a graph that is composed of many stacked lines, often to be read from the left to the right. By putting observations' sequences together, we try to arrange sequences in a way that

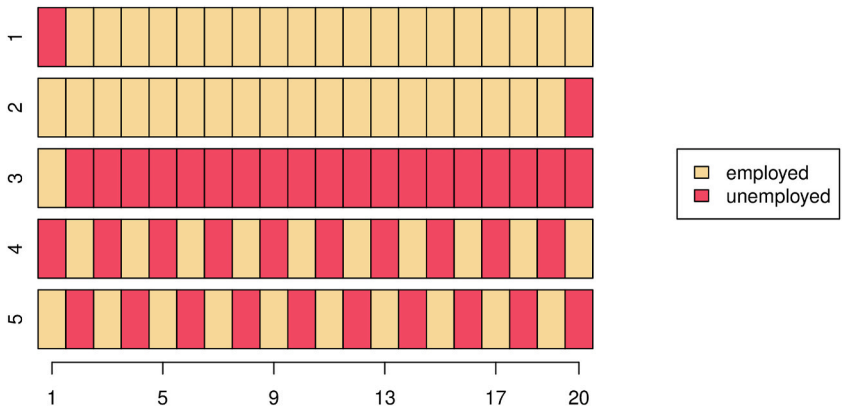

Fig. 2. Five example sequences.

regularities become apparent. In addition, researchers might be interested in aggregate frequencies or cumulations of given states at a given time.

The classic and probably the most intuitive approach to visualize sequences is the sequence index plot introduced by Scherer (2001), where each individual sequence is drawn as a single horizontal line, in which different colors represent different (categorical) states in (discrete) time units, that are sorted from left to the right. This type of visualization shows all the three regularities. See Fig. 2 of the five hypothetical sequences for a simple example of sequence index plots. The set of states used to identify the status of the unit of analysis at each time point is referred to as alphabet (of the states). Sequence index plots are for representing all individual sequences, while some other sequence graphs aim to summarize sequence information, such as relative frequency sequence plots by Fasang and Liao (2014). Since Abbott's introduction of the SA methods to the social sciences, the capacity of computers has increased tremendously. This leads to the fact that in recent SA, capacities in all three dimensions also increased: more observations, more time points, and more states. Consequently, the construction of sequence index plots was modified, with new graph types developed.

Sequence index plots that display too many observations suffer from so-called "overplotting," with the number of observations exceeding the number of available printed lines of the graph. That is why nowadays, sequence index plots are often only plotted with a selected sample of the observations to be shown in the graph. Usually, a random sample of approximately 200 individuals for each graph panel is sufficient depending on the overall size of the graph. Another important aspect of sequence index plots is the order of the individuals: If a classification already exists, sequences in a single graph panel should be ordered accordingly. Alternatively, sequences can be displayed in panels that represent a typology (as shown in the first row of Fig. 3). Another important aspect is the use of color in sequence index plots. Researchers should construct these graphs with the following three considerations. First, a colored version that may use contrasting colors for emphasizing certain states. The use of different shadings can help show commonalities between states. For example, in Fig. 3 the states "full-time work" is indicated in blue while "part-time work," red. Additionally, the property "working in training firm" is shown by the dark version of both colors while "working in another firm" by the light version of the colors. Second, sequence index plots should also be constructed in grey scales because printed journals or books may still require graphs without colors (or charge authors for color pages). Third, approximately 5–10% of human beings suffer from physical color deficiencies. It can be problematic for them to interpret multi-colored graphs because most statistical software packages rely on color palettes by default.

Fig. 3 is taken from Brzinsky-Fay et al. (2016) and shows the three classical types of sequence graphs, namely sequence index plots (first row), state proportion plots (also known as "state distribution plots," second row) and modal plots (third line), with the aim to analyze individual labor market entry processes after vocational education and training (VET) in Germany. For analytical reasons, the authors were interested in whether VET graduates stay with their apprenticeship firm or not, whether they exit employment, whether they start another apprenticeship, and whether they work part-time or full-time. This interest determined the statuses defined shown in the legend at the bottom of Fig. 3. The columns are the trajectory types that resulted from cluster analysis, their labelling is interpretative and aims at giving substantive meaning to the grouped individual sequences. For example, "continuity internal" comprises VET graduates that more or less work full-time in the apprenticeship firm. Although the individual sequences show some variation – which can be seen in the first row, the clustering summarizes them to a common type because of their computed similarities. Therefore,

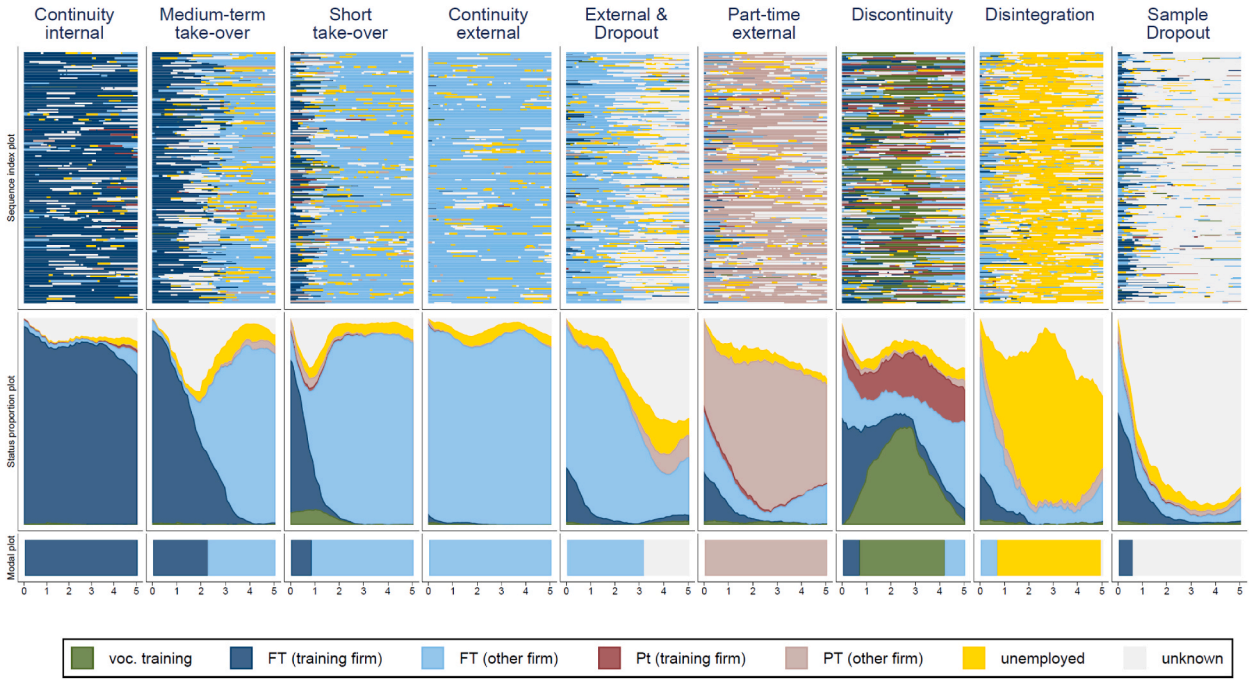

Fig. 3. Example of combination of sequence index plots, state proportion plots and modal plots; taken from Brzinsky-Fay et al. (2016).

we assigned cluster labels or names based on the prominent features of state timing, duration, and order of similar sequences in a cluster.

The second classical graph type used when visualizing sequences is the state proportion plot (or chronogram or state distribution plot), which for each time point shows the relative frequency of states in a stacked bar chart (second row in Fig. 3). This graph shows the aggregate change in the state distribution over time by disregarding individual sequence diversity. In other words, the increase in readability comes at the price of information loss (Brzinsky-Fay, 2014). The most simplified classical graph type is the modal plot (third row in Fig. 3), which consists of only one single horizontal line showing the most frequent state at each time point. Here, the information loss is even higher because all less frequent states are disregarded. However, to allow readers to understand the differences between sequence types, showing all three graph types—as shown in Fig. 3—may be the most helpful strategy because such a graph represents and summarizes sequences and their characteristics.

Because the construction of these kinds of graphs can be complicated and time-consuming, researchers applying SA invented a couple of additional graphs. Such graphs highlight certain sequence characteristics while keeping as much sequence information as possible. Sometimes standard statistical graphs are used and filled with sequence information. However, the trade-off between amount of information and readability remains. Among the most often used information, transition rates stand out. Visualizations of transition rates are constructed by using weighted scatter plots (labelled as transition plots) for overall transition frequencies (Brzinsky-Fay, 2014), by a group of line graphs (called transition pattern plots) for transition rates at each time point (Halpin, 2017), or by directed graphs with nodes and edges (Helske and Helske, 2019). Both visualization types extract information from the data for display. Likewise, the parallel coordinate plot (Brzinsky-Fay, 2014; Bürgin and Ritschard, 2014) focuses on order by focusing on typical sequencings of states and events and the variability of the observed orders.

Gabadinho and colleagues (Gabadinho et al., 2011; Gabadinho and Ritschard, 2013) proposed weighted representative sequence index plots, where not every individual sequence is depicted, but only those who are representative of particular properties, such as frequency, neighborhood density, or centrality. Similarly, Piccarreta (2012) developed smoothing techniques to avoid overplotting in sequence index plots, while Fasang and Liao (2014) invented relative frequency plots for representing while summarizing sequence information. All these additionally developed graph types perform very well with respect to summarizing sequence information although they come with the additional decisions about tuning parameters and/or new metrics that need to be interpreted by researchers who apply these graphs.

Finally, the main challenge to visualizing sequence data is the decision regarding which piece of information to show and which piece of information to disregard. To some extent, this is characteristic of all empirical research described by Occam's razor (or the principle of parsimony). Visualization as a method for SA, which itself is primarily exploratory, should begin by exploring the data using simple tools before moving on to more sophisticated ones.

3.2. *Measuring trajectory complexity*

Akin to visualization, complexity measures also portray trajectories, albeit numerically. Life course trajectories can be complex depending on the sequence length, the number of state changes (or equivalently the number of spells), and the number of distinct states visited. It is of interest to quantitatively characterize this complexity to distinguish, for example, smooth occupational or family pathways from more chaotic trajectories, or to identify deteriorating trajectories, i.e., trajectories that tend towards less favorable states. Number of state changes, number of distinct states visited, entropy of the within-sequence-state distribution, and variance of spell duration can all serve as rough indicators of sequence complexity or diversity within sequences. In Fig. 2 for example, sequences 4 and 5 appear more complex than the other three sequences because of the greater number of state changes and also because of the high entropy of the state distribution in those two sequences.

More elaborated propositions to characterize the complexity of individual sequences were first made by Brzinsky-Fay (2007) and by Elzinga and Liefbroer (2007). Brzinsky-Fay (2007) introduced two indicators—volatility and integrative capability—for analyzing school-to-work transition, and Elzinga and Liefbroer (2007) constructed a turbulence index for studying the de-standardization of family life courses. Volatility and integration capability are based on a state dichotomization into "good" and "bad" states. In school-to-work transition the objective is to find a job and the positive or "good" state is in that case "being employed." Volatility is the proportion of spells in "good" states and reflects flexibility. Integrative capability measures the tendency to integrate a positive state. The turbulence of Elzinga and Liefbroer (2007) is a composite index based on the variance of spell duration and the number of subsequences that can be extracted from the sequence of distinct successive states (DSS), the latter being an indicator of the diversity of the spells. A simpler complexity index that combines proportion of state changes and entropy was proposed by Gabadinho et al. (2010), and Brzinsky-Fay (2018) introduced an objective—not based on dichotomization—volatility, which combines proportion of state changes with proportion of visited states. Ritschard (2021) showed that Elzinga's turbulence may produce counter-intuitive results, because it ignores zero time spent in non-visited states and proposed a revised turbulence to remedy this flaw.

Complexity of a sequence as measured by the turbulence and complexity indexes is typically used as an indicator of the trajectory stability (e.g., Elzinga and Liefbroer, 2007; Christensen, 2021; Van Winkle, 2020). However, it is generally unclear whether complexity (instability) should be considered as something positive or negative unless there is a clear trend. For example, a complex sequence with multiple successively improving changes may reflect a favorable evolution while multiple successively deteriorating changes can represent an unfavorable evolution. Likewise, a full sequence in employment is typically a favorable stable trajectory, and a full sequence in unemployment is an undesired trajectory.

To distinguish between favorable and unfavorable sequences, it is necessary to take account of the nature of the states being analyzed. This is what (normative) volatility and integrative capability do by requiring a dichotomization into good and bad states.

These two indicators are, therefore, able to distinguish more favorable from less favorable evolution. A few other indicators have been proposed to exploit unexplored information such as a preference order (or at least a partial preference order) or desirability degrees of the states. These include precarity (Ritschard et al., 2018), deterioration, badness, and insecurity (Ritschard, 2021) indexes. Formal definitions of all the above-mentioned complexity indicators together with a thorough comparative study can be found in Ritschard (2021).

3.3. Dissimilarity measures

A dissimilarity (or distance) measure between sequences aims to quantify the extent to which two individuals followed dissimilar trajectories. This information is of interest for comparing the trajectories of two individuals or for comparing an observed sequence with an ideal-type sequence to determine, for example, by how much a family formation sequence departs from a traditional model. In addition, and more importantly, computed pairwise dissimilarities between observed sequences allows one to measure the diversity of a set of sequences, to group similar sequences for building a typology of trajectories, to compare groups of trajectories, to compute principal coordinates of sequences, to define neighborhoods, and to identify representative sequences such as medoids. As such, the computed dissimilarities are a key step for subsequent analysis.

Following the seminal work of Abbott and Forrest (1986), OM has been the most used method to assess dissimilarities of sequences. Technically, OM measures dissimilarity between sequences by computing the minimum cost required to transform one of the sequences into an exact copy of the other, by considering two kinds of edit operations: substitutions and insertion-deletion (indel) of states. While such edit operations can be interpreted as mutations in biology and signal changes in information science, they have no straightforward interpretations in social sciences. However, it can be shown that, according to OM, two sequences are considered as similar if they share a long common subsequence, which can be interpreted as a "common backbone" of two trajectories (Elzinga and Studer, 2015).

OM has been highly criticized for the limitations such as the difficulty to sociologically interpret substitution and indel operations, its low sensitivity to the sequencing of the states, and the high number of parameters that can be set by the user (Levine, 2000; Wu, 2000). To deal with these limitations, new distance measures have been proposed (Aisenbrey and Fasang, 2010). These developments were based on sequence properties such as order of spells (Dijkstra and Taris, 1995), within-state distribution (Deville and Saporta, 1983; Robette and Thibault, 2008), expected future (Rousset et al., 2012), set of subsequences (Elzinga and Studer, 2015), or as generalizations of the Hamming distance (Lesnard 2010), and variations of the original OM distance (e.g., Gauthier et al., 2009; Hollister, 2009; Halpin, 2010, 2014; Biemann, 2011; Studer and Ritschard, 2016).

Studer and Ritschard (2016) conducted an extensive review of these distance measures, based on the idea that a distance measure aims to compare sequences. As presented in the introduction, Studer and Ritschard (2016) identified three kinds of regularities for consideration for this comparison: The *sequencing* of the states, the *timing* of the states or transitions, and the *duration*, i.e., the time spent in each state.

As we discussed earlier in the introduction, sequences 1 and 2 in Fig. 2 are very similar in timing and duration as they are in the same state at the same time for most of the trajectory length but have different sequencing. Sequence 1 has improving dynamics while the dynamics is negative in sequence 2. On the contrary, Sequences 2 and 3 have the same sequencing but strongly differ in timing and state durations. Sequences 4 and 5 show similar sequencing and duration, but different timing. Depending on the chosen criterion, one may reach different conclusions regarding sequence similarity. One therefore needs to make a choice about the criterion to use for comparing sequences, and this choice should be grounded in one's substantive research question.

Studer and Ritschard (2016) provided guidelines based on the relative sensitivities of each distance measure to the three above-mentioned criteria. Summarizing their conclusions, OM is mostly sensitive to the duration aspect and somewhat to sequencing. The Hamming distance should be preferred when the focus is given to timing. Finally, OM of transitions (Biemann, 2011), SVRspell (Elzinga and Studer, 2015) and OM of the spells (Studer and Ritschard, 2016) are the most sensitive to sequencing, with SVRspell being the most sensitive to small perturbations in the ordering of states. The latter three distance measures can also be parameterized to hold an intermediary position between these three life course aspects.

During the second-wave SA, the social scientific interpretation of distances has been improved, and new distance measures have been proposed. SA users now have a large number of distance measure choices, and the review conducted by Studer and Ritschard (2016) provides useful guidelines for making choices. In this respect, the distance measure chapter of SA has been concluded. This was a key achievement for SA to be recognized as a mature longitudinal method.

A direct application of dissimilarity measures is for finding representative sequences such as medoids. The medoid as a popular representative is the most centrally located observed sequence, which has the smallest sum of distances to the sequences it represents (Abbott and Hrycak, 1990, p. 165). Another representative of interest is the sequence with the densest neighborhood, that is, for a given radius r, the sequence that has the greatest number of other sequences within a distance r from it. Whatever the representative, a single representative is most often insufficient to render the diversity inside the group it is supposed to characterize. Therefore, Gabadinho and Ritschard (2013) suggested using multiple representatives for each group and proposed a heuristic for identifying the smallest set of representatives of a user-defined percentage of all sequences in the group such that at least the given percentage of sequences lie within a distance r from one of the representatives. Such small sets of representatives are particularly powerful for synthetically characterizing a group of sequences by highlighting typical sequences and, at the same time, the diversity inside the group.

3.4. Group analysis of dissimilarities

Instead of assessing similarities between individual sequences, oftentimes a researcher's interest is in the differences between some fixed or otherwise observed groups, such as gender, race/ethnicity, national origin, social class, and so on, taking advantage of the dissimilarity measures presented above. Luckily, the SA toolkit provides tools for studying the association between sequences and other variables such as fixed groups more directly with two approaches: (1) the discrepancy framework proposed by Studer et al. (2011) that generalizes the principle of analysis of variance (ANOVA) and (2) the approach by Liao and Fasang (2021) that utilizes an adjusted version of the Bayesian information criterion (BIC) and the likelihood ratio test (LRT).

The ANOVA-based discrepancy analysis introduced by Studer et al. (2011) analyzes whether sequences differ across observed groups (e.g., by gender or socioeconomic status groups) by translating the dissimilarity matrix into a measure of discrepancy, which can be conceived of as a measure of variability among a set of sequences. The discrepancy represents the average distance to a group's gravity center, which is defined as the (hypothetical) sequence that minimizes the sum of distances to all sequences belonging to the respective group. It is utilized to quantify how much of the variation in the sequence data can be explained by another variable, but it is also informative by itself as it grasps the degree of variation between the sequences in a group. In the context of life course research, discrepancy can be considered an indicator measuring de-standardization (Brückner and Mayer, 2005) with higher discrepancy scores indicating higher unpredictability of sequences. The techniques proposed by Studer et al. (2011) allow one to test whether discrepancies differ significantly across groups. Thus, one could test, for instance, if male employment trajectories are significantly more standardized—i.e., having a lower discrepancy score than that of their female counterparts. Note, however, that such an analysis would not tell us anything about whether male trajectories are less complex—or less *differentiated* in life course terminology (Brückner and Mayer, 2005)—or more successful. This can only be assessed by examining the measures discussed in Section 3.2, which focus on within-sequence instead of between-sequence variability.

Although the comparison of group-specific discrepancies can be very insightful, the key benefit of the discrepancy analysis framework is that it allows for examining the relationship between sequences and covariates by means of an ANOVA-like variance decomposition. Drawing on a *pseudo-*R^2, this approach quantifies the share of sequence discrepancy that can be explained by one or multiple covariates. The *pseudo-*R^2 can also be utilized as a splitting criterion in a so-called *tree-structured analysis of sequences* (Studer et al., 2011). In this procedure the sample is partitioned into separate groups (nodes) by binary splits until a predefined stopping criterion is met. At each step, the nodes are split by the variable that maximizes the *pseudo-*R^2. This exploratory stepwise strategy is well suited to uncovering differences in the relative importance of covariates across groups. Educational attainment, for instance, might explain much more of the sequence discrepancy in the employment trajectories among men than among women.

While the discrepancy framework's adaptation of the *pseudo-*R^2 makes it attractive for many applicants that are accustomed to R^2-based measures from other contexts, its practicality for SA also spurred some critique that led to the proposition of an alternative approach for studying group differences in sequence data. While the proposition by Liao and Fasang (2021) is also a gravity center-based approach, it utilizes LRTs with bootstrapped samples instead of computationally more intense permutation tests to assess the significance of group differences. The strength of those differences is evaluated by drawing on the BIC rather than *pseudo-*R^2. Liao and Fasang (2021) showed that the BIC difference is preferable to the *pseudo-*R^2 from the discrepancy analysis because the latter often tends to be relatively low. The main clear benchmarks for evaluating the strength of groupwise differences of the BIC are those "levels of evidence" provided by Kass and Raftery (2005).

Although a simulation study by Liao and Fasang (2021) comparing the BIC/LRT approach and the discrepancy framework indicated performance differences in some specific scenarios (e.g., of varying sample sizes), the two strategies could possibly produce similar results as they both rely on the distances to the group-specific centers of gravity. Resting on fewer assumptions than group comparisons based on the results of a cluster analysis (see the next section), both approaches present viable alternatives that warrant any sequence analysts' consideration.

3.5. Cluster analysis of trajectories

As another method of assessing sequence groups, cluster analysis aims to describe a set of sequences by identifying groups of similar trajectories that possibly differ by (various degrees of) misalignments of experienced characteristics such as the differences in the timing, duration, and ordering of processual events. In the previous section, we discussed group differences for groups defined by observed variables such as gender or level of education. The aim of cluster analysis is to partition the sequences into latent groups or clusters that are as homogenous themselves and as different from each other as possible. With cluster analysis, sequence analysts estimate the membership of the groups (clusters) using pairwise dissimilarities between the sequences instead of drawing on other observed variables.

Specifically, hierarchical algorithms (for example, Ward's algorithm) can be applied to cluster cases based on the dissimilarities between sequences evaluated on a given criterion (see Section 3.3). A potential problem of hierarchical clustering algorithms, proceeding with subsequently joining cases and clusters, is the lack of flexibility. Clusters formed at each step are never split, and this might lead to possible distortions or not fully satisfactory results when there are outlying sequences. An alternative, more flexible approach is offered by the partitioning around medoids algorithm (PAM, Kaufman and Rousseeuw, 2005), an extension to the *K*-means partition algorithm when only dissimilarities (rather than measurements on variables) are available. Such an algorithm starts with a random partition of cases; each cluster is represented by its *medoid*, the trajectory most similar to all the others in the cluster. Sequences are iteratively assigned to clusters based on their dissimilarities to the clusters' medoids*, until a convergence criterion is satisfied.*

The *typology* obtained via cluster analysis allows the identification of the most relevant and distinguished temporal patterns in data.

Sequences deviating from the cluster they have been assigned to are usually ignored when interpreting the data-driven types or clusters. Indeed, deviations of trajectories from types are often attributed to sample variation, i.e., to the possible differences in the observed realizations of the same underlying temporal process (Abbott, 1995). Nonetheless, this assumption is valid only when a given partition is reliable. However, cluster analysis is an unsupervised method, which always produces a typology, even when the data are not supposed to be clustered into groups (see Levine, 2000). The quality of the obtained clusters must therefore be carefully evaluated.

Several cluster quality indices can be used for this purpose (see Studer, 2013 for a review). Quality is generally measured according to within-cluster homogeneity and "separation" (or difference) between clusters. In other words, a good typology should be composed of types that are highly homogeneous within types and very different between types. However, these indices lack thresholds values indicating whether the typology is good enough. Parametric bootstrap procedures were developed for SA to provide these threshold values (Studer, 2021). Aside from these indices, a careful evaluation of the typology regarding its within-cluster homogeneity and its between-cluster separation is recommended. It is also important to detect the presence of extreme or peculiar sequences that are weakly related to—therefore not well represented by—their cluster(s) as well as of sequences lying between different clusters. Such analysis is fundamental to identify, for example, rare types underrepresented in the data, and to ensure the quality of subsequent analysis making use of the typology (Piccarreta and Studer, 2019). Outlying sequences can be identified by evaluating the dissimilarities between sequences in a cluster and the cluster's representative sequences, such as medoids.

Aside from cluster quality, other aspects are important for validating a typology. According to Han et al. (2017), a good typology should reproduce known associations with key covariates. They proposed a procedure based on the minimization of the BIC to assess it. Hennig (2007) developed a method for measuring the stability of clustering across bootstraps to capture the idea that a typology should not be sample-dependent. Finally, the interpretability and theoretical soundness of the results is also of key importance for assessing the relevance of a typology (Piccarreta and Studer, 2019).

Cluster analysis is receiving increasing attention within the SA community in the third-wave SA. Aside from typology validation (see Section 4.2), fuzzy clustering has also been used (Dunn, 1973; Bezdek, 1981). This iterative algorithm allows cases to belong to different degrees to more than one cluster. Even if seldom applied to the study of sequences (see Studer, 2018 for application in SA), such algorithm is explicitly based upon the idea of distinguishing among core cases (those having a high degree of coherence within a specific cluster), border cases (those lying between different clusters), and peripheral cases (those having a flat degree of membership in all clusters and therefore not particularly related to any cluster).

3.6. Multidomain/multichannel analysis

In its original formulation (as we have so far reviewed in this paper), SA focuses on processes defined in a *single* domain. Nonetheless, often social scientists are interested in studying trajectories defined in more than one domain. Specifically, joint (or multichannel) SA focuses on the case when multiple trajectories are observed for each individual. The goal now becomes to study how such trajectories—describing, for example, work, family, and housing careers—unfold jointly. Methodologically, multichannel analysis is related to the study of dyadic or polyadic sequences (see Section 3.7), that is the study of the same domain observed for paired trajectories, for example, the work or the family formation histories of parents and of their children.

Broadly speaking, the study of multiple sequences of the same individual requires defining a dissimilarity measure that suitably summarizes the information arising from the set of considered trajectories. Different proposals have been introduced in the literature. A first approach (see, e.g., Aassve et al., 2007; Piccarreta and Billari, 2007; Lesnard, 2008) consists in building a *combined domain*, describing the *combination of states* experienced in each period. It is true that in the context of life course analysis the number of domains analyzed jointly will generally be limited. Even so, the combined sequences can easily be noisy and unstable. Therefore, this approach is convenient only when few, very well-connected domains are considered so that the number of state combinations in terms of the extended alphabet is not too large.

The most popular proposal consists in an extension of the OM algorithm, based on the combination of *costs* defined for the specific involved domains. This idea (Stovel et al., 1996; Blair-Loy, 1999), formalized and systematized by Pollock (2007) and Gauthier et al. (2010), consists in calculating the dissimilarity between two *cases* by averaging the substitutions (and insertion and deletion) costs needed to align the sequences in each domain. This approach, named Multichannel SA by their authors, nicely extends the rationale underlying OM to the case of multiple domains. Also, it preserves the information from each domain, as measured by the specific transformation costs. The approach can be easily extended to any measure of *edit distance*.

Dissimilarities based on multiple trajectories can be used—as in standard SA—to obtain a joint typology via cluster analysis or to apply the dissimilarity-based analyses of group differences described in Section 3.4. It is important to emphasize that a joint analysis of domains—even if reasonable and well-motivated from a substantive point of view—is effective only when the considered domains are associated.

Also, even in the presence of association, joint dissimilarities do not necessarily describe all the domains adequately. This aspect was first considered in a systematic way by Piccarreta and Elzinga (2013), who introduced several measures to quantify the extent of association between *two* domains. Subsequently, Piccarreta (2017) proposed an integrated approach to extend such measures to the case of multiple domains. In addition, Piccarreta (2017) considered the problem of the possible asymmetric performance of a joint analysis across different domains. When applying procedures based on dissimilarities for multiple domains, it is important to consider that some domains might prevail over the others, driving the analysis and influencing results to an excessive extent. This is particularly true when domains are characterized by a different degree of turbulence (or complexity): In these situations, the more stable and less turbulent domains will typically prevail over the others. It is therefore important to evaluate the quality of results both at a joint level and at a domain-specific level. In her work, Piccarreta (2017) described some criteria for assessing whether the results of the

procedures employed to analyze sequences—e.g., cluster analysis, multidimensional scaling, ANOVA, or regression trees—that are based on joint dissimilarities are satisfactory for *all* the considered domains. Piccarreta (2017) focused on cluster analysis, which is often employed to simplify the inspection of the most typical patterns in data and can therefore be particularly useful for describing the joint patterns in data and gaining insights about the relations among domains.

Building on measures introduced by Piccarreta and Elzinga (2013) and Piccarreta (2017), Fasang and Aisenbrey (2021) recently showed how the strength of the association between the two life domains of work and family is a promising indicator for addressing research questions on the extent to which domains such as work and family life courses that unfold over longer periods of time condition and constrain each other. Mantel coefficients that measure the correlation between two distance matrices from two domains show the highest correlation for Black women's work and family lives, and the lowest for White men's, indicating that events in one domain strongly condition or constrain events in the other. The quantitative measure of the strength of the association between two sequence domains can then be complemented by multichannel SA and clustering to illuminate the substantive patterns of cooccurrence behind weaker and lower domain association (Fasang and Aisenbrey, 2021).

3.7. Analyzing dyadic/polyadic sequences

This approach is based on the direct comparison between each sequence with one or more sequences linked by a specific (social) relation. In life course research, the analysis of dyadic sequences (a special case of polyadic sequences) has been prevalent, often used to compare trajectories of family members, such as siblings (Karhula et al., 2019; Raab et al., 2014), a parent and his/her child (Fasang and Raab, 2014; Liefbroer and Elzinga, 2012), or partners (Nutz and Gritti, 2022). For example, one can examine family-life trajectories of parent–child dyads, whose similarities in intergenerational transmission should be larger than similarities between trajectories of unrelated dyads.

Liao (2021) identified three approaches to analyzing dyadic sequences and contributes with a fourth. First, multichannel SA can be used (Gauthier et al., 2010; Pollock, 2007), where each member of the dyad is one of the channels. Second, the so-called grid-sequence analysis can be used (Brinberg et al., 2018), where sequence data are reshaped into grid-sequences. Third, average dyads similarities based on OM or on features-based approaches (Liefbroer and Elzinga, 2012) are compared to the average similarity of unrelated dyads. Finally, the method proposed by Liao (2021) provides an individual measure of linked lives by distinguishing between dyadic distance and degree of dyadic linkage compared with randomly constructed dyads. Further, compared to the other strategies, this method can be more easily extended to polyads, and it enables the identification of the separate effect of timing, duration, and order in the similarity between the members of a dyad or a polyad.

The dissimilarities between linked trajectories can be used in a regression framework as dependent or independent variables: For example, one could examine family-life trajectories of parent and child dyads, whose similarities in intergenerational transmission should be larger than similarities between trajectories of unrelated dyads. Alternatively, such dissimilarities can be the input of a cluster analysis to identify ideal-typical patterns of a certain process.

There are certain data requirements for this kind of analysis. First, the length of the dyadic sequences to be compared as well as the alphabet of the states for both members of the dyads must be the same. Second, the dyads under study must be unequivocally identified as such in the data. In the domain of life course research, these two conditions are satisfied by a relatively small number of data sources. However, in the case of survey data a multigenerational design is required ex ante (e.g., Longitudinal Study of Generations data used by Liao, 2021 and Fasang and Raab, 2014) while in the case of register data the identification of family members and the link between the records can be ex post (e.g., Finnish register data used by Karhula et al., 2019).

Finally, researchers must make an additional decision when analyzing polyadic sequence trajectories, that is, whether the pairwise relations between the members of a polyad should be treated as equally important when computing dissimilarities. For friendship network members, pairwise relations may be considered as equal. However, for a family triad, the resemblance between the parents may not be the same as the father–child or mother–child resemblance and if not, differential weighting should be applied (Liao, 2021).

3.8. Markovian analysis of sequences and alternatives

Abbott (1995) distinguished between the approaches considering the entire sequence as the whole unit of analysis versus a step-by-step process. In the latter case, the aim is to "fit sequences of categories by estimating transition probabilities step by step" (Abbott, 1995, p. 104). Markovian models (MM) belong to this class of methods alongside the more general class of multistate models (e.g., Meira-Machado et al., 2009). MM are stochastic models for the analysis of the transitions between successive states in sequences, with the aim to describe how the current distribution of the possible values of a variable of interest depends on the previously observed values. As such, MM are in particular of interest for analyzing the dynamics driving the unfolding of the sequences. Fitting a MM on each of the sequences in Fig. 2, we would find a high probability to stay in the previously observed state for the first three sequences while we would estimate this probability as zero for sequences 4 and 5.

The Markov chain (MC) is the simplest Markovian model possible. In its conventional formulation, the next modality of the variable depends only on the current state that is assumed to summarize the relevant history of the individual (the *Markov property*) and the probability of switching from a given state to another is assumed to remain unchanged over time (*time homogeneity*). The homogeneity assumption and having a memoryless process make the basic MC easy to compute but it can be unrealistic in many social scientific applications.

The simple first-order homogeneous MC can be extended in several ways: such as increasing the order of the process (accounting for two or more of the previous states), allowing variable-order (e.g., Begleiter et al., 2011; Gabadinho and Ritschard, 2016), and

including both time-varying and time-invariant covariates. Moreover, while traditionally MMs are in discrete time, it is possible to formulate a Markov process in continuous time. Similarly, it is possible to extend them from a categorical outcome to a continuous one.

One of the most interesting MM extensions for the analysis of sequential data in social sciences is to include a hidden or latent variable that can be time-constant or time-varying (e.g., Vermunt et al., 2008; Helske and Helske, 2019). Adding a time-constant latent variable leads to the mixture Markov model (MMM), where we assume that the data consists of latent subpopulations with varying patterns. The MMM is typically used for clustering sequences (Barban and Billari, 2012; Han et al., 2017) and can also simultaneously link covariates to clusters—with the covariates used to predict cluster memberships or transition probabilities, either time-constant or (in the case of transition probabilities) time-varying. Unlike the typical distance-based clustering methods used in SA, where each individual is deterministically assigned to one cluster, in an MMM, we have a probability distribution of state membership. The group membership is derived from the stochastic process that links the observed outcome to the latent variable.

An MM with a time-varying latent variable is called the hidden or latent Markov model (LMM, see e.g., Zucchini and MacDonald, 2009). With an LMM, we can analyze how the time dependence between observable states is governed by a latent process (e.g., Bolano and Berchtold, 2016). This is particularly useful in life course studies where many non- or hardly observable aspects such as motivations, beliefs, or levels of frailty may influence or explain the observed behavior (Bolano et al. 2019; Piccarreta and Studer, 2019; Han et al., 2020). Combining the properties of the MMM and the LMM leads to the mixture latent (or hidden) Markov model (MLMM), which has a time-constant latent variable (related to estimated latent subpopulations) and a time-varying latent variable (related to modelled time dynamics). Another extension is the double chain MM, which allows for dependencies between both the observed and the latent states (Berchtold, 1999).

While MMs are well-established approaches, some practical aspects still need to be further developed to be fully applicable to the analysis of sequence data in the social sciences. First, from a computational point of view, MM typically requires the estimation of multiple parameters. The EM algorithm is a common approach to estimating Latent MMs, an approach that, despite its advantages, heavily depends on the starting values and is quite unstable in high-dimensional settings, requiring the researcher to estimate the model numerous times from different starting values. Alternatives exist but unfortunately no universally best optimization method is available (Taushanov and Berchtold, 2017; Helske and Helske, 2019). Also, estimation time can be long for complex models, making the use of MMs less appealing to scholars. Second, in terms of representation of the results, in the presence of multiple states (either observed or latent) or with a high-order Markov process, the results are not easily readable. In the literature, some attempts have been made for providing a graphical representation of the transition probabilities with visualization tools such as the suffix trees (Gabadinho and Ritschard, 2016) or directed graphs and stacked sequence plots (Helske and Helske, 2019). However, comprehensive graphical tools for representing the empirical results need to be further developed. Finally, multivariate (multichannel) data allow the researcher the opportunity to study joint evolutions of multiple outcomes over time. In their traditional formulation, Markovian processes are used to describe the behavior of a single univariate time series. Although some extensions to analyze multivariate data with multichannel sequences have been proposed in the literature (Helske and Helske, 2019), a unique framework for such models is still missing.

To summarize, the difference between an MM and a SA approach in the analysis of sequence data lies not only in having a probabilistic model-based approach versus a deterministic data mining one but also in the way of examining the sequence of states. SA takes an overall approach with trajectories as a unique holistic unit of interest, what we may call a broad approach. MM applies instead a narrow, transition-specific, approach focusing on the different transitions happening during the life of an individual and the factors (i.e., covariates) that may explain the probability of experiencing the transitions (Bolano and Berchtold, 2021), what we may call a step-by-step approach. The two approaches should not be considered as direct "competitors" but instead different ways to understand different aspects of the phenomenon under investigation. Integrations of SA and Markov based approaches have also been proposed in the literature. For example, a two-step analysis combining SA and MM has been proposed as a workaround solution for the issue of complex models with an unknown model structure (Helske et al., 2018; Helske et al., 2021) and the Sequence Analysis Multistate Model has been used for modelling the relationship between time-varying covariates and trajectories of categorical states (Studer et al., 2018b).

3.9. Sequence analysis and life course research: theory and applications

Despite its early applications being in cultural and historical sociology (Abbott and Forrest, 1986; Bearman and Stovel, 2000), it is no surprise that SA has proliferated and matured most in the broader life course field, including expanded labor market research and family demography. As early as 1992, Abbott stated that "There are also literatures where process is of obvious conceptual importance, but where empirically no one has moved beyond the recipes of regression, the life course literature being the best example." (Abbott, 1992, p. 184). Three decades later, the life course literature can be viewed as the central driver of theoretical, conceptual, and methodological advances of SA in the social sciences.

The success of SA in life course research originates in the core theoretical ambitions of life course research and its tradition as a multidisciplinary hub for methodological development. First, trajectories of categorical states, not just metric outcomes, are of central theoretical importance when studying the life course (Billari, 2001; Fasang and Mayer, 2020). Theoretically, life courses exemplify *process outcomes* as defined by Abbott (2001, p. 176) as "… long run stabilities established by myriads of individual events… it is the whole walk that is the outcome." Second, the life course field has a strong tradition as a hub for multidisciplinary methods development for longitudinal data analysis, both qualitative and quantitative. The life course field has been characterized by a nondogmatic approach of innovatively combining different methodologies to capitalize on their potential to complement one another to inform substantive questions, rather than fighting over their relative merits and flaws (Mayer, 2009).

SA has been fruitful to both address core principles of the life course paradigm (life-long development, timing, time and place,

linked lives, agency, Elder et al., 2003) and to refine more specific theoretical arguments within each broader principle.

Considering human development as a long-term process over varying life stages, studies using SA techniques have shown for instance long-term effects of family life trajectories on income (Bolano and Studer, 2020; Muller et al., 2020) and health in midlife (Barban, 2013; O'Flaherty et al., 2016). In a similar fashion, SA has been instrumental in understanding the process of accumulation of (dis)advantage over the life course. The timing of life events and their temporal order differentiate one life trajectory from another. Looking at the (sub)sequence of states in a SA perspective allows studying their correlates and consequences (Furstenberg, 2005; Jalovaara and Fasang, 2020). Similarly, SA can be applied to the study of social norms and the "appropriate" time and order of life events.

Life course agency rests on a belief in one's capacity to achieve life course goals, and the "ability to formulate and pursue life plans" (Shanahan and Elder, 2002, p. 147) while facing internal and external constraints and opportunities. Recent theoretical advances in life course research have centered on how the predictability and stability of life courses impacts on time horizons for individual decision making, where SA can take a double role in assessing the instability and predictability of life plans and patterns of time use that can reflect individual agency on a small scale (Lesnard, 2008). SA of time use data as well as the structure of individual life courses therefore is promising to further inform theoretical ideas about agency across the life course and its manifestation in everyday life or to directly assess timing or duration in the sequential model of action phases (selecting, engaging, or disengaging) proposed by Heckhausen and Buchmann (2019) with appropriately fine-grained micro-level data on decision making processes. Macro-structural contexts and social policies set constraining or enabling opportunity structures for individual agency that jointly shape life courses. Multi-country studies have shown how and to what extent macro factors influence life trajectories both in developed (Fasang, 2012; Madero-Cabib et al., 2021) and in developing countries (Pesando et al., 2022).

Other aspects that play a role in shaping life courses and that can be successfully analyzed using SA are: (1) the interdependence among individuals (the concept of linked lives) with dyadic SA (see Section 3.7) as an appropriate way to address such interrelationships (Kalucza et al., 2021; Fasang and Raab, 2014); and (2) the intertwinement across life course dimensions (Bernardi et al., 2018; Elder et al., 2003; Huinink, 2005), such as work and family, or health and migration, that mutually condition and constrain each other (Fasang and Aisenbrey, 2021). Multichannel SA (see Section 3.6) is a way to study such relationships.

To summarize, within life course research, SA has been used to address theoretical questions about: (1) *within life course variability* in terms of the temporal structure of individual life courses, first explicitly conceptualized as life course differentiation (Brückner and Mayer, 2005); and (2) *between life course variability* looking at (dis)similarity across individuals to assess the degree of de-standardization and pluralization of life courses and to identify typologies of ideal typical life courses that can be compared across cohorts and countries (Van Winkle and Fasang, 2021; Van Winkle, 2018). The latter approach to assessing similarity between life courses also has so far untapped potential for identifying "outlier" life courses.

The need for new analytic approaches to better understanding these two sources of heterogeneity have also pushed the development of new SA tools in the young adulthood of SA, and it is expected to influence the methodological development in the field even further. Concerning *within life course variability*, initial metrics only assessed the temporal structure related to the number of transitions and the duration in different states, mostly with the sequence complexity index (Elzinga and Liefbroer, 2007; Gabadinho et al., 2010). Not attaching any values or weights to socially positive (employment) or negative states (unemployment) plagues the interpretation of these indices. Recent developments that allow attaching positive or negative weights, for example in badness or precarity indices (Ritschard, 2021) greatly improved the ability of SA to inform theoretical questions on zigzagging processes with back-and-forth movements, instability, volatility and precarity, or upward/downward mobility. Bolano and Studer (2020) proposed to identify which specific aspects of the trajectory (timing, order, and sequencing of events) are relevant for explaining an outcome later in life by combining SA with data mining techniques. These developments offer new possibilities to address path dependence and the contentious concept of "turning points" in the life course literature (Abbott, 2001 p. 251). Concerning *between life course variability*, pairwise sequence distances can directly assess the degree of de-standardization of life courses, identify ideal types, and evaluate, for example, divergence and convergence of life courses across cohorts (Liao and Fasang, 2021). Concepts at the core of structural-institutional life course research developed in Europe to study social change through cohort replacement and to link theories of modernization and grand societal transformations to life course pluralization and de-standardization. Recent proposals of combining SA with event history analysis (Studer et al., 2018b) and adapting the Bayesian Information Criterion and the Likelihood Ratio Test for sequence comparisons (Liao and Fasang, 2021) to study time-varying period effects on life course patterns and the convergence and divergence of life courses under different state systems across cohorts are good examples for methodological developments for life course research.

3.10. Sequence network analysis: theory and applications

Although a lot of SA analysts do life course research, some sequence researchers take advantage of a social network framework to analyze sequences (e.g., Bison, 2014; Cornwell 2015, 2018; Cornwell and Watkins, 2015; Hamberger, 2018), offering an alternative to classic SA methods (Courgeau, 2018; Ritschard and Studer, 2018b). With this "sequence-network" approach, analysts view sequence elements (i.e., events, activities, and actors) as nodes that are connected to one another in a network according to their sequenced (e.g., temporarily ordered) nature. Social network techniques (Wasserman and Faust, 1994) can therefore be employed to analyze sequence data to yield new insights into the interconnected structure of sequence elements. This is akin to the "narrative network" (Bearman and Stovel, 2000), "event structure" (Dixon, 2008), and "practice network" (Higginson et al., 2015) approaches, which map links among different individuals' accounts of which events, activities, or practices follow which, over a given period of time.

Recent applications use the network approach to make sense of complex time-stamped data. Given a daily time diary provided by a

person, for example, their activities throughout the day can be linked by "ties" or "arcs" (e.g., that person reports "eating" at 6 p.m., followed by "watching TV" at 6:30). As additional actors are included, there emerge more links across activity chains (e.g., the next person reports "eating" at 6 p.m., followed by "cleaning" at 6:30), thus revealing alternative pathways between elements. A larger network of sequence chains thereby emerges, revealing linkages not just among elements in contiguous sequence time periods but also among elements in noncontiguous time periods that share indirect connections. This approach allows scholars to model sequenced phenomena in a nonlinear fashion, such that events, activities, and subjects can be linked regardless of when they occur in a sequence.

Using this approach, analysts can focus not only on questions like "How similar to each other are people's everyday lives?" but also questions like "Which activities at what times serve as the points around which the similarity in people's everyday lives derive?" In network terms, this is a question of which sequence element or positions are most "central" or most connected to each other in the overall sequence network. Finally, network programs allow users to map and thus visualize otherwise hidden connections across sequences and actors (e.g., Higginson et al., 2015).

The network approach has been used to understand sequence structure in a variety of phenomena that are embedded within sequences, including eating practices (Castelo et al., 2021), travel patterns (Zhang and Thill, 2017), gendered career mobility patterns (Hamberger, 2018), organizational work practices and routines (Goh and Pentland, 2019; Mahringer, 2022; Mahringer and Pentland, 2021), and other social phenomena (for an overview, see Pentland et al., 2017). A particularly important application has been to residential electrical energy consumption (see Lőrincz et al., 2021; McKenna et al., 2020).

Due to growing environmental and energy concerns, researchers have begun to focus on approaches to reducing energy demand spikes related to people's activities at certain times and to shifting individuals' activities to different times or days to reduce aggregate energy costs. Researchers have argued that it is not enough to look only at which activities occur at which times, but also how people sequence their activities. This work relies on individual-based time use data to examine the implications of different activity schedules, or "activity sequences," for aggregate energy demand. McKenna et al. (2020) showed that people's activity sequence networks during the weekend are denser and are less centralized than are weekday activity networks. The more varied and less organized sequences of weekends signal more flexible, or less rigid, activity routines, thus providing more opportunities for household-level interventions to alleviate electricity load problems. This work also shows that how activities are sequenced within the household are heavily patterned by work-related schedules. Studies of energy consumption patterns have begun to focus on work scheduling as a potential point of intervention. The network approach is especially capable of recognizing heterogeneity in activity sequences, thus allowing for interventions that are tailored to different "cohesive subsets" of people who have varying degrees of flexibility in their schedules (see Lőrincz et al., 2021).

The emergence of the sequence-network approach has also led to the infusion of sequence concepts into the area of social network analysis (e.g., Liao, 2021). This can be seen in the growth of the "relational event framework" (e.g., Butts and Marcum, 2017; Schecter et al., 2018)—a statistical approach to modelling dependencies among temporally sequenced activities within socially networked settings. It is also evident in recent calls to use sequence approaches to understand complex over-time dynamics in the structure of networks (e.g., Kim et al., 2019; Nee et al., 2017). Future work will need to focus on making the sequence-network approach more relatable to scholars who are unfamiliar with complicated social network techniques and tightening the link between respective bodies of research on sequences and networks.

3.11. Sequence analysis and other social science research: theory and applications

A systematic complete review of the theoretical approaches and research published in the broader social sciences where SA has demonstrated to have heuristic power is beyond the scope of the current paper. The interest in the SA toolbox arises from the need to account for how processes of different kinds unfold over time and why different patterns emerge. Classical sociological theories interpret the existence, persistence, and predictability of certain patterns and routines of social action (Giddens, 1986) and roles (Parsons, 1951) as indicators of the link between individual experience in the different domains of the social worlds (Simmel, 1955) and social institutions. These considerations can be extended to other disciplines to the extent that they are concerned with the variation in temporal patterns of a given process although substantively they draw from other theoretical traditions and explanations.

In political sciences, SA sequences analysis has been applied to different units of analysis: individuals, organizations, movements, or institutional processes. Three main types of sequences are analyzed (Blanchard, 2019). First, individual careers within institutional environments (e.g., Claessen et al., 2021); second, trajectories of political participation via elections, social mobilization, or conflict (Buton et al., 2012); finally, processes involving unconventional units of analysis, such as interactions between institutional actors (e.g., Casper and Wilson, 2015) or the legislative steps for the parliamentary approval of a certain law in different countries (e.g., Borghetto, 2014).

Spatial disciplines have explored patterns in residential mobility (Stovel and Bolan, 2004), commuting patterns (Mattioli et al., 2016; Brum-Bastos et al. 2018), tourist trips (Shoval and Isaacson, 2007), changes in the social composition of neighborhood (Le Goix et al., 2019), and land use (Mas et al., 2019). In these applications, SA helped make visible the link between spatial mobility (broadly understood) and time.

In survey methodology, SA was used for different purposes with important implications for survey management and survey monitoring. For example, Durrant et al. (2019) employed SA to identify unusual interviewer calling behaviors, and Wahrendorf et al. (2019) used SA to compare the performance of record linkage of administrative data and retrospective interview data on employment trajectories. Finally, Lazar et al. (2019) studied the impact of different types of missing survey data in joint SA (see also Kreuter and Kohler, 2009).

3.12. Software for sequence analysis

At the early stages of SA, Andrew Abbott provided a self-contained program named "Optimize" for computing OM distances. The outcome of the program was then inputted in statistical systems such SPSS or SAS for running cluster analysis. OM was also implemented in the TDA (Transition data analysis) software of Rohwer and Pötter, 2002. The program "Sequence" by Dijkstra (1994) proposed several tools for dealing with sequences and computing the similarity measure described in Dijkstra and Taris (1995). In the early 2000, Cees Elzinga wrote CHESA, a program that computed OM and several alternative distance methods as well as the turbulence index. SA with these programs remained somewhat cumbersome because the user had to export the computed distances to another statistical environment for running further analyses.

It appeared that developing packages for statistical environments such as Stata and R has a couple of major advantages: It allows running complete analysis in the same environment, and it gives access to the plotting capabilities of the environments for visualizing sequences data, therefore much facilitating automatization of complete analyses. The main SA toolkits available today are thus the SQ (Brzinsky-Fay et al., 2006) and SADI (Halpin, 2017) modules for Stata and the TraMineR (Gabadinho et al., 2010) R package. All these packages offer a variety of sequence visualization tools, compute OM and other distances between sequences as well as individual sequence summary variables. TraMineR is the most versatile package with, for example, tools for ANOVA-like analysis of sequences, growing regression trees of sequences, and identifying representative sequences. In addition, TraMineR has two useful companion packages: TraMineRextras (Ritschard et al., 2022) that provides a few additional plots (e.g., survival plot and relative frequency plot) and a series of ancillary functions such as polyadic analysis and the computation of BIC values for comparing groups of sequences, and WeightedClusters (Studer, 2013) that provides, among others, tools for clustering and evaluating cluster quality of sequence data and for rendering clustering trees of sequence data. Other useful packages are MICT (Halpin, 2016), a Stata module dedicated to the imputation of missing elements in sequences, seqHMM (Helske and Helske, 2019), an R package for fitting mixtures of hidden Markov models with plots for rendering multidomain sequences and transitions within sequences, PST (Gabadinho and Ritschard, 2016), an R package that fits and renders probabilistic suffix trees of sequences, and ggseqplot (Raab, 2022), an R package for visualizing sequence data using ggplot2 (Wickham 2016). The modules and packages listed are all maintained by scholars or teams of scholars who continuously contribute to SA. TraMineR, in particular, is regularly updated two or three times a year. Therefore, we can confidently expect that these packages will continue to be supported.

4. The future of sequence analysis

4.1. Potentials of sequence analysis for theory development

As an initially exploratory technique in the tradition of data mining, SA was often seen as a rather theoryless method. Yet, to date it has proven vital in assessing research hypotheses and theoretical arguments, particularly in life course research and family demography concerned with outcomes that unfold over longer periods of time. To name just a few examples, SA studies have substantiated critiques of the second demographic transition thesis arguing that the highly educated are the vanguards of family change with increasing family complexity, a decline of marriage and delay of fertility. In fact, family complexity over longer periods of family life courses has increased most among the lower educated in most countries (Van Winkle, 2018).

Recent studies started to spell out more precise theoretical mechanisms of how country contexts shape the interplay between work and family lives (Aisenbrey and Fasang, 2017; Fasang and Aisenbrey, 2021) and how life course patterns captured in sequence typologies are associated with a fanning out of economic rewards as within cohort differentiation with age or cumulative (dis)advantage (Gruijters et al., 2022; Buyukkececi et al., 2022). In developing cumulative disadvantage and the Matthew effect in scientific careers, Merton (1988) noted early on, that any such studies require a process perspective moving beyond single time point analyses. Many outcomes of theoretical interest in life course and stratification research are not only metric (income) but categorical (employment, benefit receipt) in nature (Fasang and Mayer, 2020). SA therefore is promising for more comprehensively assessing cumulative disadvantage as a fanning out of socially valued goods and the structural conditions that shape cumulative (dis)advantage processes across countries and birth cohorts.

More generally, SA is particularly useful for testing and developing theoretical arguments about the speed, order, and timing of processes. These include questions on whether the same processes experienced at different speeds have different correlates and consequences, how different structural contexts shape longer term processes, why specific processes in multiple domains tend to cooccur or preclude each other, and for whom and under which conditions processes are particularly volatile involving back-and-forth movements.

Most SA applications have focused on individuals as units of analysis. Individuals do not change over time in ways that households or organizations do through splitting up or acquiring new members. Abbott (2016) referred to this as the "historicality of individuals" given by their bodily integrity. But there is untapped potential of SA for informing theoretical ideas that put other units of analysis at the center, which only a few studies have to date explored. A handful studies focused on couple trajectories (e.g., Visser and Fasang 2018; Lesnard 2008; Möhring and Weiland, 2022). In an earlier application, Lesnard (2008) used SA to capture off-scheduling of couples, when partners take turns being at home and working and have very little overlap in time use, to show how the lack of joint family time is associated with negative outcomes for families. Concerning the temporal order of macro-structural processes, such as the introduction of certain types of government, social policies, and economic restructuring, Abbott (1992) presented a pioneering study on the order in which countries introduced different social policies. Recently, Gjerløw et al. (2021) used SA to inform theoretical arguments about democratization and economic development via an analysis of institutional histories (see also Wilson 2014).

A core question is whether SA can inform causal mechanisms. Most SA applications quickly put forward disclaimers that their results are merely descriptive and cannot be interpreted causally. However, strictly this is the case for all nonexperimental methods using survey or administrative data. Recent innovations as the Sequence Analysis Multistate Model (Studer et al., 2018a, 2018b) for modelling the impact of time-varying covariates on the likelihood of transitioning into specific processes with frailty terms that comes as close to causal assessments as fixed effects models. Moreover, SA has been fruitfully applied to augment matching methods for causal inference, with survey data matching not only on time invariant covariates but on entire pre-treatment trajectories (Barban et al., 2017). Moreover, methods for causal inference, such as difference-in-difference models, often rely on untestable assumptions and deliver a "causal" effect of certain magnitude without being able to clearly adjudicate between different mechanisms that generate this causal effect.

Another way of linking empirical evidence to theoretical arguments about causal mechanisms is to reconstruct the order and sequencing of events that happen along a causal chain with a detailed description of processes. This may not yield a causal effect of an exact size but can be qualitatively informative about the actual mechanisms at play. Such a logic resembles process-tracking methods in small-N case studies in historical sociology or political science. SA enables taking the logic of quantitative micro-level processes by providing sophisticated descriptive evidence on processes that can either be in line with certain theoretical arguments or contradict them. As noted above, family life course studies using SA have offered a way for assessing several core tenets of the second demographic transition thesis.

4.2. Sequence analysis: methods currently being developed

Besides the advances highlighted in the previous sections regarding new algorithms for measuring dissimilarities or refined versions of synthetic complexity measures, we focus here on the advances that aim to go beyond the second-wave SA (Aisenbrey and Fasang, 2010). What can be considered as the "third-wave" SA (Raab and Struffolino, 2022) has focused so far on attempts to bring together the stochastic and the algorithmic modelling cultures and to introduce heuristics for validating cluster typologies.

4.2.1. Combining sequence analysis with other methods

Although SA has benefitted from several other methodological traditions—cluster analysis, principal coordinate analysis, network analysis (Cornwell, 2015, 2018), Markov models (Helske et al., 2018; Helske and Helske, 2019; Piccarreta and Bonetti, 2019), analysis of variance (Studer et al., 2011), and the combination of traditions (such as cluster with qualitative comparative analysis (Borgna and Struffolino 2018)—three major problems have remained unaddressed. First, in most applications it is advisable to use sequences of equal length, as a result, censored observations (e.g., individuals followed for a different number of years or otherwise with different record lengths) must be excluded, thus having several methodological and substantive implications for sample selection. Second, in the standard SA cluster analysis workflow where the cluster typology is used as independent variable in a regression analysis, it is not possible to include time-varying covariates: Typically, the researcher can estimate the effect of those variables measured prior or at the onset of the sequences only. Finally, the relatively limited ability to establish causal relationships (however, see the previous section for a detailed discussion).

A few recent contributions have addressed some of these limitations by combining SA and event history analysis (Studer et al., 2018a, 2018b; Rossignon et al., 2018). The analytical strategies described in Studer et al. (2018a, 2018b) share a key feature: They enable the inclusion of time-varying covariates, although by moving away from the holistic perspective of the standard SA approach. Instead of looking at entire sequences, these methods analyze transitions into subsequences of shorter (but still equal) lengths of the initial sequence, which can be of different length as only part of them will be used in the cluster analysis for identifying a typology of the process of interest (e.g., pathways of six-year-long after exiting unemployment or after finishing education). The time-span antecedent to the subsequence is accounted for in a competing risk model, where time-varying and time-invariant variables can be included in this time span. In Rossignon et al. (2018), cluster membership of sequences preceding an upcoming event of interest is used as a covariate in an event history analysis.

The combination of SA with matching procedures has the objective to approach causal designs, therefore enabling the estimation of causal relationships. Different types of matching procedures were used to balance characteristics between treatment and control groups. For example, Raab et al. (2014) and Karhula et al. (2019) applied exact matching to compare unrelated and sibling dyads who were identical in the covariates. Drawing directly on the SA results to implement propensity score matching, Fauser (2020) identified a typology of employment trajectories and then estimated propensity scores for selecting into different career patterns while Barban et al. (2017) used standard variable-based propensity score matching, dissimilarities of pre-treatment trajectories obtained by OM, and a combination of these two strategies to study the consequences of age at retirement on subsequent health outcomes.

4.2.2. Validating cluster typologies

A main criticism of the first-wave SA concerned the validation of cluster typologies. The reliability of a specific clustering algorithm in recovering the number of latent groups of trajectories in the data cannot be assessed in a statistical sense (see, e.g., Warren et al., 2015). This applies to every application of cluster analysis, but it is even more relevant for SA because the typically high number of observed trajectories—usually analyzed only graphically—additionally complicates the detection of peculiar or isolated cases deviating from their cluster. Cluster quality criteria (Studer, 2013) combine indicators for within-cluster homogeneity and between-cluster separation and are generally used to guide the researcher in the choice of the number of clusters to be extracted. One attempt to use these criteria for further cluster validation is to exclude cases with negative average silhouette width values to avoid sequences having a weak similarity to the others in the cluster to affect subsequent analysis (Jalovaara and Fasang, 2020). However, as noted by Studer

(2021), these criteria do not test the no-clustering solution (i.e., the lack of a null model), and the values are difficult to interpret in absolute terms when it comes to evaluate the strength of the structure identified. To address these limitations, Studer (2021) adopted the framework proposed by Hennig and Lin (2015) for parametric bootstrapping for comparing the quality of a given clustering with the quality of having no-cluster. This approach allows testing different typologies against the "null case" of non-clustered data by using bootstrapping and provides a baseline value for the interpretation of the thresholds of the cluster quality criteria.

4.3. Sequence analysis: methods to be developed

Although SA has now reached young adulthood with a solid core of well-established methods, there remain a series of open issues deserving further investigation.

4.3.1. Missing data and sequences of different lengths

First, there is the important question of missing data in sequences. It is important in SA to distinguish between data that could not be collected, because of nonresponse for example, from data that do not exist. Missing data of certain experienced states lead to holes in sequences that can possibly be filled through imputation. Nonexistent states correspond to intervals of observation after the period of data collection and typically lead to sequences of different lengths. For example, ages over 22 years in 2022 cannot be observed for cohorts born in 2000 and later.

Basic approaches often adopted for missing data consist of either dropping incomplete sequences or allowing for sequences with a limited percentage of missing states by treating the missing state as another state in the alphabet. These approaches are known to bias SA outcome, especially for states that are not missing at random.

Imputation techniques specifically designed for filling holes in sequences have been proposed in the literature. For example, Halpin (2016) proposed to infer the imputed value (or its distribution for multiple imputation) from the surrounding valid states in the sequence, and Gabadinho and Ritschard (2016) used the prediction of the probabilistic tree (variable-length Markov model) fitted on the subsequence preceding the missing state. Although such techniques have proven useful, there is a need for easily useable imputation methods that, in addition to the information on valid states in the sequences, would also take into account contextual information such as values of constant or time-varying covariates as well as the non-response mechanism.

Several SA tools are technically applicable to sequences of different lengths. For example, dissimilarity measures that allow for time warp such as OM and many of its variants can compare sequences with different lengths. Normalization of dissimilarities is another attempt. Likewise, normalized complexity indicators are comparable among sequences of different lengths. Nevertheless, comparison only makes sense when the difference in length remains limited. For example, it would not make sense to compare the complexity of a sequence of length 2 or 3 with a sequence of length 20, and the distance between a sequence of length 2 with one of length 20 would essentially reflect their difference in length rather than their difference in content. Here, there is a need for criteria to guide decisions about the maximum length difference that can be afforded given the size of the alphabet and the kind of analysis envisaged.

4.3.2. Big data sequences

SA applies without major difficulties to sets of several thousand sequences. However, some methods, the computation of pairwise dissimilarities in particular, may become impossible because of memory and computation time limitations when the number of unique sequences exceeds, say, around 20,000. This becomes an issue, for example, when the researcher wants to analyze sequences derived from census data that may involve millions of cases. A workaround is to resort to sampling approaches to analyze one or multiple randomly chosen samples. Currently, with the exception of clustering algorithms such as CLARA (Kaufman and Rousseeuw, 2005) there is a lack of tools for SA using this approach to deal with large samples. Another possible direction lies in parallel processing computation, which remains underexplored by most SA researchers and is currently being considered by a few sequence analysts. We can expect promising developments in both directions.

4.3.3. Sequence generating models and synthetic life courses

SA is essentially exploratory. According to Ritschard and Studer (2018b, p. 4), "One advantage of SA often discussed is its holistic perspective (see, e.g., Billari, 2005), meaning that SA sheds light on the entire trajectory rather than 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." Probabilistic models (e.g., multistate and Markovian models such as MM, HMM, PST, etc.) are required to model this process. While such models are useful for understanding how state changes depend on history (previously visited states), they are complex and often hardly interpretable. Moreover, their ability to take account of constant and time-varying covariates remains limited. Piccarreta and Studer (2019) stressed that there is a need for criteria for evaluating the ability of MMs to reproduce sequencing. This is also true for timing and duration. Development of generating models specifically designed for reproducing observed sequencing, timing, and spell durations would be particularly welcoming for modelling event occurrences with timing constraints and constraints based on timing and occurrence of previous events.

Panel data generally cover only a portion (a few years) of the life course and building synthetic complete trajectories from such data could help draw conclusions on complete life trajectories. In his treatment of synthetic life histories, Willekens (2014, pp. 46–50) suggested building entire life trajectories by simulating successive portions of the trajectories by means of (microsimulation) models fitted on age groups. Full life courses would be generated by using the model of the youngest age group for generating the first years, and successively the models of the next age groups for later ages. Here, we need a detailed investigation of the applicability of such an approach, especially the underlying assumption regarding the validity of the model for the younger age group for the periods when

older cohorts also lived those ages. Automatizing the approach in a user-friendly tool would also contribute to enlarging the scope of SA.

4.3.4. State and time granularity

Granularity of time and of state definition must be considered. One such an issue is, when analyzing occupational trajectories, how results change when we use the second-level ISCO code (International standard classification of occupations) instead of the first-level code. Likewise, how much do results differ when using monthly rather than yearly data is also an open question. Both distances between sequences and sequence indicators are sensitive to the chosen granularity. Another issue of granularity is about optimally reducing the size of the alphabet. There is a need for methods and guidelines for helping the user to make the best choices.

5. Conclusion

We began the substantive portion of this paper with a discussion of Abbott's substantive interest in histories and his role in the development of SA as an alternative to the so-called "General Linear Reality" represented by regression analysis dominant at the time. Some readers may have seen that, in many SA applications, regression analysis is actually employed, a feature we duly recognized in the paper. Such applications of SA and regression seem to contradict Abbott's original intention. Or do they? It is our view that combining SA with the more typical regression analysis, be it linear or nonlinear, will certainly enrich a research project relying on only one of the two methods. SA facilitates an examination and understanding of historical or life course processes in the data; regression type of analysis allows a researcher to gain certain understanding of some of the relevant mechanisms related to such processes. Therefore, combining the two methods permits researchers to take advantage of the strengths of both methodological approaches.

In this paper, we have reviewed the developments of SA from its birth to its young adulthood in Sections 2 and 3. Our review, albeit *comprehensive*, is far from being *complete*. In our view, it is neither possible nor necessary to give a *complete* review of all the developments in a scholarly field. Our objective has been to provide the reader a *comprehensive* review of the major and useful SA methods as well as the life course that SA has travelled to this day. Therefore, readers new to SA will be able to see its major available methods and tools and are referred to the many additional references where the methods are introduced and explained in greater detail. Social science researchers who already have some SA knowledge can also benefit from the paper with additionally understanding the connections between existing SA methods and the advantages and limitations of the many SA methods currently available.

It is also our sincere hope that, through Section 4, sequence analysts will be able to help us further develop SA in major directions—its further engagement with and enrichment of life course and other social scientific theories, its refinement of the current methodological toolbox, and even new directions that we have not identified in this article. SA's growth from infancy to young adulthood involved the transition from the scholarly attention and efforts of primarily just a sole scholar to the collective efforts of a contingent of devoted SA researchers, as evidenced by the authorship of the current article. Therefore, we are optimistic that future cohorts of scholars will be able to nurture SA into a strong adult in the not-too-distant future.

Acknowledgments

The authors below gracefully acknowledge the support they have received for working on the paper: T. F. Liao for the research support from the College of Liberal Arts & Sciences, the University of Illinois at Urbana-Champaign; D. Bolano for the financial support from the European Union's Horizon 2020 research and innovation program, project "DisCont—Discontinuities in Household and Family Formation" (grant No. 694262; PI: F.C. Billari); and M. Studer for the grant support of the Swiss National Science Foundation (project "Strengthening Sequence Analysis", grant No.: 10001A_204740).

References

Aassve, A., Billari, F.C., Piccarreta, R., 2007. Strings of adulthood: a sequence analysis of young British women's work-family trajectories. Eur. J. Popul. 23, 369–388.
Abbott, A., 1983. Sequences of social events. Hist. Methods 16, 129–147.
Abbott, A., 1988. Transcending general linear reality. Sociol. Theor. 6, 169–186.
Abbott, A., 1990a. Conceptions of time and events in social science methods. Soc. Sci. Hist. 23, 140–150.
Abbott, A., 1990b. A Primer on Sequence Methods. Organ. Sci. 1 (4), 373–392. https://doi.org/10.1287/orsc.1.4.375.
Abbott, A., 1992. From causes to events: notes on narrative positivism. Sociol. Methods Res. 20 (4), 428–455. https://doi.org/10.1177/0049124192020004002.
Abbott, A., 1995. Sequence analysis: new methods for old ideas. Annu. Rev. Sociol. 21, 93–113.
Abbott, A., 2001. Time Matters. Chicago: University of Chicago Press, Chicago.
Abbott, A., 2016. Processual Sociology. University of Chicago Press, Chicago.
Abbott, A., 2022. An Interview with Andrew Abbott. January 16, 2022, conducted by the corresponding author.
Abbott, A., Barman, E., 1997. Sequence comparison via alignment and Gibbs sampling. Sociol. Methodol. 27, 47–87.
Abbott, A., DeViney, S., 1992. The welfare state as transnational event. Soc. Sci. Hist. 16, 245–274.
Abbott, A., Forrest, J., 1986. Optimal matching methods for historical data. J. Interdiscip. Hist. 16, 473–496.
Abbott, A., Hrycak, A., 1990. Measuring resemblance in social sequences: an optimal matching analysis of musicians' careers. Am. J. Sociol. 96, 144–185.
Abbott, A., Tsay, A., 2000. Sequence analysis and optimal matching methods in sociology, review and prospect. Sociol. Methods Res. 29 (1), 3–33. https://doi.org/10.1177/0049124100029001001. 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. Sociol. Methods Res. 38 (3), 430–462.

Aisenbrey, S., Fasang, A.E., 2017. The interplay of work and family trajectories over the life course: Germany and the United States in comparison. Am. J. Sociol. 122 (5), 1448–1484.
Barban, N., 2013. Family trajectories and health: a life course perspective. Eur. J. Popul. 29, 357–385. https://doi.org/10.1007/s10680-013-9296-3.
Barban, N., Billari, F.C., 2012. Classifying life course trajectories: a comparison of latent class and sequence analysis. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 61 (5), 765–784.
Barban, N., de Luna, X., Lundholm, E., Svensson, I., Billari, F.C., 2017 November. Causal effects of the timing of life-course events. Sociol. Methods Res., 004912411772969 https://doi.org/10.1177/0049124117729697.
Bearman, P.S., Stovel, K., 2000. Becoming a Nazi: a model for narrative networks. Poetics 27, 69–90.
Begleiter, R., El-Yaniv, R., Yona, G., 2011. On prediction using variable order Markov models. J. Artif. Intell. Res. 22, 385–421.
Berchtold, A., 1999. The double chain Markov model. Commun. Stat. Theor. Methods 28 (11), 2569–2589.
Bernardi, L., Huinink, J., Settersten, R.A., 2018. The life course cube: a tool for studying lives. Adv. Life Course Res. 1–13 https://doi.org/10.1016/j.alcr.2018.11.004.
Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.
Biemann, T., 2011. A transition-oriented approach to optimal matching. Sociol. Methodol. 41 (1), 195–221. https://doi.org/10.1111/j.1467-9531.2011.01235.x.
Billari, F.C., 2001. Sequence analysis in demographic research. Can. Stud. Popul. 28 (2), 439–458. https://doi.org/10.25336/P6G30C.
Billari, F.C., 2005. Life course analysis: two (complementary) cultures? Some reflections with examples from the analysis of transition to adulthood. In: Levy, R., Ghisletta, P., Le Goff, J.-M., Spini, D., Widmer, E. (Eds.), Towards an Interdisciplinary Perspective on the Life Course, Advances in Life Course Research, vol. 10. Elsevier, Amsterdam, pp. 261–281.
Bison, I., 2014. Sequence as network: an attempt to apply network analysis to sequence analysis. In: Blanchard, P., Bhülmann, F., Gauthier, J.-A. (Eds.), Advances in Sequence Analysis: Theory, Method, Applications. Springer, New York, pp. 231–248.
Blair-Loy, M., 1999. Career patterns of executive women in finance: an optimal matching analysis. Am. J. Sociol. 104 (5), 1346–1397.
Blanchard, P., Bühlmann, F., Gauthier, J.-A. (Eds.), 2014. Advances in Sequence Analysis: Theory, Method, Applications. Springer, New York.
Blanchard, P., 2019. Sequence analysis. In: Atkinson, P.A., Williams, R.A., Cernat, A. (Eds.), Encyclopedia of Research Methods. Sage.
Bolano, D., Berchtold, A., 2016. General framework and model building in the class of Hidden Mixture Transition Distribution models. Comput. Stat. Data Anal. 93, 131–145.
Bolano, D., Berchtold, A., Bürge, E., 2019. The heterogeneity of disability trajectories in later life: dynamics of activities of daily living performance among nursing home residents. J. Aging Health 31 (7), 1315–1336.
Bolano, D., Berchtold, A., 2021. The analysis of inequality in life trajectories: an integration of two approaches. In: Nico, M., Pollock, G. (Eds.), The Routledge Handbook of Contemporary Inequalities and the Life Course. Routledge, pp. 63–80.
Bolano, D., Studer, M., 2020. The link between previous life trajectories and a later life outcome: a feature selection approach. LIVES Work. Pap. (82) https://doi.org/10.12682/lives.2296-1658.2020.82.
Borghetto, E., 2014. Legislative processes as sequences: exploring temporal trajectories of Italian law-making by means of sequence analysis. Int. Rev. Adm. Sci. 80 (3), 553–576. https://doi.org/10.1177/0020852313517996.
Borgna, C., Struffolino, E., 2018. Unpacking configurational dynamics: sequence analysis and qualitative comparative analysis as a mixed-method design. In: Ritschard, G., Studer, M. (Eds.), Sequence Analysis and Related Approaches. Life Course Research and Social Policies, vol. 10. Springer, Cham, pp. 167–184 doi: 10.1007/978-3-319-95420-2_10.
Brinberg, M., Ram, N., Hülür, G., Brick, T.R., Gerstorf, D., 2018. Analyzing dyadic data using grid-sequence analysis: Interdyad differences in Intradyadic dynamics. J. of Gerontology Series B 73, 5–18. https://doi.org/10.1093/geronb/gbw160.
Brückner, H., Mayer, K.U., 2005. De-standardization of the life course: what it might mean? And if it means anything, whether it actually took place? Adv. Life Course Res. 9 (4), 27–53. https://doi.org/10.1016/S1040-2608(04)09002-1.
Brum-Bastos, V.S., Long, J.A., Demšar, U., 2018. Weather effects on human mobility: a study using multi-channel sequence analysis. Comput. Environ. Urban Syst. 71, 131–152.
Brzinsky-Fay, C., 2007. Lost in transition? Labour market entry sequences of school leavers in Europe. Eur. Sociol. Rev. 23 (4), 409–422. https://doi.org/10.1093/esr/jcm011.
Brzinsky-Fay, C., 2014. Graphical representation of transitions and sequences. In: Blanchard, P., Bühlmann, F., Gauthier, J.-A. (Eds.), Advances in Sequence Analysis: Theory, Method, Applications. Cham/Heidelberg/New York/Dordrecht/London: Springer, pp. 265–284.
Brzinsky-Fay, C., 2018. Unused resources: sequence and trajectory indicators. In: International Symposium on Sequence Analysis and Related Methods, Monte Verita, TI. Switzerland, October 10–12, 2018.
Brzinsky-Fay, C., Ebner, C., Seibert, H., 2016. Veränderte Kontinuität. Berufseinstiegsverläufe von Ausbildungsabsolventen in Westdeutschland seit den 1980er Jahren. Kölner Z. Soziol. Sozialpsychol. 68 (2), 229–258.
Brzinsky-Fay, C., Kohler, U., Luniak, M., 2006. Sequence analysis with Stata. STATA J. 6 (4), 435–460. https://doi.org/10.1177/1536867X0600600401.
Bürgin, R., Ritschard, G., 2014. A decorated parallel coordinate plot for categorical longitudinal data. Am. Statistician 68 (2), 98–103. https://doi.org/10.1080/00031305.2014.887591.
Buton, F., Lemercier, C., Mariot, N., 2012. The household effect on electoral participation. A contextual analysis of voter signatures from a French polling station (1982–2007). Elect. Stud. 31 (2), 434–447. https://doi.org/10.1016/j.electstud.2011.11.010.
Butts, C.T., Marcum, C.S., 2017. A relational event approach to modeling behavioral dynamics. Pp. 51–92 in group processes: data-driven computational approaches. In: Andrew Pilny and Marshall Scott Poole. Springer.
Buyukkececi, Z., Fasang, A.E., Kraus, V., Levanon, A., Saburov, E., 2022. Cumulative (Dis)advantages in Work and Family Life Courses Among Jewish and Palestinian Women in Israel. Manuscript submitted for publication.
Casper, G., Wilson, M., 2015. Using sequences to model crises. Polit. Sci. Res. Methods 3 (2), 381–397. https://doi.org/10.1017/psrm.2014.27.
Castelo, A.F.M., Schäfer, M., Silva, M.E., 2021. Food practices as part of daily routines: a conceptual framework for analysing networks of practices. Appetite 157, 104978.
Chan, T.-W., 1995. Optimal matching analysis: a methodological note on studying career mobility. Work Occup. 22, 467–490.
Christensen, R.C., 2021. Elite professionals in transnational tax governance. Glob. Networks A J. Transnat. Aff. 21 (2), 265–293. https://doi.org/10.1111/glob.12269.
Claessen, C., Bailer, S., Turner-Zwinkels, T., 2021. The winners of legislative mandate: an analysis of post-parliamentary career positions in Germany and The Netherlands. Eur. J. Polit. Res. 60 (1), 25–45. https://doi.org/10.1111/1475-6765.12385. ISSN 0304-4130.
Cornwell, B., 2015. Social Sequence Analysis. Cambridge University Press, New York.
Cornwell, B., 2018. Network analysis of sequence structures. In: Ritschard, G., Studer, M. (Eds.), Sequence Analysis and Related Approaches. Springer, Cham, pp. 103–120. https://doi.org/10.1007/978-3-319-95420-2_7.
Cornwell, B., Watkins, K., 2015. Sequence-network analysis: a new framework for studying action in groups. Adv. Group Process. 32, 31–63.
Courgeau, D., 2018. Do different approaches in social science lead to divergent or convergent models? In: Ritschard, G., Studer, M. (Eds.), Sequence Analysis and Related Approaches. Springer, Cham, pp. 15–33. https://doi.org/10.1007/978-3-319-95420-2_2.
Deville, J.-C., Saporta, G., 1983. Correspondence analysis, with an extension towards nominal time series. J. Econom. 22 (1–2), 169–189. https://doi.org/10.1016/0304-4076(83)90098-2.
Dijkstra, W., 1994. Sequence: a program for analyzing sequential data. Bull. Méthodol. Sociol. 43, 134–142.
Dijkstra, W., Taris, T., 1995. Measuring the agreement between sequences. Sociol. Methods Res. 24, 214–231. https://doi.org/10.1177/0049124195024002004.
Dixon, M., 2008. Movements, counter movements and policy adoption: the case of right-to-work activism. Soc. Forces 87 (1), 473–500.
Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. 3 (3), 32–57.
Durrant, G.B., Maslovskaya, O., Smith, P.W.F., 2019. Investigating call record data using sequence analysis to inform adaptive survey designs. Int. J. Soc. Res. Methodol. 22 (1), 37–54.

Elder, G.H., Johnson, M.K., Crosnoe, R., 2003. The emergence and development of life course theory. In: Mortimer, J.T., Shanahan, M.J. (Eds.), Handbook of the Life Course. Springer US, Boston, MA, pp. 3–19. https://doi.org/10.1007/978-0-306-48247-2_1.
Elzinga, C.H., Liefbroer, A.C., 2007. De-standardization of family-life trajectories of young adults: a cross-national comparison using sequence analysis. Eur. J. Popul. 23, 225–250. https://doi.org/10.1007/s10680-007-9133-7.
Elzinga, C.H., Studer, M., 2015. Spell sequences, state proximities and distance metrics. Sociol. Methods Res. 44 (1), 3–47. https://doi.org/10.1177/0049124114540707.
Erickson, B.W., Watterson, D.M., Marshak, D.R., 1980. Sequence alignment of calmodulin domains by metric analysis. Ann. N. Y. Acad. Sci. 356 (1), 378–379.
Fasang, A., 2012. Retirement patterns and income inequality. Soc. Forces 90 (3), 685–711. https://doi.org/10.1093/sf/sor015.
Fasang, A., Aisenbrey, S., 2021. Uncovering Social Stratification: Intersectional Inequalities in Work and Family Life Courses by Gender and Race, Social Forces. doi: 10.1093/sf/soab151.
Fasang, A.E., Liao, T.F., 2014. Visualizing sequences in the social sciences: relative frequency sequence plots. Sociol. Methods Res. 43 (4), 643–676.
Fasang, A.E., Mayer, K.U., 2020. Life course and social inequality. In: Falkingham, J., Evandrou, M., Vlachantoni, A. (Eds.), Handbook of Demographic Change and the Life Course. Edward Elgar, pp. 22–39.
Fasang, A.E., Raab, M., 2014. Beyond transmission: intergenerational patterns of family formation among middle-class American families. Demography 51, 1703–1728.
Fauser, S., 2020. Career trajectories and cumulative wages: the case of temporary employment. Res. Soc. Stratif. Mobil. 69, 100529.
Forrest, J., Abbott, A., 1990. The optimal matching method for anthropological data. J. Quant. Anthropol. 2, 151–170.
Furstenberg, F.F., 2005. Non-normative life course transitions: reflections on the significance of demographic events on lives. Adv. Life Course Res. 10, 155–172.
Gabadinho, A., Ritschard, G., 2013. Searching for typical life trajectories, applied to childbirth histories. In: Lévy, R., Widmer, E.D. (Eds.), Gendered Life Courses - between Individualization and Standardization. A European Approach Applied to Switzerland. LIT, Vienna, pp. 287–312.
Gabadinho, A., Ritschard, G., 2016. Analysing state sequences with probabilistic suffix trees: the PST R package. J. Stat. Software 72 (3), 1–39. https://doi.org/10.18637/jss.v072.i03.
Gabadinho, A., Ritschard, G., Muller, N.S., Studer, M., 2011. Analyzing and visualizing state sequences in R with TraMineR. J. Stat. Software 40 (4), 1–37. https://doi.org/10.18637/jss.v040.i04.
Gabadinho, A., Ritschard, G., Studer, M., Muller, 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,, pp. 61–66.
Gauthier, J.-A., Widmer, E.D., Bucher, P., Notredame, C., 2009. How much does it cost?: optimization of costs in sequence analysis of social science data. Sociol. Methods Res. 38, 197–231.
Gauthier, J.-A., Widmer, E.D., Bucher, P., Notredame, C., 2010. Multichannel sequence analysis applied to social science data. Sociol. Methodol. 40 (1), 1–38.
Gjerløw, H., Knutsen, C.H., Wig, T., Wilson, M.C., 2021. One Road to Riches? How State Building and Democratization Affect Economic Development. Cambridge University Press.
Giddens, Anthony, 1986. The Constitution of Society: Outline of the Theory of Structuration, 1st pbk. University of California Press, Berkeley.
Goh, K.T., Pentland, B.T., 2019. From actions to paths to patterning: toward a dynamic theory of patterning in routines. Acad. Manag. J. 62 (6), 1901–1929.
Gruijters, R., Van Winkle, Z., Fasang, A.E., 2022. Life Course Trajectories and Wealth Accumulation in the United States: Comparing Late Baby Boomers and Early Millennials. Manuscript submitted for publication.
Halpin, B., 2010. Optimal matching analysis and life-course data: the importance of duration. Sociol. Methods Res. 38 (3), 365–388.
Halpin, B., 2014. Three narratives of sequence analysis. In: Blanchard, Philippe, Bühlmann, F., Gauthier, Jacques-Antoine (Eds.), Advances in Sequence Analysis: Theory, Method, Applications, vol. 2. Life Course Research and Social Policies, Heidelberg: Springer, pp. 75–103. https://doi.org/10.1007/978-3-319-04969-4_5.
Halpin, B., 2016. Multiple imputation for categorical time series. STATA J. 16 (3), 590–612. https://doi.org/10.1177/1536867X1601600303.
Halpin, B., 2017. SADI: sequence analysis tools for Stata. STATA J. 17 (3), 546–572. https://doi.org/10.1177/1536867X1701700302.
Halpin, B., Chan, T.-W., 1998. Class careers as sequences: an optimal matching analysis of work-life histories. Eur. Sociol. Rev. 14, 111–130.
Hamberger, K., 2018. Relational sequence networks as a tool for studying gendered mobility patterns. In: Ritschard, G., Studer, M. (Eds.), Sequence Analysis and Related Approaches. Springer, Cham, pp. 121–146. https://doi.org/10.1007/978-3-319-95420-2_8.
Han, Y., Liefbroer, A.C., Elzinga, C.H., 2017. Comparing methods of classifying life courses: sequence analysis and latent class analysis. Longitudinal Life Course Stud. 8 (4), 319–341.
Han, Y., Liefbroer, A.C., Elzinga, C.H., 2020. Mechanisms of family formation: an application of Hidden Markov Models to a life course process. Adv. Life Course Res. 43, 100265.
Han, S.-K., Moen, P., 1999. Clocking out: temporal patterning of retirement. Am. J. Sociol. 105, 191–236.
Heckhausen, J., Buchmann, M., 2019. A multi-disciplinary model of life-course canalization and agency. Adv. Life Course Res. 41, 100246.
Helske, S., Helske, J., Eerola, M., 2018. Combining sequence analysis and hidden Markov models in the analysis of complex life sequence data. In: Ritschard, Studer (Eds.), Sequence Analysis and Related Approaches. Springer, Cham, pp. 185–200 doi:10.1007/978-3-319-95420-2_11.
Helske, S., Helske, J., 2019. Mixture hidden Markov models for sequence data: the seqHMM package in R. J. Stat. Software 88 (3), 1–32. https://doi.org/10.18637/jss.v088.i03.
Helske, S., Keski-Säntti, M., Kivelä, J., Juutinen, A., Kääriälä, A., Gissler, M., et al., 2021. Predicting the Stability of Early Employment with its Timing and Childhood Social and Health-Related Predictors: a Mixture Markov Model Approach. SocArXiv https://osf.io/preprints/socarxiv/qkcxs/.
Hennig, C., 2007. Cluster-wise assessment of cluster stability. Comput. Stat. Data Anal. 52 (1), 258–271.
Hennig, C., Lin, C.-J., 2015. Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. Stat. Comput. 25 (4), 821–833.
Higginson, S., McKenna, E., Hargreaves, T., Chilvers, J., Thomson, M., 2015. Diagramming social practice theory: an interdisciplinary experiment exploring practices as networks. Indoor Built Environ. 24 (7), 950–969.
Hollister, M., 2009. Is optimal matching suboptimal? Sociol. Methods Res. 38 (2), 235–264. https://doi.org/10.1177/0049124109346164.
Huinink, J., 2005. Räumliche Mobilität und Familienentwicklung. Ein lebenslauftheore tischer Systematisierungsversuch. In: Steinbach, A. (Ed.), Generatives Verhalten und Generationenbeziehungen. VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 61–84.
Jalovaara, M., Fasang, A.E., 2020. Family life courses, gender, and mid-life earnings. Eur. Sociol. Rev. 36 (2), 159–178.
Kalucza, S., Lam, J., Baxter, J., 2021. Transformation, disruption or cumulative disadvantage? Labor market and education trajectories of young mothers in Australia. Adv. Life Course Res. 100446.
Karhula, A., Erola, J., Raab, M., Fasang, A.E., 2019. Destination as a process: sibling similarity in early socioeconomic trajectories. Adv. Life Course Res. 40, 85–98.
Kaufman, L., Rousseeuw, P.J., 2005. Finding Groups in Data. John Wiley & Sons, Hoboken.
Kass, R.E., Raftery, A.E., 2005. Bayes factors. J. Am. Stat. Assoc. 90 (430), 773–795.
Kim, D., Graham, T., Wan, Z., Rizoiu, M.-A., 2019. Analysing user identity via time-sensitive semantic edit distance (t-SED): a case study of Russian trolls on Twitter. J. Comput. Soc. Sci. 2 (2), 331–351.
Kreuter, F., Kohler, U., 2009. Analyzing contact sequences in call record data: potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. J. Off. Stat. 25 (2), 203–226.
Lazar, A., Jin, L., Spurlock, C.A., Wu, K., Sim, A., Todd, A., 2019. Evaluating the effects of missing values and mixed data types on social sequence clustering using T-SNE visualization. J. Data Inf. Qual. (JDIQ) 11 (2), 1–22.
Le Goix, R., Giraud, T., Cura, R., Le Corre, T., Migozzi, J., 2019. Who sells to whom in the suburbs? Home price inflation and the dynamics of Sellers and buyers in the metropolitan region of Paris, 1996–2012. PLoS One 14 (3), e0213169.
Lesnard, L., 2008. Off-scheduling within dual-earner couples: an unequal and negative externality for family time. Am. J. Sociol. 114, 447–490.

Lesnard, L., 2010. Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. Sociol. Methods Res. 38, 389–419. https://doi.org/10.1177/0049124110362526.
Levine, J.H., 2000. But what have you done for us lately? Sociol. Methods Res. 29 (1), 35–40.
Liao, T.F., 2021. Using sequence analysis to quantify how strongly life courses are linked. Sociol. Sci. 8, 48–72.
Liao, T.F., Fasang, A.E., 2021. Comparing groups of life-course sequences using the Bayesian information criterion and the likelihood-ratio test. Sociol. Methodol. 51 (1), 44–85. https://doi.org/10.1177/0081175020959401.
Liefbroer, A.C., Elzinga, C.H., 2012. Intergenerational transmission of behavioural patterns: how similar are parents' and children's demographic trajectories? Adv. Life Course Res. 17, 1–10.
Lőrincz, M.J., Ramírez-Mendiola, J.L., Torriti, J., 2021. Impact of time-use behaviour on residential energy consumption in the United Kingdom. Energies 14 (19), 6286.
Madero-Cabib, I., Le Feuvre, N., König, S., 2021. Gendered retirement pathways across life course regimes. Ageing Soc. 1–30.
Mahringer, C.A., 2022. Analyzing digital trace data to promote discovery – the case of heatmapping. In: Marrella, A., Weber, B. (Eds.), Business Process Management Workshops. BPM 2021. Lecture Notes in Business Information Processing. Springer, Cham, pp. 209–220.
Mahringer, C.A., Pentland, B.T., 2021. Sequence analysis in routine dynamics. In: Feldman, M.S., Pentland, B.T., D'Adderio, L., et al. (Eds.), Cambridge Handbook of Routine Dynamics. Cambridge University Press, pp. 172–183.
Mayer, K.U., 2009. New directions in life course research. Annu. Rev. Sociol. 35, 413–433. https://doi.org/10.1146/annurev.soc.34.040507.134619.
Mas, J., Nogueira de Vasconcelos, R., Franca-Rocha, W., 2019. Analysis of high temporal resolution land use/land cover trajectories. Land 8 (2), 30. https://doi.org/10.3390/land8020030.
Mattioli, G., Anable, J., Vrotsou, K., 2016. Car dependent practices: findings from a sequence pattern mining study of UK time use data. Transport. Res. Part A Pol. Pract. 89 (July), 56–72. https://doi.org/10.1016/j.tra.2016.04.010.
McKenna, E., Higginson, S., Hargreaves, T., Chilvers, J., Thomson, M., 2020. When activities connect: sequencing, network analysis, and energy demand modelling in the United Kingdom. Energy Res. Social Sci. 69, 101572.
Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., Andersen, P.K., 2009. Multi-state models for the analysis of time-to-event data. Stat. Methods Med. Res. 18 (2), 195–222.
Merton, R.K., 1988. The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. Isis 79 (4), 606–623. https://doi.org/10.1086/354848.
Möhring, K., Weiland, A.P., 2022. Couples' life courses and women's income in later life: a multichannel sequence analysis of linked lives in Germany. Eur. Sociol. Rev. 38 (3), 371–388. https://doi.org/10.1093/esr/jcab048.
Muller, J.S., Hiekel, N., Liefbroer, A.C., 2020. The long-term costs of family trajectories: women's later-life employment and earnings across Europe. Demography 57 (3), 1007–1034.
Nee, V., Liu, L., DellaPosta, D., 2017. The entrepreneur's network and firm performance. Sociol. Sci. 4, 552–579.
Needleman, S., Wunsch, C., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453. https://doi.org/10.1016/0022-2836(70)90057-4.
Nutz, T., Gritti, D., 2022. Dyadic employment biographies and within-couple wealth inequality in Britain and Western Germany. J. Marriage Fam. 84 (2), 552–569. https://doi.org/10.1111/jomf.12811.
O'Flaherty, M., Baxter, Haynes, M., Turrell, G., 2016. The family life course and health: partnership, fertility histories, and later-life physical health trajectories in Australia. Demography 53 (3), 777–804.
Parsons, Talcott, 1951. The Social System. Routledge, London.
Pentland, B.T., Pentland, A.P., Calantone, R.J., 2017. Bracketing off the actors: towards an action-centric research agenda. Inf. Organ. 27 (3), 137–143.
Pesando, L.M., Barban, N., Sironi, M., Furstenberg, F.F., 2022. A sequence-analysis approach to the study of the transition to adulthood in low- and middle-income countries. Popul. Dev. Rev. 47 (3), 719–747. https://doi.org/10.1111/padr.12425.
Piccarreta, R., 2012. Graphical and smoothing techniques for sequence analysis. Sociol. Methods Res. 41 (2), 362–380.
Piccarreta, R., 2017. Joint sequence analysis: association and clustering. Sociol. Methods Res. 46 (2), 252–287.
Piccarreta, R., Billari, F.C., 2007. Clustering work and family trajectories using a divisive algorithm. J. Roy. Stat. Soc. Ser. A 170, 1061–1078.
Piccarreta, R., Bonetti, M., 2019. Assessing and comparing models for sequence data by microsimulation. SocArXiv. https://doi.org/10.31235/osf.io/3mcfp.
Piccarreta, R., Elzinga, C.H., 2013. Mining for association between life courses. In: McArdle, J., Ritschard, G. (Eds.), Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences. Routledge, New York.
Piccarreta, R., Studer, M., 2019. Holistic analysis of the life course: methodological challenges and new perspectives. Adv. Life Course Res. 41, 1–11. https://doi.org/10.1016/j.alcr.2018.10.004.
Pollock, G., 2007. Holistic trajectories: a study of combined employment, housing and family careers by using multiple-Sequence Analysis. J. Roy. Stat. Soc. Ser. A 170, 167–183.
Raab, M., Fasang, A.E., Karhula, A., Erola, J., 2014. Sibling similarity in family formation. Demography 51, 2127–2154.
Raab, M., 2022. ggseqplot: Render Sequence Plots using 'ggplot2'. github. https://CRAN.R-project.org/package=ggseqplot.
Raab, M., Struffolino, E., 2022. Sequence Analysis. SAGE Publications.
Ritschard, G., 2021. Measuring the nature of individual sequences. Sociol. Methods Res. https://doi.org/10.1177/00491241211036156.
Ritschard, G., Bussi, M., O'Reilly, J., 2018. An index of precarity for measuring early employment insecurity. In: Ritschard, G., Studer, M. (Eds.), Sequence Analysis and Related Approaches. Springer, Cham, pp. 279–295. https://doi.org/10.1007/978-3-319-95420-2_16.
Ritschard, G., Studer, M. (Eds.), 2018a. Sequence Analysis and Related Approaches: Innovative Methods and Applications, Volume 10 of Life Course Research and Social Policies. Springer, Cham. https://doi.org/10.1007/978-3-319-95420-2.
Ritschard, G., Studer, M., 2018b. Sequence analysis: where are we, where are we going? In: Ritschard, G., Studer, M. (Eds.), Sequence Analysis and Related Approaches. Springer, Cham, pp. 1–11. https://doi.org/10.1007/978-3-319-95420-2_1.
Ritschard, G., Studer, M., Bürgin, R., Liao, T., Gabadinho, A., Müller, N.S., Rousset, P., 2022. Package TraMineRextras: TraMineR Extension. Reference Manual. CRAN. Version 0.6.3. https://CRAN.R-project.org/package=TraMineRextras.
Robette, N., Thibault, N., 2008. Analyse harmonique qualitative ou méthode d'appariement optimal? Une analyse exploratoire de Trajectoires professionnelles. Population 63 (4), 621–646. https://doi.org/10.3917/popu.804.0621. http://www.jstor.org/discover/10.2307/27736782?uid=3739256&uid=2129&uid=2&uid=70&uid=4&sid=21102253276241.
Rohwer, G., Pötter, U., 2002. TDA User's Manual. Software. Ruhr-Universität Bochum, Fakultät für Sozialwissenschaften, Bochum. http://www.stat.ruhr-uni-bochum.de/tda.html.
Rossignon, F., Studer, M., Gauthier, J.-A., Le Goff, J.M., 2018. Sequence History Analysis (SHA): estimating the effect of the past trajectories on an upcoming event. In: Ritschard, G., Studer, M. (Eds.), Sequence Analysis and Related Approaches. Springer, Cham, pp. 83–101. https://doi.org/10.1007/978-3-319-95420-2_6.
Rousset, P., Giret, J.-F., Grelet, Y., 2012. Typologies de Parcours et dynamique longitudinale. Bull. Sociol. Methodol./Bull. Méthodol. Sociol. 114 (1), 5–34. https://doi.org/10.1177/0759106312437142.
Sankoff, D., Kruskal, J., 1983. Time Warps, String Edits, and Macromolecules. University of Chicago Press, Chicago.
Schecter, A., Pilny, A., Leung, A., Poole, M.S., Contractor, N., 2018. Step by step: capturing the dynamics of work team process through relational event sequences. J. Organ. Behav. 39 (9), 1163–1181.
Scherer, S., 2001. Early career patterns: a comparison of great Britain and west Germany. Eur. Sociol. Rev. 17 (2), 119–144.
Shanahan, M.J., Elder, G.H.J., 2002. History, agency, and the life course. Nebraska Symposium on Motivation. Nebr. Symp. Motiv. Paper 48, 145–186.
Shoval, N., Isaacson, M., 2007. Sequence alignment as a method for human activity analysis in space and time. Ann. Assoc. Am. Geogr. 97 (2), 282–297. https://doi.org/10.1111/j.1467-8306.2007.00536.x.

Simmel, G., 1955. Conflict & the Web of Group-Affiliations. Translated by Kurt H. Wolff and Reinhard Bendix. Free Press, New York.
Stovel, K., Bolan, M., 2004. Residential trajectories using optimal alignment to reveal the structure of residential mobility. Sociol. Methods Res. 32 (4), 559–598.
Stovel, K., Savage, M., Bearman, P., 1996. Ascription into achievement: models of career systems at Lloyds Bank 1890–1970. Am. J. Sociol. 102, 358–399.
Studer, M., 2013. WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R. https://doi.org/10.12682/lives.2296-1658.2013.24. LIVES Working Papers 24, NCCR LIVES, Switzerland.
Studer, M., 2018. Divisive property-based and fuzzy clustering for sequence analysis. In: Ritschard, G., Studer, M. (Eds.), Sequence Analysis and Related Approaches: Innovative Methods and Applications. Springer, Cham, pp. 223–239. https://doi.org/10.1007/978-3-319-95420-2_13.
Studer, M., 2021. Validating sequence analysis typologies using parametric bootstrap. Sociol. Methodol. 51 (2), 290–318. https://doi.org/10.1177/00811750211014232.
Studer, M., Ritschard, G., 2016. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. J. Roy. Stat. Soc. Ser. A 179 (2), 481–511. https://doi.org/10.1111/rssa.12125.
Studer, M., Ritschard, G., Gabadinho, A., Müller, N.S., 2011. Discrepancy analysis of state sequences. Sociol. Methods Res. 40 (3), 471–510. https://doi.org/10.1177/0049124111415372.
Studer, M., Liefbroer, A.C., Mooyaart, J.E., 2018a. Understanding trends in family formation trajectories: an application of competing trajectories analysis (CTA). Adv. Life Course Res. 36, 1–12. https://doi.org/10.1016/j.alcr.2018.02.003.
Studer, M., Struffolino, E., Fasang, A.E., 2018b. Estimating the relationship between time-varying covariates and trajectories: the sequence analysis multistate model procedure. Sociol. Methodol. 48 (1), 103–135.
Taushanov, Z., Berchtold, A., 2017. A direct local search method and its application to a Markovian model. Stat. Optim. Inf. Comput. 5 (1), 19–34.
Van Winkle, Z., 2018. Family trajectories across time and space: increasing complexity in family life courses in Europe? Demography 55 (1), 135–164.
Van Winkle, Z., 2020. Family policies and family life course complexity across 20th-century Europe. J. Eur. Soc. Pol. 30 (3), 320–338. https://doi.org/10.1177/0958928719880508.
Van Winkle, Z., Fasang, A.E., 2021. The complexity of employment and family life courses across 20 th century Europe. Demogr. Res. 44, 775–810.
Vermunt, J.K., Tran, B., Magidson, J., 2008. Latent class models in longitudinal research. In: Menard, S. (Ed.), Handbook of Longitudinal Research: Design, Measurement, and Analysis. Elsevier, Burlington, MA, pp. 373–385.
Visser, M., Fasang, A.E., 2018. Educational assortative mating and couples' linked late-life employment trajectories. Adv. Life Course Res. 37, 79–90.
Wahrendorf, M., Marr, A., Antoni, M., Pesch, B., Jöckel, K., Lunau, z., Moebus, S., et al., 2019. Agreement of self-reported and administrative data on employment histories in a German cohort study: a sequence analysis. Eur. J. Popul. 35 (2), 329–346. https://doi.org/10.1007/s10680-018-9476-2.
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? 1. Am. J. Sociol. 120 (6), 1809–1856.
Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, New York.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis, second ed. Springer. https://doi.org/10.1007/978-3-319-24277-4. 2016.
Willekens, F., 2014. Multistate Analysis of Life Histories with R. Use R!. Springer, Heidelberg.
Wilson, M.C., 2014. Governance built step-by-step: analysing sequences to explain democratization. In: Advances in Sequence Analysis: Theory, Method, Applications, Edited by Philippe Blanchard, Felix Bühlmann, and Jacques-Antoine Gauthier. Springer, New York, pp. 213–227.
Wu, L., 2000. Some Comments on Sequence analysis and optimal matching methods in sociology: review and prospect. Sociol. Methods Res. 29, 41–64.
Zhang, W., Thill, J.-C., 2017. Detecting and visualizing cohesive activity-travel patterns: a network analysis approach. Comput. Environ. Urban Syst. 66, 117–129.
Zucchini, W., MacDonald, I.L., 2009. Hidden Markov Models for Time Series: an Introduction Using R. Chapman and Hall/CRC.