SAGE researchmethods

Sequence Analysis

Foundation Entries

SAGE Research Methods Foundations

By: Philippe Blanchard
Published:2019
Length: 10,000 Words
DOI: http://dx.doi.org/10.4135/9781526421036
Methods: Sequence Analysis
Online ISBN: 9781526421036
Disciplines: Business and Management, Geography, Health, Political Science and International Relations, Sociology
Access Date: February 7, 2020
Publishing Company: SAGE Publications Ltd
City: London

Abstract

Sequence analysis (SA) is the systematic descriptive and causal study of sequences, that is, successions of standard categorical states or events. SA is a unique method for representing, comparing, and clustering sequences; for extracting prototypical sequences; and for mining sequence populations. Its core tool, the optimal matching algorithm, screens and discriminates longitudinal processes according to the nature of events, their duration, and their order. Numerous fields in the social and political sciences deal with sequences including life-course analysis, the sociology of professional careers, political sociology, the study of political regimes, and various geographical or ethnographic practices. The entry firstly defines sequences, provides examples of them, and presents SA's underlying sociological concepts and objectives. It shows how SA differs from other social scientific methods, especially longitudinal, statistical methods. A brief historical account of the method's development highlights the original intuition, its standardisation, and the dialectic between unity and diversity of the field. The second section relies on a range of concrete applications to present the main usual steps in SA studies: conceptualisation, data collection and preparation, exploration, calculation of dissimilarities, and postdissimilarity treatments. The third section shows the limitations of the standard approaches, and the available variations and variants, including alternative calculations of dissimilarities, multichannel SA, and prototypical sequences.

Introduction

Definitions

Sequence analysis (SA) is the systematic study of samples of sequences. A sequence is a succession of *states* chosen from a finite *alphabet* that covers all possible states in the sample. Sequences may also be regarded as time series, but with categorical data, which makes their modelling more complex than usual, continuous time series. States chain up into *subsequences*, some of which are recurrent and may constitute meaningful patterns. Subsequences made of one or more occurrences of a unique state are *episodes*. States may have a duration, such as a position, or not, such as transition to another position.

Figure 1. Example of sequences—An international comparison of academic careers.

Years (count before professorship)	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
Sequence for individual 1	a	a	a	b	b	b	b	b	d	d	f
Sequence for individual 2		a	b	b	e	e	e	d	d	d	f
Sequence for individual 3			c	c	e	e	e	c	d	d	f

Alphabet	
State a: Contract at pre-PhD level	State d: Teaching and research at PhD level
State b: Teaching contract at PhD level	State e: Abroad
State c: Research contract at PhD level	State f: Professor

Figure 1 is an unpublished extract from the ERC project "Disconex" by J. Angermuller, P. Blanchard, F. Dufour, and A. Zelaznyaiming at understanding the various pathways to professorship in France, Germany, and the United Kingdom. Each individual is represented by one sequence of yearly positions, from the first academic contract to access to professorship. The alphabet records the positions according to the formal qualification requirement of the contract (which does not always correspond to the effective degree held). All three sequences share a common subsequence <ddf> at the end, which includes an identical transition <d→f> from a research and teaching contract to professorship. Sequences 2 and 3 also share a 3-year episode <eee>of postdoctoral activity abroad at the same moment of their career.

Standard examples of individual sequences are work careers made of positions or ranks in organisations, family trajectories made of marital statuses, and residential trajectories listing regions or towns. Otherwise, cases can be collective entities, like countries with evolving policies or legislations, or social movement mobilisations observed through their stages of development. More abstract processes are also relevant to SA, such as sequences of domestic or international conflicts, artistic performances, or symbolic social practices (Abbott, 1995).

The overarching goal of SA is to understand social phenomena as they unfold over time. This goes further than comparing a given sample of individuals, groups, or social artefacts over a few observation points such as repeated sampling through panel questionnaire surveys or database extracts. Sequences require a sufficient number of time points so as to describe precisely enough the richness of social time as it progresses, step by step. Reversely, processes that are too complex, too different between cases, or that evolve too continuously—and therefore cannot be mapped by a finite alphabet—cannot be treated as sequences.

SA addresses, sometimes redefines and enriches, a range of concepts that social theory has always had an interest in but without analysing them with the systematic nature of empirical statistics. These concepts include the following:

- *transitions*—moving from one state, status, or position to the next;
- *stages*—recurring steps that tend to be regarded as normal;
- *turning points*—major transitions, which reorient the whole of a process, whether an individual's life course, an institution's development or a macrogroup's political–historical evolution;

- *time patterns*—subsequences with certain meaning or theoretical bearing that repeat themselves across cases, time, and/or space;
- *generations*—groups of cases that share not only a close starting point along a predefined time axis but also the experience of certain events that unite their perception of being contemporaneous to each other.

To this list can be added a range of concepts that describe some degree of continuity in pathways, such as

- *legacies*—where the present has to manage the past,
- *path dependencies*—where past situations condition present ones,
- *projects*—the present preparing the future, and
- *strategies*—present arrangements to best control a future situation.

In the earlier example of academic careers, stages can be local academic contracts during one's PhD, then postdoctoral contracts abroad, before accessing to professorship; a turning point might often be obtaining a PhD, as this degree grants access to a market of stable academic contracts; younger generations tend to experience longer postdoctoral episodes, across the three countries; and careers are definitely conditioned by individual projects and strategies.

At a more operational level, SA is a set of tools and procedures, implemented through software programmes, with five main objectives (Blanchard, 2011):

1. describing and visualising sequences with measures and plots both in order to facilitate their analysis and to display sequential structures in efficient ways to readers;
2. comparing sequences and classifying them;
3. mining sequences, that is, adding the sequential dimension to already available data mining techniques (e.g., cluster analysis, correspondence analysis);
4. extracting prototypical sequences, so as to summarise concretely clusters of similar sequences and link these clusters with nonstatistical data (typically, in-depth interviews or case studies); and
5. linking sequence profiles with exogenous variables, which brings in causal explanation, both of and by sequences.

Depending on their topic and purpose, SA studies may aim at all or only part of these five objectives. But by opening the sequence toolbox, they access a range of concepts and tools that enable, possibly, to address them all.

Objectives

Sequence analysis developed in the social sciences in response to a major discrepancy between a widespread interest in social dynamics and processes on one side, and a limited ability to fulfil this interest with existing methods and tools, on the other (Abbott, 2001). The methods and tools that were available by the early 1980s to analyse populations of sequences systematically (i.e., distinct from, among other methods,

small-*N* life history interviews, prosopography, and in-depth biographies) can be grouped in three categories: basic, multivariate, and longitudinal. *Basic methods* included summary statistics on individuals and groups, such as frequency of a given state at a given point in time, mean and standard deviation of length spent in one given position over a given period, and elementary sequence graphs, such as by means of "conditional formatting" functions in spreadsheet software. *Multivariate methods* may have appeared better equipped to grasp this complexity: Clustering algorithms or correspondence analysis, key approaches to thin-grained, descriptive data analysis, including with categorical variables, seem to be able to handle long sequences in a holistic way. Synchronic regression models are able to integrate multiple time points as independent variables, with the purpose of explaining a current outcome from factors more or less remote in the past. Specific *longitudinal methods* deal with sequences, from several angles: the distribution of the successive values taken by several continuous variables and their direct and/or lagged interactions over time (time series analysis); the probability for a given event to happen, or not, and within a certain time span, given its past occurrences and the past evolution of a range of continuous independent variables (a set of tools known as event history analysis); and a range of other regression models handling time, such as dynamic panel, hazard-rate, latent trajectory, and diffusion models.

All these approaches present limitations with regard to treating sequentiality. Basic methods propose a fine-grain, controlled approach to sequences, but they meet obvious limits regarding the efforts and time involved, the repetition of tasks, and their ability to address slightly complex data structures such as sequences with large alphabets and complex sequential patterns. Multivariate methods embrace the multifactorial nature of social phenomena; however, they are strictly dependent to a column-by-column data set structure. As they regard columns as interchangeable, they miss the ordered chaining of events, and the fact that positions persist over time, or not, and may have consequences over various time spans. Longitudinal methods sound like the best candidates, yet each of them is limited for one or more of the following reasons: They focus on continuous variables, which are not predominant in the social sciences, except in the fields of economics and psychology; they focus on explaining only a given event (outcome) or set of events (Aisenbrey & Fasang, 2010); or they apply a point-by-point, analytical, short-sightseeing decomposition of sequences, thereby missing the dynamic nature of sequences. Additionally, most of these three kinds of traditional methods struggle to handle data censorship, that is, the fact that the left and/or right tail of sequences is missing due to the nature of the data or the way they were collected. Finally and most importantly, all of them address the order of events only superficially, if at all.

SA intends to overcome these limitations by means of a combination of five characteristics. The first characteristic, probably the central one, is that, thanks to the flexibility of sequence comparison algorithms (see Calculation of Dissimilarities section later in this entry), SA takes jointly into account *three dimensions of social time* that other methods only address separately:

1. the duration of sequences and episodes within sequences, which materialise for example social stability, persistence, survival, or continuity;
2. the timing of events, that is, their position along the time axis, typically age or historical time, and

their subsequent concurrence with other events; and
3. the order of events.

For instance, being unemployed can be a very different experience within a life course, with different causes and consequences, whether it lasts 2years or 2months; whether it takes place early in the career when a worker is acquiring experience, or later when he or she is supposed to be fully employable; and whether it happens before or after a period of stable employment.

Second, SA is *holistic*. It considers sequence objects as consistent entities, with a narrative logic from beginning to end and stages of development in-between (whether these stages are known or simply hypothesised). There may be breaks and turning points on the way, which lessen the continuity of trajectories, but these only appear once objects are analysed in a holistic way, without being decomposed beforehand.

Third, SA has *a strong descriptive ambition*, realised by means of metric and visual outputs. This ambition stems from a recognition in principle that social scientific objects are not self-evident but require some descriptive work to be defined and understood. This characteristic may be regarded as trivial, as all social scientific objects deserve proper observation before moving to causal analysis. However, this is specifically true because sequences are novel objects in the social sciences, and social scientists still have a limited ability to make sense of, represent and interpret them. The last two characteristics of SA are partly a consequence of the previous one. Unlike the general linear model (Abbott, 1992) that dominates social statistics, the applicability of SA is not restrained by a range of complex and hardly ever met assumptions such as about the linearity of relationships between variables or the distribution of errors. SA is actually not focused on modelling reality, in the strict sense of deriving from theory or observation an ideational, simplified representation, on which to test hypotheses. It rather progresses by means of algorithms that convert complex materials into outputs that are more easily read and interpreted—however, using some theoretical and empirical knowledge all along this data reduction process. Finally, SA can accommodate various kinds of cases, various sample sizes, and various lengths of sequences—as long as these parameters are appropriate to the research at hand. This empirical versatility explains why SA has been applied successfully to a range of empirical objects, within several disciplines (see Variations and Variants to Standard Approaches section later in this entry).

Development of the Method

The introduction of SA in the social sciences owes a lot to the American sociologist Andrew Abbott. He published a series of articles on the theory and epistemology of sequences in the social and historical sciences (mostly collected in Abbott, 2001). At the same time, he undertook empirical applications, among which five articles on distinct objects are as many introductions to SA: steps in 18th-to19th-century British Midlands “Morris” folk dances (Abbott & Forrest, 1986; Forrest & Abbott, 1986); narrative variations in the many versions of the “Star Husband” Native American tale scattered across the United States and Canada (Forrest & Abbott, 1986); careers of 18th- and 19th-centurycourt and town German musicians (Abbott & Hrycak, 1990); order and precociousness of introduction of major social legislations in 18 Western countries

over a century and a half (Abbott & DeViney, 1992); and the progressive historical crystallisation of the rhetorical structure of the modern scientific article in a leading American sociology journal (Abbott & Barman, 1997). Through these applications to different structures of sequential data and theoretical frameworks, Abbott was demonstrating the promising scope and versatility of the method, as he made it more explicit in two landmark literature reviews (Abbott, 1995; Abbott & Tsay, 2000). At the same time, the dense and creative methodological developments contained in these empirical papers were acknowledgement of the challenges that remained to be taken up in order to turn the study of sequences into a fundamental ingredient of the social scientist's cookbook for longitudinal data. As Abbott explains (2001), introducing a new method requires not only to develop new perspectives but also to convince those who believe in the superiority of other methods such as event history analysis, time series or various regression models incorporating the time factor, that the newly developed method is making a significant contribution (Blanchard, 2016).

The development of SA faced several criticisms, which crystallised around the year 2000 (Levine, 2000; Wu, 2000, commenting on Abbott & Tsay, 2000). Some have been overcome through fruitful empirical applications, others through methodological and technical clarifications (Aisenbrey & Fasang, 2010). Other criticisms remain open puzzles to date either being ignored by sequence analysts or generating open disagreements and stimulating improvements (see Abbott's reply to Levine & Wu: Abbott, 2000). Beyond the assessment of the method itself, the controversy is revealing about the wider conditions of possibility of methodological innovation, sometimes their arbitrariness, and the underlying conceptions of quantification in the social sciences (Blanchard, 2016).

The debate also shows how Abbott's work is still relevant to current research. A systematic bibliographic review of uses of the method across the social sciences demonstrates that his work on sequences is still twice more frequently cited than the work of any other author in the field. Life-course studies have largely contributed to develop and refine Abbott's original propositions and to normalise the method's "core programme" (Gauthier, Bühlmann, &, & Blanchard, 2014; see the following section), even if at the expense of more peripheral objects and tools. Together with progress in computer power, availability of large biographical surveys and ad hoc software, they explain why, from the mid-1980s to the mid-2010s, the number of journals publishing SA studies has grown linearly, and publications, exponentially.

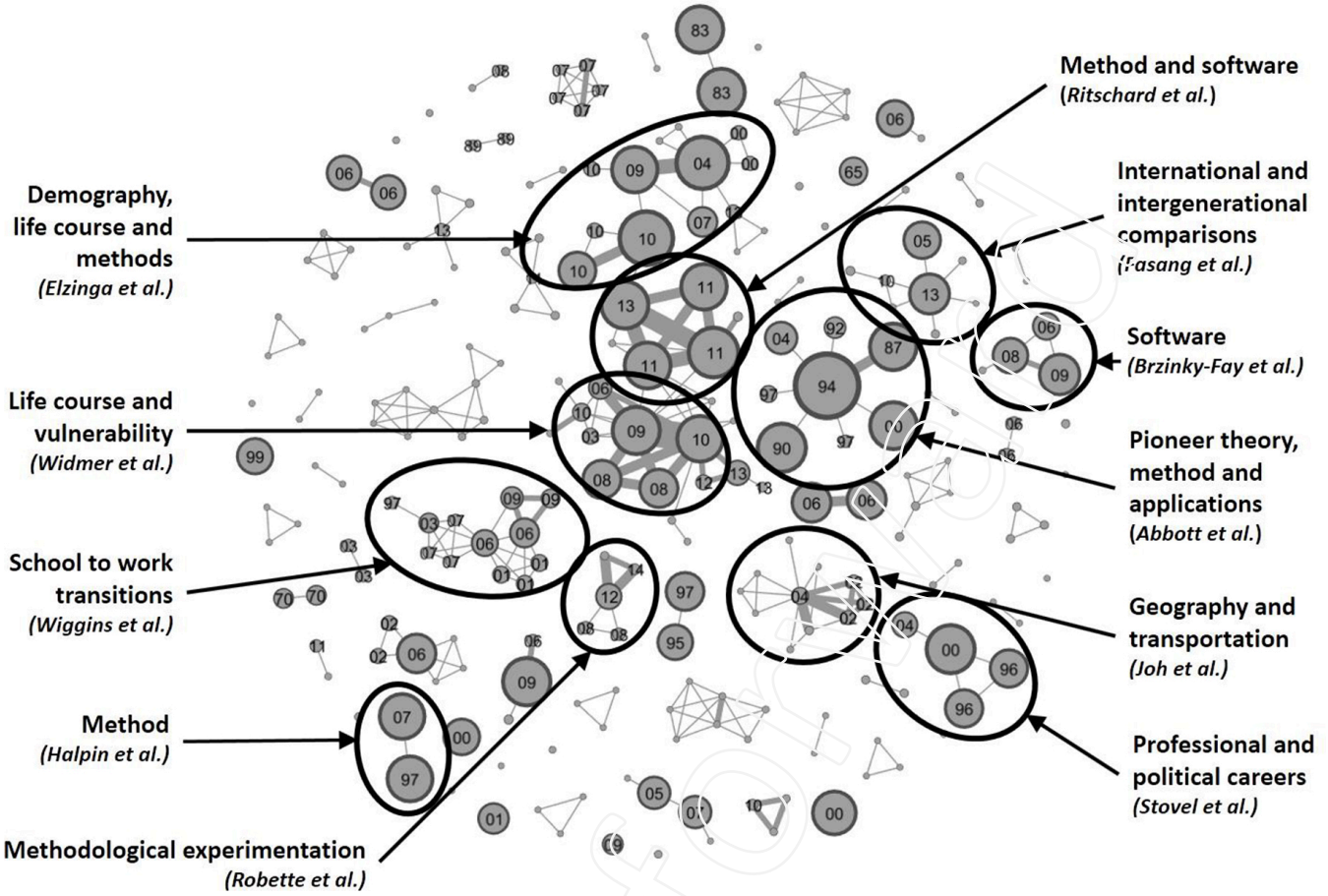
Nonetheless the diversity of the field should not be neglected. It can be measured in terms of empirical objects but also supporting journals, both those specialised in methods or life-course studies and those publishing more general social science research. Authors' disciplinary backgrounds are also varied, from the sociology of family, work, or art, to spatial disciplines (geography, urban science, transportation, and tourism), history, political sociology, international politics, and management. One may also distinguish two strands of contrasted approaches to SA (Blanchard, 2016). The *modelling approach* prioritises formal consistency, sample inference, numeric outcomes, and the possibility of replication. It relies on large individual- or household-level data sets from surveys and censuses, administrative records, and, more recently, simulated data for methodological improvements. A more *descriptiveapproach* insists on sociological and theoretical imagination, with a strong concern for data relevance and precision. It implements Abbott's (2001, 1995) core

project, which does not regard cases as conjunctions of abstract variables with sometimes limited meaning for sociological theory, and "fixed entities with attributes … monotonic causal flow … [and] univocal meaning" (pp. 40–47), but rather turns "from units to context, from attributes to connections, from causes to events" (p. 93). This approach often uses smaller, ad hoc samples, each case conceived as a conjunction of processes and interactions, with a strong descriptive ambition. It does not necessarily aim at inferring to a wider population or sometimes has access to the full population. These two approaches partially overlap with each other, yet they reveal partially diverging approaches to sequences.

In 2015, 30 years after its inception, the SA community is structured in collaborations along disciplinary and geographical lines, plus an archipelago of smaller, less connected teams and researchers (Figure 2). Clusters are distinguished by divides between generations (as defined by dates of publication in the field), but some of them show collaborations that bridge generations. The analysis of cross-citations within the same network (not displayed here) shows a high overall connectivity, much stronger than for collaborations. The level of overall cohesiveness of the domain of SA and the related academic community showed by this network analysis can be related to three factors: the uniqueness of sequence data, the specificity of the concepts and tools used to treat them, and the necessity for a small academic community to keep working together in order to establish itself.

Figure 2. Network of coauthorship and main empirical applications among users of sequence analysis.

Nodes represent authors. The size of nodes is their cumulated citation score in SA studies over 30 years (with a spline transformation that smooths out differences). Numbers in nodes are the average dates of each author's publications (only indicated for authors with more than fivecitations). Links between nodes materialise the number of coauthored pieces. Ellipses delineate the main clusters of authors, and labels indicate the main topics and most cited author within each cluster. Same corpus as in Figure 1.



Standard Steps in Sequence Analysis: The Core Programme

The coherence of the research strand around sequences comes from a series of standard steps of treatment, which may be called the "core programme" (Gauthier, Bühlmann, & Blanchard, 2014). This programme structures 80–90% of all studies and contributes to the process of collective methodological accumulation and to the consistency of the field. Six steps can be distinguished, each with specificities by comparison with nonsequential research: elaborating relevant research questions, collecting data, building a sequential data set, exploring this data set, comparing the sequences by pairs, and finally synthesising and interpreting the output of this comparison.

Conceptualisation

The first step consists of making sure that the research question at hand requires the use of SA. For this, the project needs to meet two conditions. The first condition is a focus on concepts such as processes, narratives, trajectories, biographies, accidents, transitions, turning points, and path dependency. All these concepts involve a combination of the three core aspects of social time mentioned earlier: duration, timing, and order. *Processes* and *narratives* develop over a certain duration, with certain elements happening in

a certain order. *Trajectories*, especially biographies, are processes made of states chosen or encountered by the actors from a set of possibilities that may be clearly defined and known, such as military ranks, and articulated around strategic orientations. Alternatively, states may be the result of more uncertain interactions and contexts and reconstituted ex post by the analyst. *Accidents* indicate sudden and unpredicted changes in processes or trajectories, whereas *transitions* are more or less progressive changes within the same. *Turning points* mark major transitions or multiple concurrent transitions that will modify substantially the following steps of the processes or trajectories. An accident may trigger or reveal a turning point. *Path dependency* expresses the fact that current stages in a process depend on previous stages, through institutional *legacies*, *promises* made to others, *strategies* that require time to bring outcomes, or *projects* that will only realise themselves once certain conditions are met.

The second condition for a research to require a sequential approach is that none of the more traditional methods be adequate. For example, one should not only aim at assessing how much change occurred between two time points, whether this is due to theoretical reasons or because full sequential data are not available. Neither should one only look at trends in continuous data nor how continuous data explain a given outcome or set of outcomes. In all these cases, other, more appropriate methods should be preferred to SA. However, if the conditions are met both for a sequence approach and for other methods, then a multimethod strategy is probably relevant.

As indicated earlier, the example of academic trajectories of access to professorship engages with several of the concepts mentioned. It also requires SA as a method because it addresses the variety of year-by-year changes over years- to decades-long periods. Finally, its alphabet is more complex than a unique scale, as it combines three dimensions: the contract held, as per the degree level required to access to it; the tasks involved (teaching, research, or a combination); and the location in or out of the national academic sector.

Data Collection

The second step is the collection of sequence data. Typical sources are individual or household surveys, either through classical, linear questionnaire or life history calendar; archives or interviews that are rich enough and systematic enough across cases to enable reconstituting biographies, trajectories in an organisation or social processes of some kind; extractions from administrative databases, such as civil service records for health, taxes or schooling, or company records for careers or continued education; and, increasingly, online or connected records of purchase activity, website visits, log-in location, mobile calls, or social media contributions. Data collection requires one to conceptualise the trajectories of interest, before firming up an alphabet, that is, a list of all possible states in all sequences. An alphabet needs to be welldefined and finite. It should reach a good level of exhaustiveness in the description of the empirical trajectories under study and at the same time match clearly the concept to be implemented.

A range of parameters also has to be decided upon, including the time axis, usually biographical or historical; the time unit, whether linear (e.g., years, months) or not (if defined by successive events happening at various intervals); the time limits, e.g., a fixed historical period, an age range, or the time span between two given

biographical events (e.g., graduation, marriage, appointment); and whether the sequences will all be of equal length or not. In case some degree of censorship is observed, whether to the left (e.g., for individuals who were not born yet or too young to be considered, or with undocumented periods) and/or to the right (e.g., for those who have not reached the upper age limit, or with unfinished trajectories at the time of data collection), the source of the censorship has to be identified, as well as its effects on the strategy of comparison between sequences (practical consequences of this are presented in the "Calculation of Dissimilarities" section).

Data Preparation

The third step usually consists of converting the data collected into a sequential data set. A first choice regards the time unit, which should be as precise as the data measured allow. In any case, it needs to document precisely enough the process under study. If precise dates are missing, or if the duration of episodes is not important to the research, then one may opt for an irregular time axis. In this case, each episode lasts 1, whatever its effective duration.

Within the previous example, based on a regular, yearly time axis, consider sequence S1=<aaabbbbbddf>. This is the format that results from the manual compilation of data extracted from university and personal websites, as well as from professional networking platforms like LinkedIn. S1 usually appears as S1'=<2000-02/a,2003-06/b,2007-08/d,2010/f>. This situation requires reformatting S1' into S1, and aligning it on the final event, f, which is the first year one is ranked as a professor. Moreover, some episodes may overlap. For example, consider S1''=<2000-2/a,2002-6/b,2007-8/d,2010/f>, where Individual 1 cumulates two contracts in 2007, at pre-PhD and PhD levels. Then, a decision has to be made about the nature of the combination of a and b. One may code a or b, if a hierarchy is established between the two states—in this case, it may be decided that b is more relevant than a because it marks a noticeable career progress and because it is known that the transition from a to b happened early in the year, therefore Individual 1 spent more time in b than in a. Otherwise one may choose to create a compound state ab = "Simultaneous contracts at pre-PhD and PhD levels"; the alphabet then becomes [a ab b c d e f]. More rules may be elaborated to deal with more complex combinations of states.

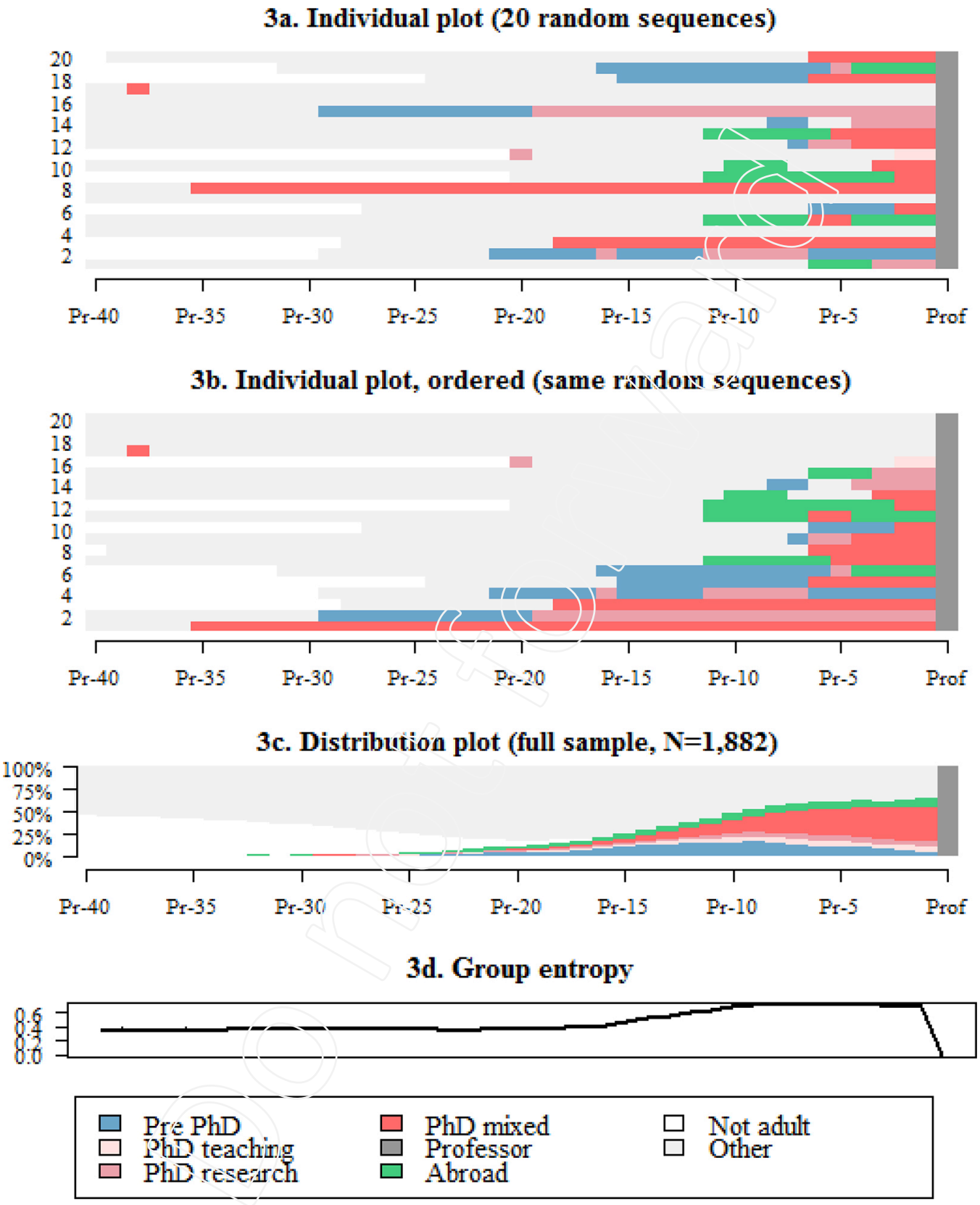
At the data preparation stage, one also needs to fix a strategy for data weaknesses. What is the threshold of missing values above which a case should be discarded? Within retained cases, will uncertain and unknown values be treated as missing? Or can positive values be imputed to them, for example using knowledge about neighbouring states and values realised in similar neighbourhoods? If two states share a given time slot, one may retain the longest one or decide on a cut-off point within the unit of time; for example, as civil years are usually split in August or September in academic careers, one may decide to retain the second code because it marks important change for the individual's work experience, both in the months preceding September (job application) and following it (new contract, possibly in a new university). Alternatively in the same example, one may decide that the time unit will be academic years, that is, 09.2000-08.2001, 09.2001-08.2002, and so on. When more complex arbitrations have to be made between multiple positions and transitions occurring at different moments within the chosen time unit, ad hoc trade-offs can be explored so that the chosen code

reflects best what happens during most of each time unit.

Exploration of Sequences

This fourth standard step can be crucial to figure out the dominant states in sequences, their overall arrangement, recurring order between states and possibly more obvious sequential patterns. Four kinds of tools may be used for this. Most useful is the individual (or “index”) sequence plot, which displays sequences as many horizontal bars (Figure 3a). Such a graph provides a precise visualisation of sequentiality, within and across sequences. In some cases, within-sample contrasts relevant to the research question may be materialised by ordering the sequences vertically. This order may be determined by a fixed covariate (e.g., life satisfaction at the end of the period), a time-varying covariate (e.g., age) or an index based on overall dissimilarities between sequences (Figure 3b, following Piccarreta, & Lior, 2010). Other kinds of plots are available, such as state distribution plots, which materialise for each time unit the distribution of all states in the sample, and are equivalent to a cross-tabulation of the alphabet by time (Figure 3c).

Figure 3. Summarising a sample of sequences.



Nongraphical exploratory tools can also be used. For example, the distribution of states over the whole sample gives an indication of which experiences are dominant in the sample, and which ones are marginal. One may also explore sequential data by means of composite, numeric indicators that vary over time. The

most popular ones evaluate a sequence's "complexity," or its variant, the "entropy" (Elzinga, 2010): the more a whole sequence departs from a (real or fictitious) nonvarying, one-state, stable sequence, the more complex and entropic it is deemed to be. Time-varying complexity indices over the whole sample (Figure 3d), or compared between groups, may also be calculated; in this case, the more variety of states between sequences at a given time point, the more complex the sample is. A last option for exploring sequential data is to look for the frequency of specifically relevant subsequences by means of mining formulas (Gabadinho, Ritschard, Studer, & Müller, 2011).

Calculation of Dissimilarities

The fifth standard step in SA has attracted the most attention due to its degree of sophistication and its key role in dealing with the combinatorial complexity of sequences. It consists of comparing sequences by pairs (say S1 and S2) and estimating their distance, or, more properly, their dissimilarity. Pairwise comparison of sequences is regarded quite consensually as the most relevant way to operate a controlled and meaningful reduction of the information contained in sequences and calculate a robust index of dissimilarity D(S1,S2). To do this, two families of methods have gathered significant amount of interest so far. The first one was theorised as a "non-aligning" technique (initially in Elzinga, 2003), based on combinatorial measures, and implemented by Alexis Gabadhino and colleagues (2011). It consists of counting the number of "common chunks" between S1 and S2, using one of several available metrics, such as the longest common subsequence (LCS) or the number of distinct common subsequences. These methods have the advantage of being intuitive both in their procedures and outcomes. The high calculation costs they require are now handled well by microcomputers. However, they are not yet completely implemented and not in all software. More importantly, little empirical feedback has been gathered so far about how well they retain the essential sequential information; in other words, scholars are unsure how well they discard noise (i.e., subsequences that are idiosyncratic to specific individuals in the sample). These are probably the main reasons why they are largely underused, by comparison with the second family of methods for sequence comparison: optimal matching (OM).

The OM algorithm (Sankoff & Kruskal, 1983) calculates dissimilarities through a state-by-state, left-to-right comparison of S1 and S2, using a set of elementary operations and associated "costs," or weights. Elementary operations are usually substitutions (a state in a given position in S2 replaces a different state in that position in S1), insertions (an element is inserted), deletions (an element is deleted), and matches (a state in S1 matches identically to the state in the same position in S2). Insertions and deletions costs, abbreviated as *indel costs*,or *icosts*, are typically set at 1, substitution costs (*scosts*)at 2 (as they are equivalent to an insertion plus a deletion), and matches at 0. Figure 4a is a strictly column-by-column comparison of two sequences (S1 and S2) that uses only substitutions and matches. In Figure 4b, insertions and deletions are also allowed; this increases the number of matches and reduces the overall cost of transforming one sequence into another. The total dissimilarity in this example goes down from 12 to 9.

Figure 4. Comparing sequences: The optimal matching approach.

4a. Comparing sequences: the straight way (using substitutions and matches only)

Years	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	Costs of operations
S1	**a**	**a**	**a**	**b**	**b**	**b**	**b**	**b**	**d**	**d**	**f**	
1 substitution	a-Ø											2
1 match		a-a										0
1 substitution			a-b									2
1 match				b-b								0
3 substitutions					b-e	b-e	b-e					3*2
1 substitution								b-d				2
2 matches									d-d	d-d		0
1 match											f-f	0
S2	**Ø**	**a**	**b**	**b**	**e**	**e**	**e**	**d**	**d**	**d**	**f**	**12**

4b. Comparing sequences: the optimal way (allowing insertions and deletions)

													Costs of operations
S1	**a**	**a**	**a**	**b**	**b**	**b**	**b**	**b**	**d**	**d**	**Ø**	**f**	
2 insertions	a-Ø	a-Ø											2*1
1 match			a-a										0
2 matches				b-b	b-b								0
3 substitutions						b-e	b-e	b-e					3*2
2 matches									d-d	d-d			0
1 deletion											Ø-d		1
1 match												f-f	0
S2	**Ø**	**Ø**	**a**	**b**	**b**	**e**	**e**	**e**	**d**	**d**	**d**	**f**	**9**

Ø indicates the absence of a state in the position

The OM algorithm can prove even more flexible if substitution costs are allowed to vary. For example, if some information is available about the theoretical degree of dissimilarity between a and b, then the analyst may set "objective" or "theoretical" *scosts* that reflect this knowledge; otherwise, *scost*(a,b) may be calculated as the inverse of the rate of transitions between a and b in the sample at hand, in a reference sample or in the population, following the principle that substitution of states between which transitions are rare should cost more.

In brief, OM allows time-warping (Halpin, 2014), that is, aligning together sections of S1 and S2 that are identical or similar but that happen at different moments. This flexibility is needed to seize, for example, the similarity between stable lives interrupted at different ages by periods of unstable work or couple situations or between democratic regimes interrupted in different years by periods of institutional crisis, and these interruptions may last more or less. OM accounts for the fact that social times and rhythms (Lesnard, 2014) do not match the standard, linear time inherited from the scientific industrial revolution; it can be stretched or compressed by accidents or agency.

A second algorithm is chained up with OM. Named after its authors S. B. Needleman and C. D. Wunsch (1970), it identifies the optimal sequences of operations to transform S1 into S2, that is, the ones that minimise the total sum of weighted costs and produces the optimal D(S1,S2). This optimal sequence of transformations follows two concomitant aims: using as much as possible the "cheapest" elementary

operations, and maximising the number of matches, in other words, spotting S1 and S2's common subsequences. Other chains of operations than the one used in Figure 4b may lead from S1 to S2, but they will not be "cheaper" than 9.

The power of OM has been demonstrated by a wealth of feedback from empirical studies. OM reveals time-lagged common patterns including through the noise that would hinder the manual (visual) identification of common patterns. Through the setting of costs, it can be adjusted so as to emphasise one or two of the fundamental dimensions of time—duration, timing, and order. SA is an analytical decomposition of social time, followed by a reconstitution of its consistency and structuring role for individuals and groups (Lesnard, 2014).

Postdissimilarity Treatments

The sixth typical step in SA consists of interpreting the dissimilarity matrix (DM). DM creates a structure inside the population of sequences, with a symmetry along a main diagonal made of 0s. Each sequence is more or less close to any other sequence. Pairs of sequences composed of very different states, and/or in very different proportions, and/or in a very different order, will have higher values in the DM, whereas identical sequences have a distance of 0. Each row r_i in the DM describes a profile of distances between case i and the (N−1) other cases. Two sequences with proportionate r_i profiles are structurally equivalent. The less identical or structurally equivalent sequences in a given sample, the higher the mathematical dimensionality d ($1<d<N$) of the sample, the more complex it is to reduce the DM into a meaningful summary.

Clustering is the most popular treatment of the DM because its principle is intuitive, its procedure easy to tailor to various research objectives, and its results are nearly always interpretable (Abbott & Barman, 1997; Forrest & Abbott, 1986; Han & Moen, 1999). A standard clustering algorithm is called "ascending" (or agglomerative) and hierarchical. It proceeds by progressive aggregation of cases, from the smallest branches of the clustering tree, to the trunk, which represents the whole sample. All the rows are therefore compared by pairs along a certain arithmetic aggregation algorithm, so as to calculate all pairwise indices of distances. The two cases with smallest distance are grouped and next assimilated as a compound case. Then, the same aggregation algorithm is applied to the (N−1) cases, more cases are grouped, and so on, until a sensible number of clusters is reached. These clusters form a partition of sequences whose quality may be assessed by means of various measures (Studer, 2013), before clusters are named and described by the sequence exploration tools listed in the Explorations in Sequences step: various kinds of sequence graphs comparing clusters, nongraphical exploratory tools, complexity indices, and sequence mining to determine the subsequences specific to each cluster, as well as unspecific ones. Nonsequential covariates may also be used in cross-tabulations and regression modelling, using the typology either as an independent categorical variable, when testing if trajectories (e.g., country regime changes) explain present outcomes (e.g., current degree of democracy) or trajectories over a later period (e.g., democratic upholding after initial democratic onset), or as a dependent variable, when testing if a one-time or fixed characteristic (e.g., a revolution) determines a trajectory (e.g., regime change).

Clustering reduces a multidimensional space of sequences to a one-dimensional typology. Other methods

provide more open and stimulating interpretation of sequential dissimilarities. For example, the DM may be represented as a network of sequences. Each case is a network node. The closer and the more linked two nodes, the more the sequences are similar. This produces a more flexible geography of the sequence space than clustering. Multidimensional scaling (MDS) may also be applied to the DM (Abbott & Forrest, 1986; Abbott & DeViney, 1992). MDS extracts a hierarchy of continuous factors that structure the DM. MDS factors do not refer to any concrete variable, but they may correspond to objective determinants of trajectories predicted by theory or intuition. The two or three main factors may then be graphed in scatterplots, with cases as dots, covariates as dot graphical properties, and compared across other covariates, so as to visualise efficiently the link between dimensions and covariates (Piccarreta & Lior, 2010).

Matthias Studer and colleagues (2011) pushed further the analysis of the relationship between sequences and covariates. Extending the principles of analysis of variance based on the sum of squared distances, they use the DM to calculate the "discrepancy" of a sample and of groups, and how much of these a given (static) covariate explains, over time. This approach "allows for studying the links between individuals' trajectories and their contexts, while at the same time preserving the notion of between-individual variability" (Studer et al., 2011, p. 491). With its multifactorial extension, this method is a crucial improvement to the widely used regression of DM-based cluster membership on covariates.

Variations and Variants to Standard Approaches

Successes and Limitations of the Core Programme

The standard approach to SA has enabled crucial progress in current understanding of a range of phenomena. The following are noticeable applications:

- the variety of educational and professional careers, how they differ between sexes, generations or countries, how they unveil different tracks of social mobility, and their link to life satisfaction, health or income (e.g., Abbott & Hrycak, 1990; Stovel, Savage, & Bearman, 1996; Wiggins, Erzberger, Hyde, Higgs, & Blane, 2007);
- the timebudgets of working couples and their impact on family cohesion (e.g., Lesnard, 2008);
- the historical deinstitutionalisation, destandardisation, differentiation and individualisation (Aisenbrey & Fasang, 2010) of life courses, especially the transitions from adolescence to adult age, and from work to retirement, how well various individuals cope with these transitions, and their outcomes in terms of satisfaction, stability, and success (e.g., Han & Moen, 1999);
- the careers of political and economic elites, their links with elite networks, and how these careers and networks shape the economic order and its crises (e.g., Blanchard, Dudouet, & Vion, 2015);
- sequences of political participation through elections, social mobilisation, conflict, and violence (e.g.,

Casper & Wilson, 2014);

- residential mobility, commuting patterns and tourist trips (e.g., Shoval & Isaacson, 2007); and
- sequences of web visits, social media networking, and online interactions (e.g., Hay, Wets, & Vanhoff, 2010).

The standard treatment process, as described in the previous section, should not hide methodological developments, variants, and criticisms that have not received all the attention they deserve. This is because the core programme (Gauthier, Bühlmann, & Blanchard, 2014) relies on a number of assumptions that may not be valid for all SA projects. SA would mainly deal with life trajectories of individuals (rather than collective entities or human artefacts), in contemporary times (rather than in the distant past), and documented through survey questionnaires (rather than various nonintrusive methods for data collection, such as administrative data, web traces, and archives). It is also often assumed that sequences are of equal lengths (whereas so many sequence data are made of trajectories of uneven length or are truncated), aligned on a linear time axis, age or historical time, with a fixed time unit, usually year or month (whereas social time is known to usually lack the mathematical properties of linearity, forwardorientation, and regular density).

The core programme also relies on the OM approach to sequence comparison (at the expense of the approach through combinatorial measures), with a range of specific options: It uses only the four basic elementary operations of matching, inserting, deleting, and substituting (excluding, for example, swaps, that transform <ab> into <ba>); *icosts* are unique and fixed for all insertions and deletions (excluding *icosts* that would vary with time, between states or depending on the length of sequences); matches cost zero (negative costs could reward pairs of sequences with more matches); substitution costs are usually derived from transition rates or objective information (sometimes the anti-identity cost matrix, which keeps all substitution costs equal, could work as well, with less theoretical justification required); and the optimal sequence of operations introduced into the DM is supposed to be the “cheapest” one (whereas more elaborate alternatives could reflect a more theoretically informed degree of dissimilarity). Finally, post-OM treatments are often centred on hierarchical cluster analysis based on simple or squared Euclidean distance and Ward aggregation algorithm, in spite of the availability of a range of alternatives. The discussion on clustering focuses unduly on statistical criteria used to justify the cutting off of the clustering tree, instead of selecting the clusters that maximise internal sequential homogeneity, while minimising external homogeneity, through visual and statistical examination.

In brief, the literature on SA, partly due to its limited volume by comparison with older and more established methods, and for the sake of simplicity and formal consistency, has come to follow some options that may not be optimal for all projects. Several interesting directions of development remain at the stage of possibilities or theoretical propositions. With time, more of the aforementioned alternatives will probably be tested. Here, a few open debates and unresolved puzzles that pave future pathways for SA are discussed.

The Complexities of Sequence Comparison

Is OM effectively the best tool for sequence comparison? This algorithm has been applied for over 30 years to

a range of topics and data, with valuable cumulative feedback. It is proven able to reveal time-lagged common patterns through a certain degree of noise. It is flexible regarding a range of data issues: Missing, uncertain, and multiple values may be treated in specific ways that will reduce their impact on the analysis, yet preserve the information that they may still convey. Finally, OM may be adjusted to pinpoint more or less each of the three fundamental aspects of social time: duration, timing, and order.

However, OM is a complex mechanism. As an algorithm, it delivers a sequence of operations that transform sequence S1 into sequence S2, but not an equation linking S1 and S2, a mathematical model that embeds all sequences in the sample or a model of the data-generating process. Several distinct pairs of sequences may end up with the same dissimilarity measure D, if different sequences of weighted transformations add up to the same total cost. Unlike some genetic models, OM does not attempt to model the effect of some causes on trajectories; "rather it is a computationally-efficient heuristic that captures significant features of the phenomenon" (Halpin, 2014, p. 101). A small tweak in the cost settings can increase or decrease dissimilarities in unpredictable proportions. In other words, converting the high dimensionality of a sample of sequences into a unique numeric scale is a drastic data reduction, that should be performed with caution. Hence, what methods for dissimilarity calculation, or "metric," should be used in which circumstances? An answer to this question consists in testing known metrics on benchmark data sets and comparing how well they detect various kinds of sequential patterns. For example, Brendan Halpin (2014) uses monthly labour market data from the British Household Panel Study to compare OM with combinatorial measures and with a variant of OM that takes into account the context of aligned subsequences. Other systematic studies tend to rely on simulated data sets, each designed to test metrics against specific sequential patterns.

Unsurprisingly, such studies conclude that no metric is universally superior and that some are more relevant than others to analyse a given aspect of social time. However, they rely on fairly standard data formats, such as sequences with even length, with limited room for data weaknesses, and based on fairly big samples. They miss part of the picture, such as the fact that, if the lengths $l1$ and $l2$ of S1 and S2 are such that $l1<l2$, then OM indels are a crucial parameter to treat at least ($l2$-$l1$) states; this situation is core to understanding intergenerational life-course differences. More generally, the exponential number of combination of elementary sequential patterns means that tests, either based on real or simulated data, are hardly universal. Benchmark comparisons deliver useful guidance at the exploratory stage of research, but the combinatorial complexity involved in most metrics prevents building efficiently models with a method, SA, that was precisely theorised and designed as an alternative to modelling. The key intuition is again Abbott's (2001): "Mathematically, the space of all possible sequences is largely empty because most things that could happen don't" (pp. 16, 166–169). This emptiness justifies a procedure that searches for local orders inside the space of sequences, instead of disassembling the states or events into "combinations of (supposedly) independent variable properties," as the linear model does (Abbott, 2001, p. 169). The real sequences contained in a given population can be compared to fish swimming in a lake. A systematic three-dimension search, like a random inspection a sonar would perform, is not fruitful because fish follow a limited number of routine paths around weeds and rocks at some preferred depth and speed. It is more efficient to track the fish, study their most frequent paths and paces, and see how these paths and paces differ from

each other (Blanchard, 2011).

Future simulation studies might bring more decisive help in sequence comparison. As for now, it appears difficult to predict a best metric for a given project. Some highly structured data will speak easily across metrics, whereas others will need many trials and errors. Empirical publications point at an encompassing approach, one that considers four concurrent dimensions:

1. *Data specifics*. For example, if successive states are known, but many dates are missing, then one may rely on LCS with discarded duration; if lengths are very uneven, then one should set OM *icosts* so as to adequately weigh this unevenness.
2. *Data quality*. If many occurrences of a given state are missing, one may decide to exclude it from the calculations or to assign a certain role and cost to it, depending on assumptions or hypotheses made about its meaning.
3. *Research question*. If testing the impact of the first steps on subsequent trajectories, such as education on subsequent work, a mix of the longest common prefix and LCS may work. If timing is of crucial concern, such as with hypotheses rooted in age or historical landmark events, one may choose OM with high *icosts* so as to use more substitutions at close dates. If the focus is order, then LCS is a good option.
4. *Feedback*. If previous studies, using same or similar data, appear revealing, one may use the same parameters. For example, using exclusively substitutions with time-varying costs has proved successful for studying sequences of linked lives, such as people living in a couple (Lesnard, 2008).

Choice of a metric often results from a combination of these four dimensions, in varying proportions. However, a fifth dimension emerges, seldom explicit, yet quite present: *pragmatic efficiency* of the metric chosen. For example, one may test OM with fixed *scosts* before moving to more complex metrics only if the outcome is not meaningful. More generally, sequence comparison is essentially a descriptive method, as it does not involve constraining assumptions such as linearity or normality. Its purpose is to reveal some regularities or some order within a sample. Hence, as long as such regularities or order emerge from the treatments based on the DM, and the analyst is able to interpret them, then the method may be regarded as adequate.

To conclude, a successful project entails articulating judiciously these five dimensions, without falling into common traps: the seduction of formal sophistication, a-theoretical induction, and blind reproduction of existing studies.

Multichannel Sequence Analysis (MCSA)

A comprehensive description of sequential phenomena often requires more than one alphabet. This is the case when considering laws passed in distinct domains of social legislation (Abbot & DeViney, 1992), branch and position levels in professional careers (Stovel, Savage, & Bearman, 1996), or couples' linked schedules (Lesnard, 2008). MCSA, also known as multiple SA or multidimensional SA, addresses this more complex variant of the method. For sociologists, MCSA is probably the most realistic approach to sequences, as it

is widely recognised that the life course unfolds in multiple social "worlds" or "spheres." These spheres are partly autonomous, but also interact dynamically, which requires a joint approach. Shin-Kap Han and Phyllis Moen (1999) were the first to address this challenge explicitly, in their study of transitions to retirement of U.S. large company employees through their occupations (64 types), work statuses (11), and organisations (5). More robust and universal solutions were later proposed, based on employment, housing, and family trajectories in the United Kingdom (Pollock, 2007) and family and occupational trajectories in Switzerland (Gauthier, Widmer, Bucher, & Notredame, 2010).

The difficulty is that MCSA combinatorics increase exponentially in comparison to monochannel SA (mcSA), making it difficult for the analyst to represent and control a sequence treatment which, as demonstrated earlier, already resists a proper modelling strategy. For a simplified case of c channels andalphabets, each composed of s different states, observed over a t-long time span, there are $(s^c)^t$potential multichannel (MC) trajectories, compared to only s^t distinct trajectories for a one-channel design. This is further complicated by the increased sociological complexity of MC trajectories: At each moment of his or her trajectory, an individual holds several positions that may be linked—legally, socially, symbolically, existentially—to each other, as well as to the multiple positions occupied previously and subsequently. Hence, the usual tools of SA need to be reassessed to fit objectives made more ambitious by MC complexities.

Four main MCSA strategies have been proposed so far. One consists in merging alphabets and falling back onto a mcSA configuration (Aassve, Billari, & Piccarreta, 2007). It is adequate where the resulting MC alphabet remains of sensible size and complexity. Another strategy is to treat each channel separately through sequence comparison and subsequent treatments, before comparing the channel-specific results with each other and with covariates. This is appropriate when each channel deserves separate inspection, before looking at their interactions. Richard D. Wiggins and colleagues (2007) take this path in their study of labour market, relationship, and housing histories of 300 British residents whose social, dietary, health, and anthropometric conditions had been surveyed in their childhood in the 1930s, and who were interviewed again in the 1990s. They establish ideal-type trajectories in the three life domains and contrast in each life domain a structurally advantaged idealtype with other kinds of trajectories. Then, they compare the role played by personal histories and current conditions in self-reported quality of life at old age, thereby concluding in a robust way about the biographical legacies that lead to more or less happy retirement times. A third MCSA strategy also relies on mc calculations of dissimilarities, but DMs are then combined linearly so as to fall back onto a unique, MC DM, which is then treated by means of usual mc techniques. Han and Moen (1999) apply this strategy to their three-channel design (occupations/work statuses/organisations), before treating the MC DM by means of analysis of varianceon within- and between-cluster distances, and regressions of covariates on gender, cohort, and clusters. Finally, a fourth solution was proposed by Gary Pollock (2007). It consists of developing an MC version of OM that most successfully transfers the principle of the aligning principle to MC configurations. At each step of the comparison of two individual biographies, c elementary operations are selected, one for each channel. This means examining together c tables identical to the one shown in Figure 4. The optimal path to transform S1 into S2 is the one that minimizes the total of the dissimilarity costs, possibly weighted.

In sum, all strategies may work under specific circumstances. The first one depends on fairly stringent conditions regarding data, but it has the advantage of being concrete and intuitive. It also simplifies greatly subsequent treatments. The next two strategies enable the control of MC results by means of intermediary mc outcomes. However, mc DMs may account very incompletely for MC sequential dynamics. This is even more true for mc typologies, which may in addition generate more cluster combinations than can be made sense of. The last solution may be more abstract and less traceable, but it has the advantage of expressing the multidimensional dynamic of sequences. The difficulty however is to interpret MC outputs. For example, an MCSA-based typology often comprises a few, large, trivial clusters, and many small, idiosyncratic clusters, missing the medium-size, meaningful clusters, that would most advance theory.

The complexity involved in MCSA makes it all the more necessary to work with simple tools and analytical strategies in the first place. This includes examining each channel separately, so as to detect and address data weaknesses, understand the sequential dynamic in each channel, and, on this basis, elaborate the most appropriate MCSA strategy. In the case of three or more channels, monochannel explorations may lead to merging two channels, before proceeding to MCSA. Graphically, individual plots at mc level, which help in understanding the contribution of each channel to each cluster, can be combined with graphs at MC level (Blanchard, 2011; Fillieule & Blanchard, 2013), which materialise the concurrence of patterns between channels, and help imagine their interactions.

Prototypical Sequences

Prototypical sequences provide useful summaries of groups of sequences (typically, obtained through clustering), alternative to group-level statistics and graphs. Well-chosen cases enable visualising on single cases the sequential characteristics of the group. They also enable connecting SA with other treatments, such as interviews or other in-depth biographical research about these single cases. Several methods are available to build prototypical sequences. One method consists of piecing together the modal states at each time point, resulting in a fictitious, ideal-typical, modal sequence (Aassve, Billari, & Piccarreta, 2007; Gabadinho et al., 2011). This approach is intuitive and straightforward, yet it may not account well for tied or close-to-tied modes, nor for modal transitions. Otherwise one may identify the (real) medoid sequence, that is, the one that minimises the average distance to all other sequences in the group (Aassve et al., 2007; Gabadinho et al., 2011). This has the advantage of grounding the extraction of prototypes in the same calculation of dissimilarities that are used for other treatments, although it is all the more theoretically meaningful that this distance has been theorised for the case at hand. A more encompassing solution is based on the tree of all sequential variations in the group, from the left (the trunk) to the right (the smallest branches), with exuberant trees pruned down to branches beyond a given threshold (Aassve et al., 2007). This method better preserves the diversity of pathways and enables visualising clearly how they diverge from each other. Finally, a more sociologically informed approach consists of interpreting the group with all relevant sequential variables and covariates, before identifying the cases that fit best this interpretation (Fillieule & Blanchard, 2013).

Recent and Future Developments

Other noticeable future directions for SA includes the role of *visualisations*. Visualisations are moving slowly away from state distribution plots (Figure 3b), which are made of easy-to-grasp, conveniently ordered shades, but that sometimes miss or even contradict sequentiality itself. Meaningfully ordered and simplified individual plots, complemented with tailored metric sequential summaries, well-chosen or well-built prototypical cases and adequate subsequence mining, deliver stronger depictions of time patterns. Statistical and graphical improvements are also brought to standard hierarchical clustering procedures, which sometimes reveal themselves as excessive and arbitrary data reductions. Interesting developments include two-dimensional maps of clusters and regression trees. These are landmarks towards more encompassing and manipulatable representations of complex sequence spaces, that allow a flexible search and comparison of the diverse time patterns. They enable more robust clustering decisions and interpretations, and facilitate the exploration of the role of covariates.

Another direction is the *relationship between sequences and events*. Each transition between two states is an event, so event sequences are equivalent to state sequences and state-SA tools can be recycled for event-based SA. However, it is crucial to move further and combine phenomena that last (i.e., one or more time units) with others that do not (i.e., they are shorter than one unit), for life-course sequences (positions and statuses together with births, illness, or bereavements) and even more for historical and political sequences (stable periods with one-off crises or accidents). One solution is the use of an individual event transversal to a whole sample as axis origin to align sequences graphically, but more can be imagined. More attention should also be paid to events that are exogenous to sequences, yet impact some of them. This would help introduce organisational or historical contexts, more precisely than the wide periods that demographers sometimes refer to as contexts of sequences.

More work is also needed on sequences that may develop jointly because of *observed links*, for example, members of households or organisations, and individuals or groups with network relationships. Such samples require to test the impact of these links compared to covariates that do not bind cases objectively. Ongoing explorations include representing sequences by means of network analysis tools, with states as nodes and transitions as oriented edges; testing associations between sequence complexity and network centrality measured on the same cases; and explaining the evolution of positions and roles of individuals in a network by means of their sequential characteristics. The sequence-network methodological conundrum contributes to the understanding of longitudinal networks, and symmetrically of sequences within networks.

Conclusion

Sequence analysis has developed quickly over 30 years, with more diverse objects treated, from more diverse disciplinary and paradigmatic angles and by a wider intellectual community. Theorists of the method, programmers and users have engaged a fruitful collaboration. More should probably be done to integrate some highly sophisticated algorithmic propositions into concrete case studies. To this aim, it is important to

keep a logic of picking and tailoring existing tools to one's theoretical and empirical needs.

Some objects of obvious sequential nature, with available data and pending theoretical puzzles, remain largely to be explored by means of SA, such as the dynamics of rhetorical and argumentative interactions through topics, arguments or keywords in face-to-face and online conversations; the dissemination of discursive structures in formalised discourses such as political speeches, literary fiction, or media stories; the recurrence of modes of action and interaction along the development of a mobilisation for a cause; the timing of policy processes, within the constraints of law and institutions; migration and health life course, which come in complement to work and family trajectories; and comparison of life courses between periods, countries, and generations.

Last but not least, online or connected big data, collected ex postor live by public, commercial, and third-sector organisations, or through ad hoc sensing devices, now provide abundant longitudinal material, sometimes exceptionally precisely dated, about known social phenomena, as well as new ones. Turning this material into proper sequential data sets requires addressing unprecedented challenges regarding the size, quality, and readability of such data. This is likely to require the use of sophisticated algorithms for noise filtering, imputation of missing values, mining, sequence comparison and visualisation, including by making use of machine learning, simulations and dynamic programming. However, the ability to tailor simple and robust tools to the data's characteristics, in line with the descriptive tradition of SA mentioned earlier, should also remain a key asset for sequence analysts.

Further Readings

Biemann, T. (2011). A transition-oriented approach to optimal matching. *Sociological Methodology*, 41, 195–221. doi:10.1111/j.1467-9531.2011.01235.x

Billari, F., **Fürnkranz, J.**, & **Prskawetz, A.** (2000). Timing, sequencing and quantum of life course events: A machine learning approach. *European Journal of Population*, 22, 37–65. doi:10.1007/s10680-005-5549-0

Colombi, D., & **Paye, S.** (2014). Synchronising sequences. An analytic approach to explore relationships between events and temporal patterns. In **P. Blanchard**, **F. Bühlmann**, &, & **J.-A. Gauthier** (Eds.), *Advances in sequence analysis: Methods, theories and applications*. London: Springer.

Fasang, A., &, & **Liao, T. F.** (2014). Visualizing sequences in the social sciences: Relative frequency sequence plots. *Sociological Methods &Research*, 43, 643–676. doi:10.1177/0049124113506563

Henriksen, L. F., & **Seabrooke, L.** (2015). Transnational organizing: Issue professionals in environmental sustainability networks. *Organization*, 1–21. doi:10.1177/1350508415609140

Joh, C. H., **Arentze, T. A.**, & **Timmermans, H. J. P.** (2001). Multidimensional sequence alignment methods for activity-travel pattern analysis: A comparison of dynamic programming and genetic algorithms.

Geographical Analysis, 33, 247–270. doi:10.1111/j.1538-4632.2001.tb00447.x

Massoni, S., **Olteanu, M.**, & **Rousset, P.** (2009). Career-path analysis using optimal matching and self-organizing maps. Retrieved from http://hal.archives-ouvertes.fr/docs/00/40/91/14/PDF/ wsom09_V2.pdf

Piccarreta, R. (2015). Joint sequence analysis: Association and clustering. *Sociological Methods & Research*, 46, 252–287. doi:10.1177/0049124115591013

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

Robette, N., **Bry, X.**, **Lelièvre, E.**, **Elzinga, C. H.**, **Fan, W.**, **Moen, P.**, **Fasang, A. E.**, et al. (2015). Symposium: Life-course sequence analysis. *Sociological Methodology*, 45, 1–100.

References

Aassve, A., **Billari, F.**, & **Piccarreta, R.** (2007). Strings of adulthood: A sequence analysis of young British women's work-family trajectories. *European Journal of Population*, 23, 369–388. doi:10.1007/s10680-007-9134-6

Abbott, A. (1992). From causes to events: Note on narrative positivism. *Sociological Methods and Research*, 20, 428–455.

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21, 93–113. doi:10.1146/annurev.so.21.080195.000521

Abbott, A. (2000). Reply to Levine and Wu. *Sociological Methods and Research*, 29, 65–76. doi:10.1177/0049124100029001004

Abbott, A. (2001). *Time matters: On theory and method*. Chicago, IL: University of Chicago Press.

Abbott, A., & **Barman, E.** (1997). Sequence comparison via alignment and Gibbs sampling. *Sociological Methodology*, 27, 47–87. doi:10.1111/1467-9531.271019

Abbott, A., & **DeViney, S.** (1992). The welfare state as transnational event: Evidence from sequences of policy adoption. *Social Science History*, 16, 245–274.doi:10.1017/S0145553200016473

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

Abbott, A., & **Hrycak, A.** (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology*, 96, 144–185. doi:10.1086/229495

Abbott, A., & **Tsay, A.** (2000). Sequence analysis and optimal matching methods in sociology: Review and

prospect. *Sociological Methods and Research*, 29, 3–33. doi:10.1177/0049124100029001001

Aisenbrey, S., & **Fasang, A. E.** (2010). New life for old ideas: The "second wave" of sequence analysis. Bringing the "course" back into the life course. *Sociological Methods & Research*, 38, 420–462. doi:10.1177/0049124109357532

Blanchard, P. (2011). *Sequence analysis for political science*. Working Papers of the Committee on Concepts and Methods, International Political Science Association. Retrieved from http://www.concepts-methods.org/WorkingPapers/PoliticalMethodology

Blanchard, P. (2016). Les vicissitudes de l'innovation méthodologique. 'Validité, falsifiabilité, parcimonie, consistance, précision,etc.' [The challenges of Methodological Innovation: "Validity, falsifiability, parsimony, consistence, precision and so on"]. In **M. Jouvenet** & **D. Demazière** (Eds.), *Andrew Abbott, sociologue de Chicago. Héritages, dépassements, ruptures* (pp. 151–171). Paris, France: EHESS.

Blanchard, P., **Dudouet, F.-X.**, & **Vion, A.** (2015). Le cœur des affaires de la zone euro. Une analyse structurale et séquentielle des élites économiques transnationales [The core of the Eurozone business. A structural and sequential analysis of transnational economic elites], *Cultures et Conflits*, 98, 71–99.

Casper, G., & Wilson, M. (2014). Using sequences to model crises. *Political Science Research and Methods*, 3, 381–397. doi:10.1017/psrm.2014.27

Elzinga, C. H. (2003). Sequence similarity: A nonaligning technique. *Sociological Methods and Research*, 32, 3–29. doi:10.1177/0049124103253373

Elzinga, C. H. (2010). Complexity of categorical time series. *Sociological Methods & Research*, 38, 463–481. doi:10.1177/0049124109357535

Fillieule, O., & **Blanchard, P.** (2013). Fighting together. Assessing continuity and change in social movement organizations through the study of constituencies' heterogeneity. In **N. Kauppi** (Ed.), *A political sociology of transnational Europe* (pp. 79–108). Basingstoke, England: ECPR Press.

Forrest, J., & **Abbott, A.** (1986). The optimal matching method for anthropological data: An introduction and reliability analysis. *Journal of Quantitative Anthropology*, 2, 151–170

Gabadinho, A., **Ritschard, G.**, **Studer, M.**, & **Müller, N.** (2011). *Mining sequence data in R with the TraMineR package: A user's guide*. Geneva, Switzerland: Department of Econometrics and Laboratory of Demography, University of Geneva.

Gauthier, J.-A., **Bühlmann, F.**, & **Blanchard, P.** (2014). Introduction: Sequence Analysis in 2014. In **P. Blanchard**, **F. Bühlmann**, & **J.-A. Gauthier** (Eds.), *Advances in sequence analysis: Theory, methods, applications* (pp. 1–17). London, England: Springer.

Gauthier, J.-A., **Widmer, E.**, **Bucher, P.**, & **Notredame, C.** (2010). Multichannel sequence analysis applied

to social science data. *Sociological Methodology*, 40, 1–38. doi:10.1111/j.1467-9531.2010.01227.x

Halpin, B. (2014). Three narratives of sequence analysis. In **P. Blanchard**, **F. Bühlmann**, & **J.-A. Gauthier** (Eds.), *Advances in sequence analysis: Methods, theories and applications* (pp. 75–103). London, England: Springer.

Han, S.-K., & **Moen, P.** (1999). Clocking out: Temporal patterning of retirement. *American Journal of Sociology*, 105, 191–236. doi:10.1086/210271

Hay, B., **Wets, G.**, & **Vanhoof, K.** (2010). Segmentation of visiting patterns on websites using a sequence alignment method. *Journal of Retailing and Consumer Services*, 10, 145–153.

Lesnard, L. (2008). Off-scheduling within dual-earner couples: An unequal and negative externality for family time. *The American Journal of Sociology*, 114, 447–490. doi:10.1086/590648

Lesnard, L. (2014). Using optimal matching analysis in sociology: Cost setting and sociology of time. In **P. Blanchard**, **F. Bühlmann**, & **J.-A. Gauthier** (Eds.), *Advances in sequence analysis: Methods, theories and applications* (pp. 39–50). London, England: Springer.

Levine, J. (2000). But what have you done for us lately? Commentary on Abbott and Tsay. *Sociological Methods and Research*, 29, 34–40. doi:10.1177/0049124100029001002

Needleman, S. B., &**Wunsch, C. D.** (1970). A general method applicable to the search for similarities in the amino-acid sequence of two proteins. *Journal of Molecular Biology*, 48, 443–453.

Piccarreta, R., &**Lior, O.** (2010). Exploring sequences: A graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society Series A*, 173, 165–184. doi:10.1111/j.1467-985X.2009.00606.x

Pollock, G., (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society Series A*, 170, 167–183. doi:10.1111/j.1467-985X.2006.00450.x

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

Shoval, N., & **Isaacson, M.** (2007). Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers*, 97, 282–297. doi:10.1111/j.1467-8306.2007.00536.x

Stovel, K., **Savage M.**, & **Bearman P.** (1996). Ascription into achievement: models of career systems at Lloyds bank, 1890–1970. *The American Journal of Sociology*, 102, 358–399. doi:10.1086/230950

Studer, M., (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers*, 24. doi:10.12682/lives.2296-1658.2013.24

Studer, M., **Ritschard, G.**, **Gabadinho, A.** & **Müller, N. S.** (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research*, 40, 471–510. doi:10.1177/0049124111415372

Wiggins, R. D., **Erzberger, C.**, **Hyde, M.**, **Higgs, P.**, & **Blane, D.** (2007). Optimal matching analysis using ideal types to describe the lifecourse: An illustration of how histories of work, partnerships and housing relate to quality of life in early old age. *International Journal of Social Research Methodology*, 10, 259–278. doi:10.1080/13645570701542025

Wu, L. (2000). Some comments on sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods and Research*, 29, 41–64. doi:10.1177/0049124100029001003