CHAPTER 2

DESCRIBING AND VISUALIZING SEQUENCES

Before turning to numerical and graphical sequence description, this chapter introduces some basic concepts and definitions that are widely used in the sequence analysis (SA) literature and throughout this book. Note that this overview is selective in that it only covers concepts that are deemed relevant for the SA tools considered in later chapters. We will not, for instance, introduce the terminology of Markov models. Yet we will discuss most concepts and definitions that are pertinent to methods often applied in tandem with SA (e.g., cluster analysis, multinomial logistic regression, and event history analysis) in the respective chapters.

2.1 Basic Concepts and Terminology

Sequences are ordered lists of a discrete set of elements. In most social science applications, sequences are temporally ordered, but SA can also be applied to "timeless" sequences such as preference orders, cognitive schemas, or spatial orders (Cornwell, 2015). The set of elements constituting a sequence is called a state space or alphabet A. Following the notation of Elzinga and Liefbroer (2007), we can define a sequence x of length k as $x = x_1, x_2, \ldots x_k$ with $x_i \in A$ and i indicating the position of a state within the sequence x.

2.1.1 Sequences With Recurrent States

Most social science applications relate to *recurrent sequences*, where the elements of the alphabet can occur repeatedly in each sequence. Table 2.1 illustrates such sequences by presenting partnership trajectories of length $k = 6$. In our illustrative example of partnership biographies, the alphabet consists of the following states: single (S), living apart together (LAT), cohabiting (COH), and married (MAR).

Table 2.1 Example of two sequences

	x_1	x_2	x_3	x_4	x_5	x_6
Sequence A	S	S	LAT	COH	COH	MAR
Sequence B	COH	MAR	MAR	MAR	COH	COH

2.1.2 Episodes and Transitions

In both sequences shown in Table 2.1, certain states appear multiple times. A series of consecutively repeated states, such as *(S, S)* and *(MAR, MAR, MAR)*, is called an episode or spell. Note that even states that appear only once constitute episodes. According to this definition, Sequence A consists of four and Sequence B of three episodes.

Given the short length of the two example sequences, it is easy to recognize the episodes at a glance and it is possible to display the sequences in a table. Longer sequences call for more condensed notation. Our pairfam example data are particularly suitable to illustrate this point. The data provide monthly accounts of family formation and labor market participation biographies covering ages 18 to 40 years. The following example uses information on the respondents' partnership biographies applying the alphabet introduced earlier. This gives us recurrent sequences with a maximum of four different states and a length of $k = 22 \times 12 = 264$ months. A typical sequence can be written as follows:

LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-S-S-S-S-S-S-LAT-LAT-LAT-LATLAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-MAR

It is immediately evident that this notation, which is termed the state-sequence (STS) format, makes it difficult to identify episodes within sequences. This can be easily solved by listing only one distinct state for every episode:

LAT-S-LAT-S-LAT-S-LAT-COH-MAR

According to Gabadinho, Ritschard, Müller, and Studer (2011), this type of sequence representation is called distinct-successive-states (DSS) sequence format. It is an accessible account of the episodes and maintains the order of the original state sequence, but it lacks any information on the duration of the observed episodes. If the researcher's interest is only in the order of states, analyzing DSS sequences alone will suffice. In most applications, however, researchers are interested in the duration of episodes and the timing of events, and therefore want to utilize the full information stored in their data. Hence, avoiding the somewhat lengthy STS format, Aassve and colleagues (2007) have suggested the much more parsimonious state-permanence-sequence (SPS) notation, which is similar to the DSS format but also provides information about the length of a sequence's episodes by simply adding the duration of each episode. Thus, our example sequence would be as follows:

(LAT,13)-(S,6)-(LAT,33)-(S,24)-(LAT,41)-(S,35)-(LAT,10)-(COH,14)-(MAR,88)

As in DSS notation, here, it is easy to recognize the distinct episodes. Further, this notation shows that this 22-year-long sequence is clearly dominated by a marriage spell that lasted more than 7 years (88 months). The sequence also comprises two rather long LAT spells in the respondent's early and mid-twenties, each lasting roughly 3 years.

In addition to showing the episodes—which can be understood as building blocks of sequences—DSS and SPS notation also refer to another important concept in SA, namely, transitions. The example sequence comprises nine episodes; this implies that it includes eight transitions. The transitions from being single to LAT and from LAT to single occur three times each, while we observe only one transition to cohabitation and marriage. It is plausible to assume that the other sequences in the data show a similar pattern—that is, that "single" or "LAT" constitute origin or destination states of transitions considerably more often than "cohabitation" or "marriage." Likewise, we can expect similarities in terms of transition timing, with transitions in and out of LAT relationships mainly occurring in the early and mid-twenties and transitions to marriage coming in the late twenties or early thirties. The frequency and timing of transitions are not only interesting for deriving substantive insights into the process under study (see Section 2.3) but also relevant for data-driven approaches of measuring the similarity of sequences (see Chapter 3, Section 3.3) or assessing sequence complexity (see Section 2.5).

2.1.3 Subsequences

In addition to episodes and transitions, subsequences are another important component of sequences. According to Elzinga and Liefbroer (2007), a subsequence u of sequence x is defined as an ordered list in which the elements of alphabet A, that is, the set of categorical states constituting the sequence, appear in the same order as in sequence x (i.e., $u \subseteq x$). All sequences share two specific subsequences: the empty subsequence λ and the original sequence x. As a result, sequence $x = A, B, C$ already has eight distinct subsequences ($\lambda; A; B; C; AB; AC; BC; ABC$), which can be written as $\phi(x) = 8$.

Within the SA toolkit, subsequences are used to measure the degree of sequence turbulence (see Section 2.5) and for pairwise sequence comparisons (see Chapter 3, Section 3.4). Generally speaking, the more subsequences a sequence has, the more complex it is, and the more subsequences two sequences share, the more similar they are.

A Note on Data Formats

Sequence data come in a variety of formats. The most common ones—the wide, long, and spell data format—are depicted in the following table. Software packages for analyzing social science sequence data differ in their ability to work with different data formats. TraMineR is the most versatile package and is well suited for working with sequence data stored in the episode, long, or wide format. Stata's SQ (Brzinsky-Fay et al., 2006) requires the data to be stored in the long format, and SADI (Halpin, 2017) only works for sequence data stored in the wide format. Both R and Stata allow the data to be reshaped if it is not in the desired format (see companion webpage material for Chapter 2).

(a) Long format

id	**obs.**	***x***
1	1	A
1	2	A
1	3	B
2	1	B
2	2	C
2	3	C

(b) Wide format

id	***x1***	***x2***	***x3***
1	A	A	B
2	B	C	C

(c) Episode/spell format

id	**episode**	***x***	**start**	**end**
1	1	A	1	2
1	2	B	3	3
2	1	B	1	1
2	2	C	2	3

2.2 Defining Sequences

As with all methods, the results of SA depend heavily on the researcher's analytical decisions. Being a rather new method, SA—and optimal matching in particular—has faced a lot of criticism since it was first introduced to the social sciences. Aisenbrey and Fasang (2010) have provided an excellent overview of the initial critique and subsequent methodological advances. Much of the early criticism was concerned with measuring similarity between sequences. Although this is indeed a critical step in many SA applications, the first analytical choices must already have been made when the data are actually being defined as sequences. More specifically, the researcher has to decide on the states that should be included in the alphabet, the point at which the sequences should begin and end, and the intervals at which the states should be measured; it is also necessary to determine how to deal with gaps, missing data, and sequences of unequal length. Unfortunately, there are no recipes for all of these analytical steps. Instead, the process is iterative and should be mainly guided by theoretical and substantive considerations. As a general heuristic, however, we recommend starting with a fine-grained analysis before turning to techniques that reduce the complexity in the data, such as collapsing rare states of the alphabet or aggregating monthly data into yearly data. In the following subsections, we discuss each of these analytical choices.

2.2.1 The Alphabet

The first analytical choice virtually always entails defining the alphabet (a.k.a. state space), that is, the set of categorical states that constitute the sequences. This step is largely determined by the data available to the researcher. In our example data on partnership biographies, for instance, the coding scheme allows us to distinguish between persons without a partner and persons who are in a partnership but are not cohabiting (LAT). Depending on the research question, this might be an important distinction, but it can rarely be made because many data sources only provide information on coresidential partnerships (Raab & Struffolino, 2019). Our data would have allowed for an even more extended alphabet that not only distinguished between different partnership states but also considered the partner with whom each state was experienced. Depending on the number of considered partnerships—the maximum number of partners *p* observed in pairfam is 14—this would result in a very extended alphabet (S, LAT*p*1, LAT*p*2, . . ., LAT*p14*, COH*p*1, . . ., COH*p14*, MAR*p*1, . . ., MAR*p14*) with sparsely occupied cells and arguably little potential for producing additional insights. This illustrates that a more nuanced alphabet is not necessarily conducive to a better analysis unless the research question calls for such a fine-grained specification.

Therefore, the main goal of the alphabet specification stage is to properly balance parsimony and detail. Many applications achieve this goal by collapsing categories if they occur rarely in the data or if they are considered irrelevant for theoretical reasons. In general, we consider this a reasonable strategy, because one of SA's main goals is to reduce complexity in order to enable the identification of relevant empirical regularities. Further, this strategy also acknowledges the limited capacity of human working memory, which makes it demanding and tedious to cognitively process large alphabets. Finally, the emphasis in SA on data visualization also favors small alphabets, because it becomes very difficult to come up with print-friendly qualitative color palettes that comprise more than nine states.

That said, we call for caution when lumping states together. While collapsing states ensures that the analyst does not lose sight of the most salient patterns hidden in complex data, it may obscure meaningful regularities that point to important minorities or deviant subgroups that warrant further consideration. We therefore recommend starting with a comprehensive state space and then testing whether this can be sensibly reduced based on substantive considerations regarding similarities between states. This implies an iterative process of sequence definition and further analysis, such as optimal matching and cluster analysis. Of course, this procedure should be guided by theoretical and substantive considerations.

If SA is applied to secondary data, the researcher's freedom to choose between different alphabet specifications is often extremely restricted and the differences in the results derived from diverse alphabets tend to be rather modest. The surprisingly weak impact of alphabet specification has even been shown with anthropological sequence data on figures in ritual dances, which leaves room for very divergent definitions of the coding scheme (Forrest & Abbott, 1990). The robustness of the results to changes in the alphabet can be assessed by inspecting the degree of resemblance of the dissimilarity matrices obtained by analyzing differently defined sequences. This can be achieved by inspecting the correlation of the matrices by calculating the Mantel coefficient or applying permutation tests for the similarity of matrices (Forrest & Abbott, 1990; Piccarreta, 2017; Piccarreta & Elzinga, 2013). We will illustrate these techniques in Chapter 5.

2.2.2 Sequence Length and Granularity

In previous sections, we broadly introduced sequences as ordered lists of (categorical) elements. In most applications, these lists are temporally ordered, which implies that sequences are defined with reference to some sort of time axis. Building on a classification suggested in the literature on event history analysis (Blossfeld & Rohwer, 2001), we distinguish between a calendar time axis or a process time axis. In the first setting, the beginning and end of the sequences are defined by a fixed time point, such as a specific year, chosen by the researcher. In the second scenario, the start of the sequence is defined by the occurrence of a specific event, such as a transition like leaving school or a specific birthday. Employment trajectories, for instance, are often set to begin once the respondents leave the school (Brzinsky-Fay, 2007; Struffolino, 2019). According to this definition, the sequences for some respondents might start at age 16 while for others they may begin 2 or 3 years later. This, of course, has important repercussions for any subsequent analysis and requires a constant awareness on the part of the analyst of the age differences of the analyzed sample. In many applications, the process time is defined by referring to a transition that marks the beginning of the observation period. In the second step, the analyst decides on the length of the respective observation period, such as the first 10 years after leaving school. Of course, sequences can also be defined with reference to their end date, such as entry into retirement or death (Raab et al., 2018).

The definition of the start and end dates and the granularity (measurement accuracy of the process time; e.g., measured in years, months, weeks, or days) of the available data determine the length of the sequences. For instance, the example sequences used in this book cover a period of 22 years (process time from age 18 to 40). Given that our data provide monthly information on status changes during this observation period, our sequences are of length $k = 22 \times 12 = 264$. Compared to most published social science studies using SA, these are rather long sequences. Even with our small alphabet of four different states, the data would allow for an overwhelming number of possible sequence realizations (i.e., $4^{264} \approx 8.79 \times 10^{185}$). Although researchers will observe only a fraction of this sequence universe, this data setting means they are very unlikely to observe the same sequence multiple times. Indeed, our example data of 1,866 cases contains 1,834 unique sequence realizations for partnership biographies.

This illustrates nicely how long sequences allow for extensive heterogeneity. As was also true when specifying large alphabets, the heterogeneity of long sequences is neither bad nor good when seeking to conduct a sound analysis. It is rather a question of what the analyst is interested in. Most SA applications aim at reducing complexity by searching for patterns. Accordingly, any analytical decisions made when specifying the sequences should be evaluated in light of this goal and the substantive question at hand. Some research questions call for nuance in order to identify rare regularities. In most applications, however, the goal is to uncover the most salient patterns hidden in the data. If this applies, the researcher should consider reducing the complexity when defining the sequences.

Turning back to our example, one could argue that it is not relevant to capture every single fleeting affair and that researchers should instead focus on serious relationships that last longer than a couple of months. In general, there are two approaches to reducing complexity in such a scenario:

1. *Data manipulation by recoding sequence states.* This data manipulation procedure should be based on a set of rules that is clearly communicated by the researcher. In our example, this strategy could be enacted by imposing a threshold rule that defines a minimum length for partnership spells. If a spell falls below this threshold value, it is discounted and the respective states coded as being single rather than in a relationship. Applying a threshold of 12 months as a minimum relationship duration only slightly reduces the number of distinct sequences from 1,834 to 1,787, but still leads to somewhat less complex sequences. Note that this recoding strategy is different from the one discussed in the previous section insofar as it does not alter the size of the alphabet.
2. *Reduction of the sequence length by aggregation*. Unlike the first approach, this strategy entails both a substantive and a structural change in the sequence data, because it reduces sequence length. Executing this technique requires a rule specifying how the data should be aggregated. A predefined number of adjacent states (e.g., 12 months) can be summarized according to the first, the last, or the most frequently observed state. In the SA literature, the process of aggregating successive positions is sometimes referred to as changing the time granularity of sequences. Applying this strategy to our example data, we aggregated monthly to yearly data using each year's modal value. This reduced the sequence length from 264 to 22 and the number of distinct sequences from 1,834 to 1,432, although the observation window still covered the same time period as before.

In simplifying the original data, neither approach fully utilizes the available information; they should thus be applied with some caution. In general, the second strategy is more invasive and has a greater impact on further analysis. It also reduces the computing load and is therefore particularly suitable for large data sets and lengthy sequences with complex alphabets. Table 2.2 illustrates how the two approaches modify the original data for a subset of our partnership sequences.

Table 2.2 Comparison of different approaches toward defining sequence data

Initial sequence (states affected by the data reduction techniques printed in boldface)	**Strategy 1: recode** (only considers partnerships lasting at least 1 year)	**Strategy 2: aggregate** (monthly to yearly data using modal values)
*(S,89)-(LAT,26)-(COH,14)-**(LAT,6)-**(S,34)-**(LAT,4)-**(MAR,91)*	*(S,89)-(LAT,26)-(COH,14)-(S,44)-(MAR,91)*	*(S,7)-(LAT,3)-(COH,1)-(S,3)-(MAR,8)*
(LAT,13)-(S,6)-(LAT,33)-(S,24)-(LAT,41)-(S,35)-(LAT,10)-(COH,14)-(MAR,88)	*(LAT,13)-(S,6)-(LAT,33)-(S,24)-(LAT,41)-(S,45)-(COH,14)-(MAR,88)*	*(LAT,1)-(S,1)-(LAT,2)-(S,2)-(LAT,4)-(S,3)-(LAT,1)-(COH,1)-(MAR,7)*
(S,56)-(LAT,69)-(COH,47)-(MAR,92)	*(S,56)-(LAT,69)-(COH,47)-(MAR,92)*	*(S,5)-(LAT,5)-(COH,4)-(MAR,8)*
***(LAT,4)-**(S,134)-**(LAT,9)-(COH,3)-**(MAR,52)-**(LAT,5)-**(COH,25)-(MAR,32)*	*(S,150)-(MAR,52)-(S,5)-(COH,25)-(MAR,32)*	*(S,12)-(MAR,5)-(COH,2)-(MAR,3)*

Because it changes the granularity of the data, the second approach has greater potential to reduce the complexity of the original data. Apart from the sequence depicted in the last row of Table 2.2, however, the results obtained by the two strategies are surprisingly similar, although the sequence data structure of the approaches varies considerably ($k = 264$ vs. $k = 22$). Both approaches tend to reduce the number of LAT spells, because these relationships are characterized by short durations. Otherwise, the manipulated data are pretty similar to the original data and the changes are much less severe than might be expected. To give a comparison, changing the alphabet by recoding LAT relationships to being single would affect the sequences much more.

2.2.3 Sequences of Unequal Length: Censoring and Missing Data

The examples considered up to this point only include sequences of equal length. In a few cases, the fact that some realizations of the same process are of different durations is an empirical finding that can be relevant to specific research questions: For example, when looking at the timing of democratization, and accounting for prior political regimes' histories, one of the outcomes of interest may be the very fact that countries differ with regard to the length of the democratization process (Wilson, 2014). Another example is coming from the survey methodology literature, which is increasingly analyzing paradata capturing information about field processes. These data comprise information on contact histories of unequal length summarizing the number, timing, and outcome of contact attempts. Recent applications have analyzed these process data using SA to gain valuable insight for improving survey monitoring and survey management (Kreuter & Kohler, 2009).

In the vast majority of the cases, however, the SA literature consists of studies that use data in which each unit of analysis is observed for the same period and the sequences do not have any missing value. This is not because the data are free of missing values; rather, it points to the absence of well-established ways of working with missing or censored data in SA (Cornwell, 2015; Piccarreta & Studer, 2018). For this reason, most studies do one of the following: They replace gaps with valid values using some sort of imputation, add an additional missing state to the alphabet, or delete cases with missing data.

In general, there are two sources of missing values in sequences: censoring and gaps. Sequences are censored if not every unit of analysis was observed for the same time period. This is a typical scenario for panel studies, in which respondents often enter and leave the survey at different time points. While censoring pertains to the boundaries of sequences, missing values also manifest as gaps surrounded by valid information. In surveys, this is usually the result of temporary item or unit nonresponse. Simply abandoning the missing states would result in sequences of unequal length. Although the SA toolkit provides techniques for normalizing sequence distances, this is not a viable strategy for dealing with sequences of unequal length (Elzinga & Studer, 2019). A strategy of replacing missing values with a dedicated missing state in the alphabet also cannot be recommended. Although this produces sequences of equal length, it hampers subsequent analysis. In pairwise sequence comparisons, for instance, sequences sharing many missing states would be considered similar, although the missing category is only a placeholder for all states of the alphabet (Piccarreta & Studer, 2018).

Halpin (2016a) has proposed a technical solution to this problem. Instead of treating missing states as identical, he introduced the notion of "non-self-identical" missing values. According to this approach, sequences that share many missing values are regarded as being dissimilar to each other, which is much more plausible in most applications than the assumption that they are similar.

Imputation techniques are another promising approach to handling missing values. One common method for dealing with gaps can be applied in the data manipulation stage of the analysis. Here, the researcher simply defines a set of rules to close gaps in the data. In our partnership biographies, for instance, if there was a gap of 5 months surrounded by two single episodes, it could be closed by imputing the state single. If there was a gap surrounded by different states, both states could be used to close the gap in a symmetric fashion. Yet, this strategy might result in underestimating the true volatility of sequences by simply replacing the missing states with the surrounding states, although the gap could actually point to one or multiple episodes spent in a different state. From a statistical point of view, this kind of data imputation is based on strong assumptions as it does not reflect the uncertainty about the imputed values. It simply assumes that the real values are known.

A statistically correct alternative requires multiple imputations for each missing observation. Multiple imputation techniques are widely available in the most commonly used statistical software packages but mainly refer to the imputation of cross-sectional data. Halpin (2016b) has proposed a tweaked imputation procedure specifically tailored for use with categorical time series data. The algorithm fills the elements of the gap successively by using the first valid values surrounding a missing state. In this stepwise procedure, the technique uses values that have already been imputed to predict the remaining missing values. Halpin has provided a dedicated software package for Stata, named MICT, that facilitates the application of his ideas. He has also tested how his imputation procedure compares to the alternative imputation procedures in a simulation scenario with a random missing pattern (missing at random = MAR). The results obtained by MICT tend to perform better in retaining the "longitudinal consistency" of the sequence data than traditional multiple imputation, multiple imputation by chained equations, or the more recently proposed two-fold fully conditional specification (Nevalainen et al., 2009) for imputing missing values in longitudinal data. MICT, however, is currently unable to include additional variables with missing values in the imputation model. This is a major drawback because standard imputation techniques call for a joint imputation of all missing values of the variables used in the analysis. Given that SA is virtually always applied in tandem with other multivariable methods, the occurrence of additional missing values is very likely, and analysts must rely on more established imputation procedures, despite the fact that MICT is arguably better suited for sequence data.

An additional problem arises if the dissimilarity matrices retrieved by analyzing multiple imputed sequence data should be examined further by some sort of cluster analysis. Contrary to regression analysis, multiple imputation is rarely applied in the context of cluster analysis, and the literature still lacks well-established guidelines on how to proceed in such a case (see Basagaña et al., 2013, for some recommendations). In sum, the handling of missing data in SA remains a problem that warrants further research. Given the lack of clear guidelines, most SA applications still pertain to the analysis of completely observed sequences and try to close gaps in the data manipulation stage of the project rather than in the context of a statistically sound imputation procedure.

2.3 Description of Sequence Data I: The Basics

A sound sequence analysis is based on a good descriptive understanding of the analyzed data. The basic description of sequence data does not require specific techniques as it mainly provides counts and averages of different sequence properties.

2.3.1 Time Spent In Different States and Occurrence of Episodes

An overview of the average time spent in the different states of the alphabet as well as the average number of state-specific episodes is a good starting point for describing the data. Using the monthly example data on partnership trajectories, Table 2.3 shows that the respondents in the sample spend 36% of the time in marriages (95 of 264 months), while the corresponding figures for time spent in LAT relationships or coresidential unions are only half as high. Time spent outside relationships ranks between these two extremes. Although the respondents spent the shortest amount of time in LAT relationships, the corresponding number of episodes is (slightly) higher than for the other states. People experience more than twice as many LAT spells than marriage episodes (1.8 vs. 0.8) but spend only half the time in these partnerships, suggesting comparatively short average durations of approximately 2 years per LAT relationship.

2.3.2 Transition Rates

Once the average number of distinct episodes is known, the obvious next step is the examination of transitions between episodes and states. In the monthly sequence data, the average number of transitions is 4.3, whereas the same figure for the yearly data equals 3.3.

Moving beyond simple averages, transition rates between consecutive time points can provide even more insight into how the sequences unfold. In most social science applications, however, sequences are characterized by a considerable amount of stability. This means that most individuals do

Table 2.3 Average time spent in different states and number of spells

State	*Time spent in state x in months*			*Number of episodes*	
	Mean	**SD**	**Rel. freq.**	**Mean**	**SD**
S	72.5	69.8	0.27	1.6	1.2
LAT	48.0	43.9	0.18	1.8	1.3
COH	48.6	53.3	0.18	1.0	0.8
MAR	95.0	78.9	0.36	0.8	0.5

not change their status between two observations of a categorical time series. Using sequence data of a finer granularity, such as monthly versus yearly data, further contributes to a low share of transitions between states. As a result, transition matrices for sequence data with recurrent states often are not very informative apart from demonstrating the salience of the concept of path dependency.

For instance, the transition rates of both the monthly and yearly sequence data depicted in Table 2.4 show that only a (small) minority of people change their partnership status between two consecutive observations. That being said, the table also illustrates that coding the sequences with a yearly granularity produces slightly more interesting results. While for the monthly data, virtually all transitions are recorded on the main diagonal—indicating stability in partnership status from one month to the next—the yearly data reveal more transitions between consecutive observations. Particularly in the case of LAT relationships, the transition matrix corroborates earlier findings on the rather temporary nature of this type of partnership by demonstrating that 32% of persons in LAT relationships are observed in a different status 1 year later. While 12% of these LAT relationships end in separation, 20% are further institutionalized, either as coresidential unions (16%) or marriages (4%).

Examining the transition rates of sequences stored in the DSS format can provide further insights by removing recurrent appearances of the same state in consecutive positions of the sequence. A transition matrix based on this format provides transition rates between episodes of distinct states. Accordingly, the cells on the diagonal equal zero. By definition, this approach yields higher transition rates between different states compared to the previous procedure, even with data of fine granularity.

The transition matrix for the monthly DSS sequence data shown in Table 2.5 illustrates this point well. Interestingly, this matrix shows that more than half

Table 2.4 Transition matrix of sequences stored in STS format

State at *t*	***State at t + 1***							
	Monthly granularity				***Yearly granularity***			
	S	LAT	COH	MAR	S	LAT	COH	MAR
S	0.98	0.02	0.00	0.00	0.81	0.14	0.04	0.01
LAT	0.02	0.96	0.02	0.00	0.12	0.68	0.16	0.04
COH	0.00	0.00	0.98	0.01	0.04	0.02	0.80	0.14
MAR	0.00	0.00	0.00	1.00	0.01	0.01	0.00	0.98

Table 2.5 Transition matrix of sequences stored in DSS format

State at *t*	***State at t + 1***			
	S	LAT	COH	MAR
S	0.00	0.91	0.07	0.02
LAT	0.42	0.00	0.50	0.08
COH	0.20	0.12	0.00	0.68
MAR	0.44	0.46	0.11	0.00

(57%) of those whose marriages ended swiftly moved on to the next relationship without an intermediate single episode. When interpreting these results, one should be aware that the analysis of DSS transition matrices only examines *movers*, because all *stayers* are removed from the data.

2.3.3 State Distribution and Shannon Entropy At Different Positions

Although the tools introduced earlier build a strong descriptive foundation, they are limited in the sense that they only produce aggregate indicators summarizing entire sequences without providing information on how these sequences develop across time.[1] The inspection of the state

[1] Although it would be possible to calculate transition matrices or other measures at different positions of the sequences, we do not consider this a very promising approach. The transition rates at different positions, for instance, are usually very low if the sequence length is not very small. In addition, and particularly in the case of long sequences, the transition rates at a specific position—such as from month 126 to 127—are not very meaningful from a substantive point of view.

Table 2.6 State distribution at selected positions

State	***Age***						
	18	20	24	28	32	36	40
S	0.65	0.52	0.36	0.25	0.18	0.15	0.14
LAT	0.31	0.32	0.25	0.17	0.12	0.10	0.05
COH	0.03	0.11	0.23	0.25	0.22	0.15	0.13
MAR	0.01	0.05	0.17	0.33	0.48	0.60	0.68

distribution at different positions of the sequence addresses this limitation. Table 2.6 displays the distribution of partnership states at ages 18, 20, 24, 28, 32, 36, and 40.

The table reveals that until their late 20s, most respondents report that they are currently without a partner. While singles remain a notable minority until the end of the sequences, marriage becomes the modal, or most prevalent, state at this age, and two thirds of the respondents indicate that they are married at age 40.

Table 2.7 complements these figures by describing the heterogeneity of the respective distributions using Shannon entropy. The entropy index S is defined as

$$S = -\sum_{i=1}^{a} p_i \times \log\left(p_i\right)$$

where p_i denotes the proportion of cases in state i and a is the size of the alphabet (Gabadinho, Ritschard, Müller, & Studer, 2011). At a given position of the sequence, S equals 1 when each state of the alphabet is observed equally often and 0 when only one state is observed.

Table 2.7 Shannon entropy at selected positions

	Age						
	18	20	24	28	32	36	40
Entropy	0.58	0.78	0.97	0.98	0.9	0.8	0.69

For our example data, the inversely u-shaped pattern of the entropy values substantiates the results reported in Table 2.7. The heterogeneity in the state distribution is lowest at age 18 and increases until the respondents are in their late 20s before it decreases again.

Before turning to the visualization of sequences, a word of caution regarding the interpretation of state distributions is in order. State distributions depict aggregated cross-sectional snapshots at different positions of the sequence. They do not convey any direct information on how individual sequences unfold across time. In the monthly partnership sequences, for instance, an inspection of the state distributions across all positions of the sequence indicates that the share of singles never falls below 13%. A thorough longitudinal inspection of the individual biographies, however, reveals that the share of persons who never experienced a partnership episode is less than 1%.

2.3.4 Modal and Representative Sequences

Some SA applications draw on the state distribution to report the modal sequence, that is, a sequence composed of the most prevalent states at each position of the sequence. In accordance with the results reported earlier, this sequence comprises only two of the four partnership states. Written in SPS notation, the sequence of modal states derived from the monthly state distributions is *(S,102)-(MAR,162)*. Note that the modal sequence is of limited use because it is virtually always a hypothetical sequence based on aggregated cross-sectional state distributions that is not actually observed in the data.

Therefore, we recommend other approaches for identifying those sequences that best represent the data. These techniques are based on a matrix measuring the pairwise dissimilarities between sequences. Chapter 3 introduces several different approaches for computing such dissimilarity matrices. Since there are a number of different takes on measuring sequence dissimilarities, the identification of a set of representative sequences hinges on the chosen dissimilarity measure. For now, however, it suffices to understand that a dissimilarity matrix allows differentiating sequences that are (very) dissimilar to the remaining sequences in the data from those that are more central, that is, less dissimilar to the other sequences.

After obtaining the dissimilarity matrix, the identification of a set of representative sequences requires additional analytical choices. The analyst has to decide how many representative sequences should be extracted from the universe of observed sequences as well as which representativeness criterion should be used for the extraction. Based on these decisions, the algorithm identifies a subset of nonredundant sequences in an iterative procedure.

Table 2.8 displays a set of representative sequences using the example data on yearly partnership biographies. These sequences were extracted by applying the neighborhood density criterion. According to this criterion,

Table 2.8 Set of representative sequences

Sequence	Coverage	Assigned
(S,1)-(LAT,2)-(MAR,19)	5.7	6.5
(S,20)-(MAR,2)	4.4	25.2
(S,4)-(LAT,1)-(COH,1)-(MAR,16)	3.8	5.3
(LAT,3)-(COH,2)-(MAR,17)	3.1	11.4
(S,2)-(LAT,2)-(COH,3)-(MAR,15)	2.7	17.1
(S,5)-(LAT,2)-(COH,2)-(MAR,13)	2.7	23.5
(COH,2)-(MAR,20)	2.6	3.0
(S,1)-(LAT,5)-(MAR,16)	2.3	8.0
Total Coverage	27.5	100.0

sequences are considered neighbors if their pairwise dissimilarity falls below a predefined threshold value—in this case, 10% of the maximum possible distance value. The share of neighboring sequences indicates the coverage of a sequence. We aimed at identifying a subset of nonredundant representative sequences with a total coverage of at least 25%. This criterion was met after the extraction of eight sequences.[2] The coverage of the first sequence depicted in Table 2.6 amounts to 5.7%. This means that every 20th person in our data set is described accurately by a sequence comprising a very short period without a partner *(S,1)* followed by a short LAT spell *(LAT,2)*, which is converted into a stable marriage *(MAR,19)* already in the early 20s. Given that marriage was the modal state in the second half of the sequences, it comes as no surprise that all representative sequences end in marriages. That said, the table reveals a considerable level of variability in the pathways leading to this predominant destination state. Contrary to the first sequence, for instance, the second representative sequence is characterized by a very long single spell *(S,20)*.

[2] Note that the representative sequences in Table 2.8 are sorted by coverage, whereas the algorithm extracted the sequences in a different order. As a result, the representative extracted last had a coverage of 2.7%, leading to a total coverage exceeding the threshold of 25% by 2.7 percentage points.

Note that the algorithm assigns every sequence in the data set to its closest representative even if the respective sequence does not meet the predefined representativeness criterion. As a result, the share of sequences assigned to a specific representative usually exceeds the share of sequences that are covered by it (see Table 2.8). High discrepancies between the two values often suggest that many of the assigned sequences are not characterized that well by their representative. TraMineR provides more sophisticated quality measurements than this simple visual inspection. For a detailed discussion of the identification and assessment of representative sequences, see the comprehensive introduction by Gabadinho, Ritschard, Studer, and Müller (2011).

This section clearly demonstrated that representative sequences are a more complex but also a much more useful tool for exploring and summarizing a large body of sequences compared to presenting a simple sequence of modal states. While we used this tool for tabular description, most applications use it to graphically illustrate sequence data—particularly if the data comprise too many cases to plot all sequences without visual artifacts (see Fasang & Liao, 2014, and Section 2.4.2 on relative frequency sequence plots in this chapter). Moreover, representative sequences are not usually extracted for the full data set but only for different subgroups, for instance, groups obtained after OM and cluster analysis or substantively interesting subpopulations, such as persons with different educational attainment or ethnic origin.

In doing so, many applications report only one single representative per group, the so-called *medoid sequence*. The medoid is defined as the most central object of a given subset of sequences, that is, it has the lowest sum of dissimilarities to all other sequences. Comparing the medoid partnership trajectory of women—*(S,3)-(LAT,2)-(COH,4)-(MAR,13)*; coverage = 3.4—to men's medoid sequences—*(S,7)-(LAT,4)-(COH,3)-(MAR,8)*; coverage = 1—shows that the most notable differences occur at the beginning and the end of the observational window. Compared to the "medoid woman," the "medoid man" remains single for 4 additional years during early adulthood and marries 5 years later.

2.4 Visualization of Sequences

The tabular inspection of sequences conveys a lot of information that can easily become hard to interpret for both the analyst and the recipients. This is due to the high level of complexity of sequence data, which arises from the categorical level of the measurement of sequences and from repeated measurements per unit of analysis. Contrary to cross-sectional numerical data, sequences cannot be satisfactorily summarized by presenting the mean or the standard deviation of a single numerical indicator such as income. Instead, a thorough description requires the exploration of the distribution of multiple categorical states at different positions of the sequence (see Tables 2.6 and 2.7) as well as the inspection of sequence-specific composite measures. As a result, the tabular description of sequence data runs the risk of producing an overwhelming number of figures that, despite their accuracy, hamper the recognition of regularities in which the researcher is actually interested.

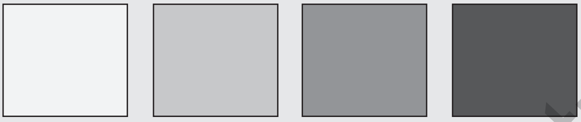
In view of this limitation, this chapter explores how graphical tools can complement the tabular description and demonstrates that they are frequently capable of communicating the same level of information in a more efficient and effective manner (Healy & Moody, 2014; Tufte, 1983). Visualization has played a prominent role in the SA literature since the proliferation of dedicated software packages for Stata and R, and recent contributions constitute notable additions to the visualization toolkit (Bürgin & Ritschard, 2014; Fasang & Liao, 2014; Piccarreta, 2012). This section presents a small selection of those tools that we consider most useful and that figure most prominently in the applied SA literature. We mainly draw on the excellent overviews provided by Brzinsky-Fay (2014) and Fasang and Liao (2014), to which we also refer for more detailed coverage of sequence visualization.

Following Fasang and Liao (2014), we distinguish two groups of graphs for sequence visualization. *Data summarization graphs* (Section 2.4.1) aggregate and summarize the information stored in the sequences. These graphs visualize one or two dimensions of information stored in sequence data (Brzinsky-Fay, 2014) by presenting the different categories of the alphabet and providing some information on the (temporal) order of the observed states. *Data representation graphs* (Section 2.4.2) add a third layer of information by plotting individual sequences rather than only aggregated summary measures. As a result, they are richer in information but also more demanding for the viewer.

Using colors to depict the different states of the alphabet can simplify the interpretation of complex sequence graphs considerably. Although the increase in electronic publishing in recent years has contributed to the widespread use of colored figures, printing costs sometimes still prohibit the use of color in print outlets. Due to this common restriction, we illustrate how one might visualize sequence data in grayscale. While it is possible to produce most of the figures in grayscale, we recommend using color figures whenever possible. For guidance on choosing appropriate colors, see the contributions by Zeileis and colleagues (Zeileis et al., 2009; Zeileis et al., 2019). The companion website at **https://sa-book.github.io** provides further details on using predefined and optimized color palettes for visualizing sequences using HCL color palettes.

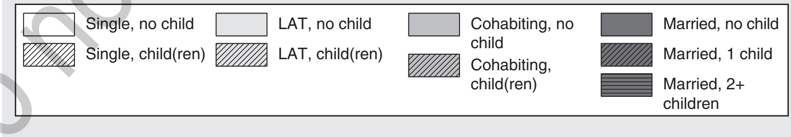
A Note on Grayscale Figures

Visualizing sequences in grayscale rather than in different colors restricts the options available to the analyst. Palettes of gray—like the one depicted in the following figure—are sequential palettes.



That is, they suggest that the data convey some sort of ordinal information. SA applications, however, often work with categorical alphabets, and the notion of a hierarchy between states might result in distorted visualizations.

Although the alphabet of partnership states analyzed in this book is also categorical, one could impose a hierarchy reflecting the partnership's degree of institutionalization (Single < LAT < Cohabitation < Marriage). Accordingly, the usage of a sequential palette of grays does not present much of a problem for this specific application. This is also true because the alphabet comprises only four different states. If the number of categories exceeds this level, gray palettes run the risk of becoming unsuitable for printing. Yet in some applications, the addition of shading lines might be a viable option for producing high-quality grayscale visualizations when facing larger alphabets. The following example and figure are a good illustration of this approach. The initial alphabet of four partnership states is extended by adding information on the parental status of the respondents. For each partnership state, we distinguish between childless persons and parents. For married persons, we add even more nuance by differentiating parents with one child from parents with multiple children. This results in an alphabet of nine different states, which can be visualized by four "colors" (white and three shades of gray) depicting the partnership states and by shading lines indicating the parental status.



The companion website features detailed instructions on how to produce such a grayscale color palette in R.

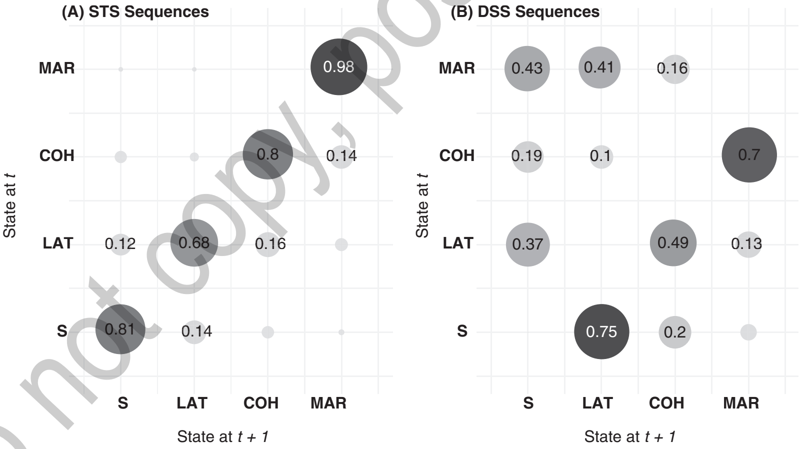
The following sections complement Section 2.3 by presenting graphs corresponding to the tables presented earlier. If not indicated differently, the plots are based on the monthly partnership data. Except for the transition plot, which was made using the ggplot2 package (Wickham, 2016), all figures were generated with the TraMineR package's visualization functions.

2.4.1 Data Summarization Graphs

Transition Plot

The transition plots in Figure 2.1 are based on transition matrices of yearly partnership data (see Table 2.4) using sequences stored in two different formats. The left panel is based on STS sequence data and therefore is dominated by high values on the main diagonal that indicate that even in modern societies, people rarely change their partnership status on a monthly basis. The right panel is based on DSS sequence data that do not allow for subsequent repetitions of the same state. Accordingly, the plot visualizes how the outflow transitions are distributed among the four partnership states.

Figure 2.1 Two transition plots of yearly partnership sequences



The size of the circles and the intensity of the gray shading increase with the prevalence of transitions between the two depicted states. If the share of outflow transitions in a given row reaches a threshold of 10%, the circles are labeled accordingly. Compared to the tabular presentation, this kind of plot emphasizes the most prevalent transitions. Note, however, that neither the transition matrices nor the transition plots provide any information on how many persons experience each transition. Similarly, it remains unclear when the transitions are taking place because the reported transition rates are averaged across all positions of the sequences. Screen presentations allow the researcher to address this issue by rendering transition plots at different positions of the sequences and displaying them sequentially, for example, as an animated GIF file (see companion website for an illustration).

Modal State Plot

In Section 2.3, we identified the modal sequence *(S,102)-(MAR,162)* for the entire sample of the example data. Without further information, the modal sequence contributes little to the understanding of the underlying data. Therefore, we strongly recommend adding a graphical depiction of the relative frequencies of the modal states at each position of the sequence.

Figure 2.2 demonstrates the added value of such an illustration. The figure not only displays the modal sequences for men and women, it also shows how the numerical dominance of the modal state(s) varies across sequence positions. For instance, in addition to showing that marriage is the modal state for women and men in their 30s, the figure also reveals that the dominance of this modal state is much more pronounced for women. Two thirds of the women are already married in their mid-30s, whereas the corresponding numbers for men never reach such a high level. Moreover, the plots show that the modal states at the beginning and the end of the sequences are more dominant, whereas the phase in between is characterized by more volatility—particularly for women.

State Distribution Plot

The state distribution plot (for an early application, see Blossfeld, 1987) represents the natural extension of the modal state plot. Technically speaking, it visualizes the distribution of all states by plotting a series of stacked bar charts at each position of the sequence. Figure 2.3 represents two state distribution plots using sequence data on family formation with yearly and monthly granularity ($k = 22$ vs. $k = 264$). Compared to the modal plot, Figure 2.3 provides additional insights by also displaying the distribution of the nonmodal states and using an enlarged alphabet that incorporates information on the parenthood status. At age 40, for instance, roughly half of the nonmarrieds are already parents.

Figure 2.2 Modal state plots for women and men

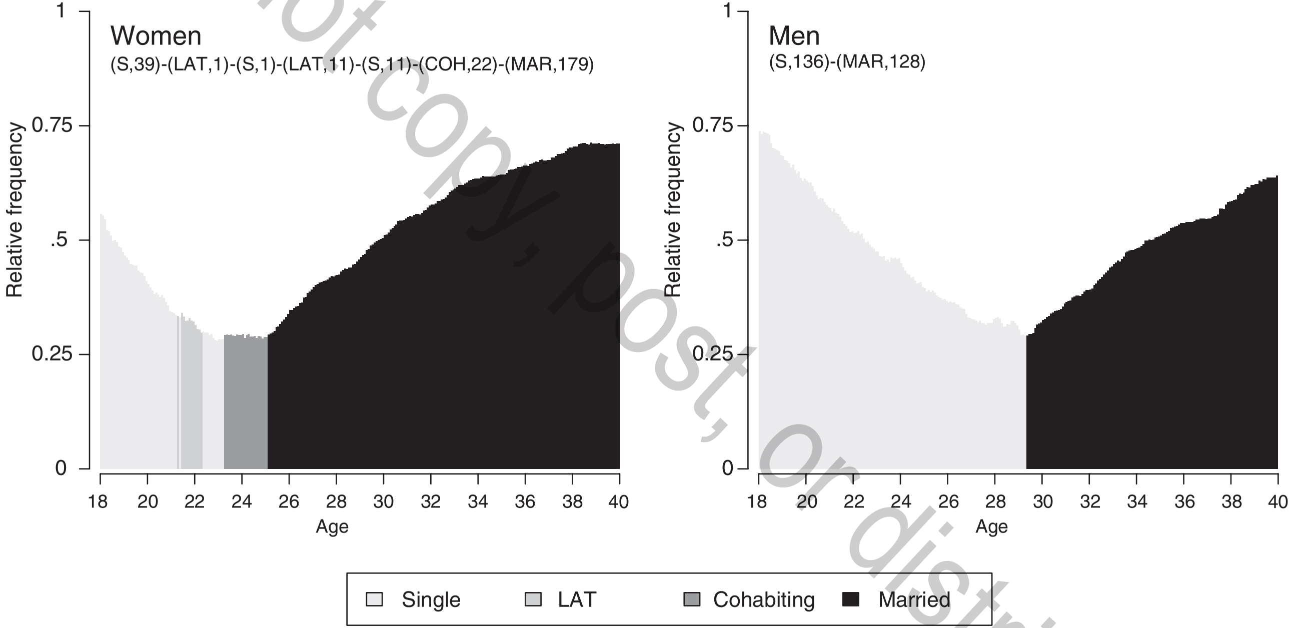
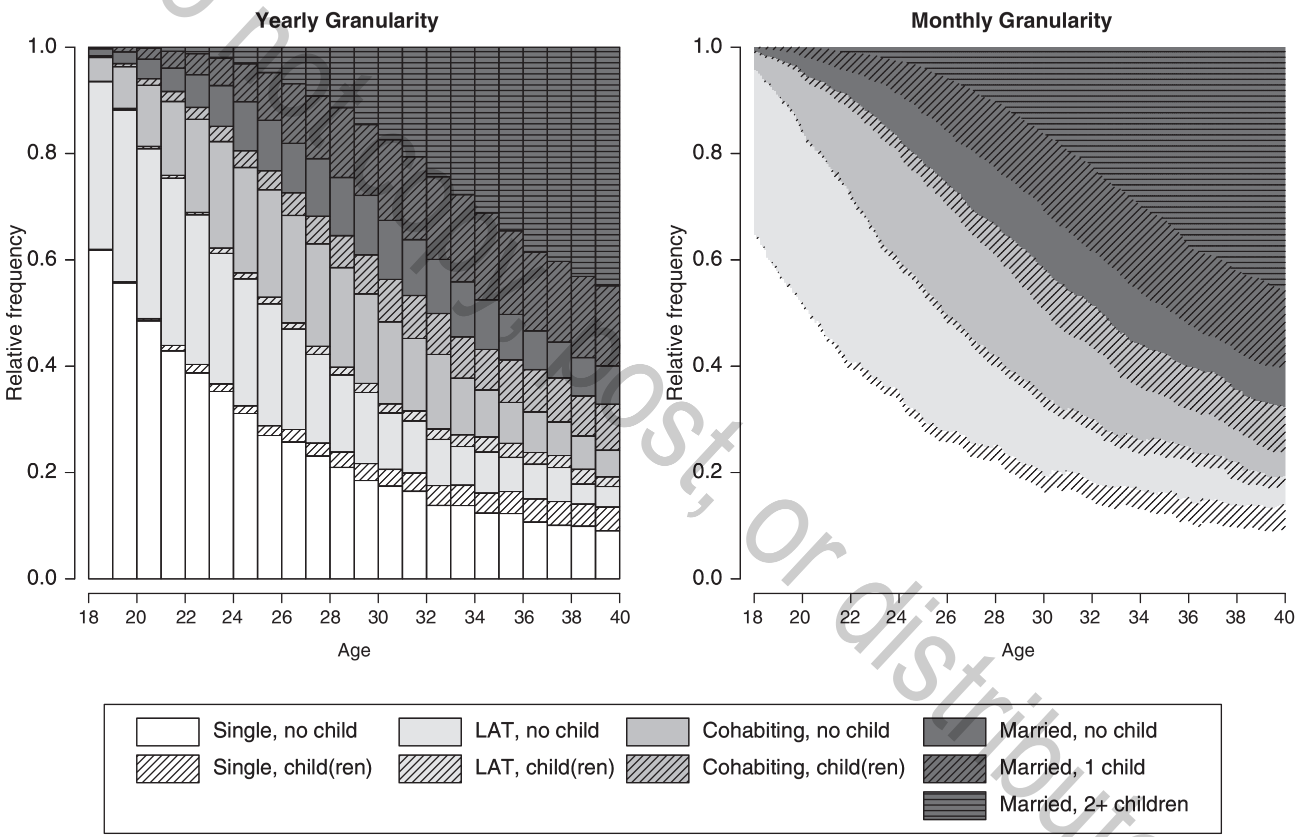


Figure 2.3 State distribution plots with different levels of granularity



Although the aggregated yearly data are less accurate, they do not change the interpretation of the results. Based on the similarity of results and the fact that the yearly data are less demanding from a computational point of view, this figure can be interpreted to support using the yearly data for further analyses. That said, we recommend such an approach without reservations only if hardware restrictions make this necessary.

Concluding the section on summarization graphs, Figure 2.4 shows a slightly enhanced version of the standard state distribution plot that also includes information on the state entropy at the different positions of the sequences. Note, however, that we only recommend this type of visualization for displaying sequences with small alphabets; hence, the figure is focusing on the partnership trajectories and neglects the information on parenthood status. The gender-specific entropy distributions point to notable differences. While the general pattern is the same for men and women, the plots reveal differences in the temporal shape and the level of entropy. Among males, for example, the entropy at the end of the observation period exceeds the initial entropy at age 18. By contrast, the entry and exit levels of entropy among women seem to be very similar.

2.4.2 Data Representation Graphs

The previous section illustrated a variety of approaches toward visualizing summary statistics of sequence data. Although the presented figures provide valuable information in an accessible fashion, they lose sight of how individual sequences unfold and thus are not well suited for uncovering and visualizing groups of similar sequences. Moreover, figures and tables of aggregated data must be interpreted with caution in order to avoid the risk of committing the ecological fallacy. This is nicely illustrated in Figure 2.5, which shows three visualizations based on two constructed data sets, each comprising 10 sequences of length $k = 5$.

Next to a state distribution plot, the figure displays two sequence index plots. In the latter, each horizontally stacked bar represents one individual sequence. Although the sequences are sorted by the first occurrence of the partnership status “married,” the index plots are visually less structured than the state distribution plot in the left panel of the figure. Communicating a higher amount of information (i.e., actual individual sequences) comes at the price of more visual complexity. The two index plots literally represent the full data, while the distribution plot summarizes them. The left panel of Figure 2.5 shows that the two very different samples of sequences can be summarized by one single state distribution plot. If the distribution plot is accurately interpreted, this does not pose a problem. For instance, the distribution plot shows neither that 20% of the observed

Figure 2.4 State distribution plots and entropy by gender

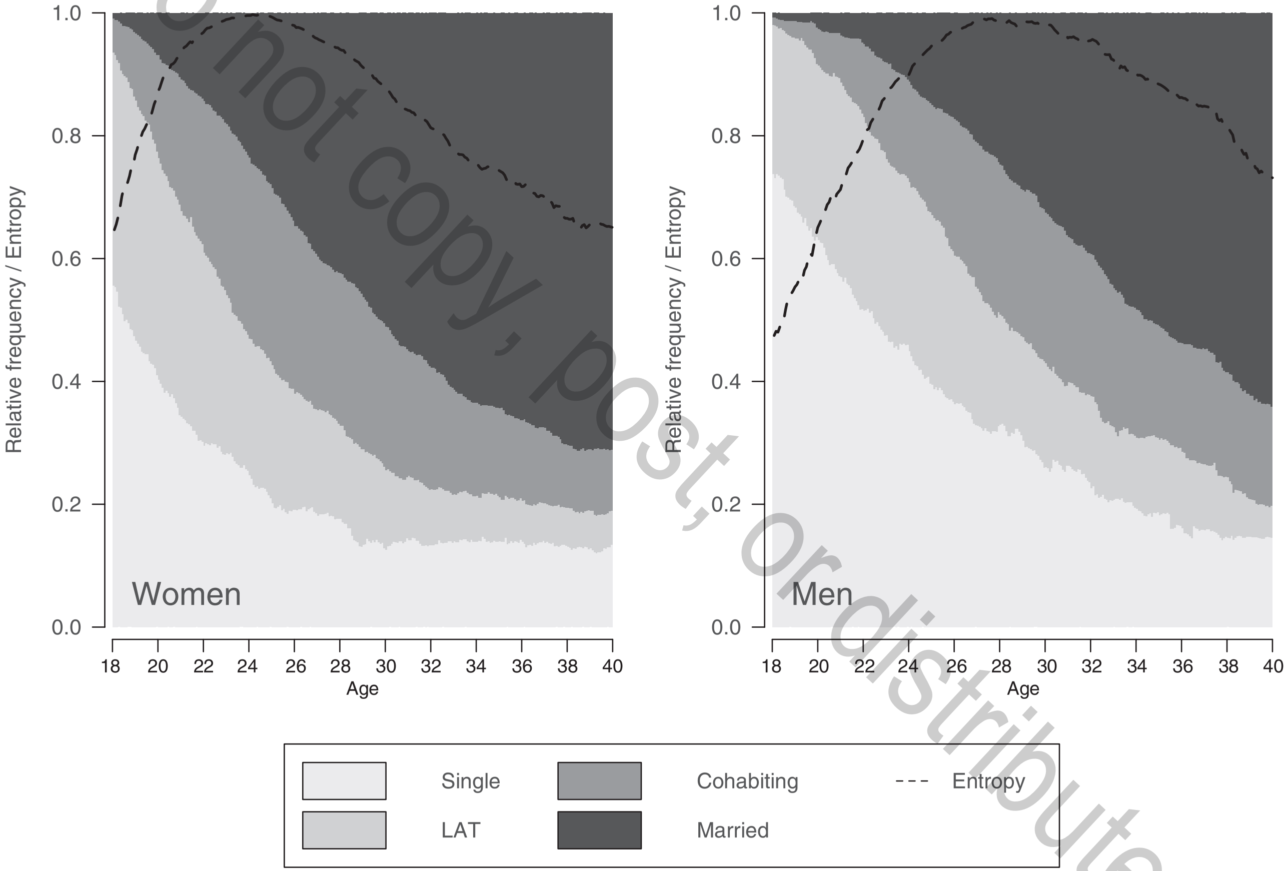
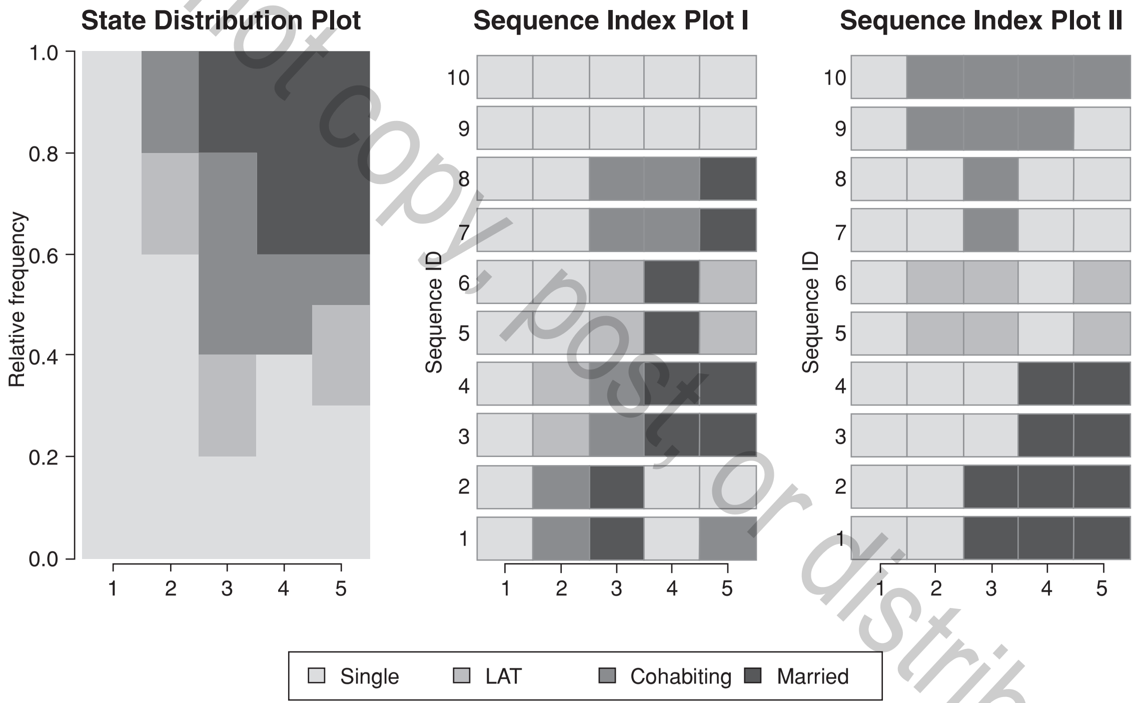


Figure 2.5 Different data producing the same state distribution plot



persons are permanently single nor that 60% never marry. Instead, it depicts that at every position of the sequences at least 20% currently do not have a partner and that 40% of the sample is married at the sequences' two final positions. The two index plots demonstrate that this aggregate picture can be brought about by two distinct populations. In the first sample, 80% of the individuals are married at some point. In the second sample, not a single case stays without a partner throughout the entire observation period.

By fully displaying the sequences, index plots are a valuable explorative visual tool for detecting structure in sequence data (Brzinsky-Fay, 2014). The accuracy and usefulness of index plots, however, hinge on the complexity of the visualized data. If the alphabet becomes too large, the resulting graph is difficult to decipher, particularly if the use of color is not an option. Keeping an eye on the number of plotted observations is an even more significant issue than the size of the alphabet. If too many sequences are displayed in one index plot, the issue of overplotting arises. That is, due to a lack of space in the plot region the stacked bars (or lines) depicting the individual sequences are partly plotted on top of each other and thus produce an inaccurate representation of the data. Depending on the plot size, overplotting arises if more than 300 to 400 sequences are displayed in one plot (Brzinsky-Fay, 2014).

The problem can be alleviated slightly by sorting the sequences such that similar sequences are plotted next to (and partly on top of) each other, rather than in the order they appear in the data set. This is not a satisfactory solution, though, because overplotting of (similar) sequences still occurs. Thus, reducing the number of plotted sequences is a more promising approach. This can be achieved by applying different strategies such as plotting (a) the most frequent sequences, (b) a random sample of sequences, or (c) a sample of representative sequences.

The first strategy is viable only if a few sequences represent a large share of the data, which is rarely the case in social science applications. Even our yearly sequence data of 1,866 partnership trajectories with a moderate sequence length of $k = 22$ and an alphabet with only four different states already comprises 1,432 unique sequences. Therefore, the second strategy, plotting a random selection of sequences, usually produces a more representative visualization of the sequences. The sampling, however, might distort the visualization slightly. Therefore, sampling and plotting should be applied repeatedly to estimate the extent of potential misrepresentation (Brzinsky-Fay, 2014).

Instead of visualizing random samples of the data, more sophisticated strategies aim at extracting and plotting only those sequences that represent the data best. According to Fasang and Liao (2014), some of these approaches, such as representative sequence plots (Gabadinho, Ritschard, Studer, & Müller, 2011) or the smoothing techniques suggested by Piccarreta (2012) try to improve the visualization of sequences by either removing redundant information or plotting only the most relevant sequences. Taking a different approach, relative frequency sequence plots, developed by Fasang and Liao (2014), aim at reducing visual complexity and the problem of overplotting while maintaining an accurate representation of the data across the full spectrum of observed sequences. This goal is achieved by the following procedure.

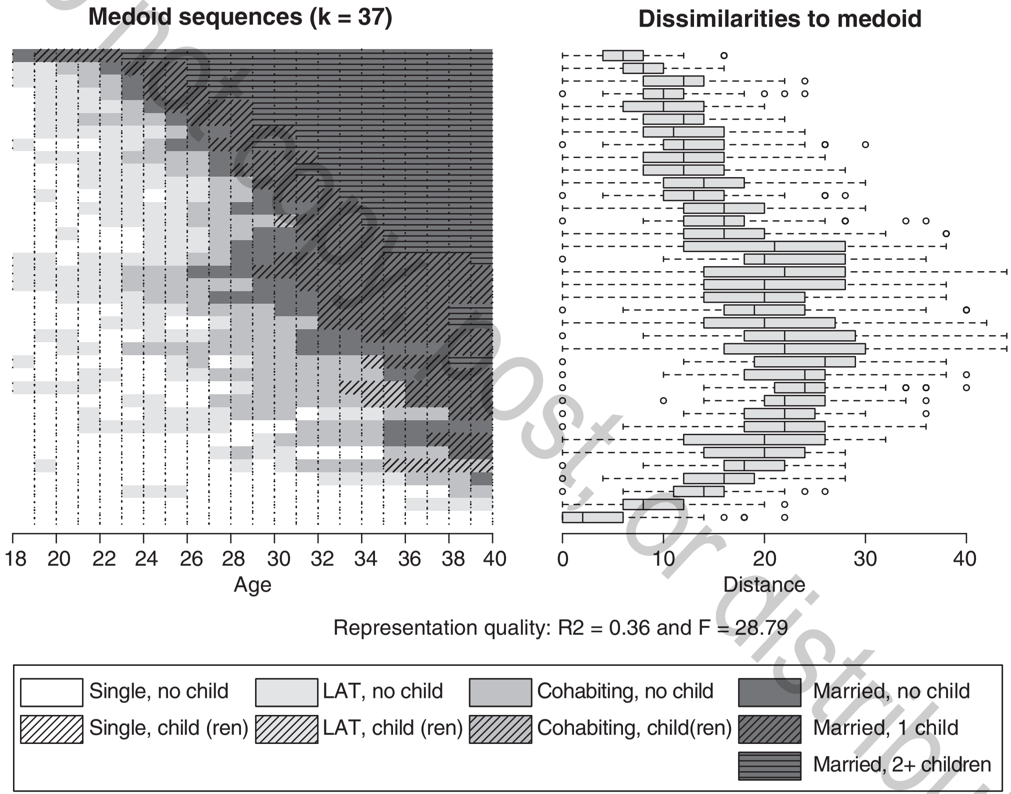
First, the sequences are sorted according to a substantively meaningful principle such as the timing of a specific transition (e.g., age at first marriage) or the score on the first factor obtained by multidimensional scaling (MDS) of a dissimilarity matrix (Piccarreta & Lior, 2010). In our experience, the latter sorting strategy often produces better results, although the quality heavily depends on the chosen dissimilarity measure (see Chapter 3).

Although regular index plots are considerably improved by sorting the sequences, the order of sequences is even more critical in the context of sequence frequency plots. In regular sequence index plots, initially introduced by Scherer (2001), sorting eases the recognition of patterns without changing what is displayed. That is, irrespective of the sorting order, all sequences are rendered. In relative frequency index plots, however, the rendered medoids and the goodness of fit of the plot differ depending on the sorting order (see companion page for an example).

Once the data are sorted, they are divided into k similarly sized frequency groups. The next step extracts the medoid sequence for each of these groups. The resulting sample of k medoid sequences is rendered in an index plot. The index plot can be complemented by box-and-whisker plots, which visualize the distribution of the dissimilarities to the medoids in the k frequency groups. Finally, an R^2 statistic and an F test can be calculated to assess the goodness of fit of the relative frequency sequence plot. As the quality of the results is affected by several analytical choices (sorting criterion, chosen dissimilarity measure, number of frequency groups), we recommend evaluating and comparing different solutions.

When dealing with large amounts of sequences, relative frequency sequence plots are a very powerful visualization tool ensuring readability and visual appeal by rendering a representative selection of medoids instead of all sequences, as would be the case in a regular index plot. As general heuristic, Fasang and Liao (2014) recommend dividing the sample into approximately 100 frequency groups. If colored figures are not an option, it is reasonable to reduce the number of rendered medoid sequences further, particularly if the alphabet comprises more than four states. However, this is a feasible strategy only if the resulting plot still accurately represents the data. Figure 2.6 presents such a parsimonious version of a

Figure 2.6 Relative frequency sequence plot and boxplot of dissimilarities to medoids



relative frequency sequence plot, which renders only 37 medoid sequences, each representing approximately 50 sequences. On the companion page, we present alternative specifications with more frequency groups that produce results very similar to those presented here.

Unlike the state distribution plots presented earlier, the relative frequency sequence plot in the left panel of Figure 2.6 provides information on the temporal order of states. The plot indicates, for instance, that the family trajectories in the pairfam sample are characterized by a close link between marriage and the transition to parenthood, which typically is observed within the first 3 years of marriage. The figure also shows that marriage usually is preceded by a cohabitation spell, which tends to be longer among those who marry later.

The box-and-whisker plot on the right panel of Figure 2.6 shows how well the 37 medoids represent their respective frequency groups. The medoids are most similar to the other sequences in their frequency group among those who marry before they turn 30 and have two or more children within these marriages, and among those who neither marry nor become parents and instead remain single most of the time. The more turbulent medoid sequences in between show higher distances to the sequences they ought to represent.

Comparing the state distribution plot (Figure 2.3) with the relative frequency sequence plot (Figure 2.6) reveals marked differences in the distribution of states at different positions of the sequences. At age 40, for instance, the proportion of unmarried persons in the distribution plot is approximately twice as high, while at age 18, the share of singles among the medoids is much higher than the corresponding proportion in the distribution plot. These differences call for caution when interpreting (the prevalence of states in) relative frequency plots. The frequency groups with higher distances to their medoid most likely comprise a notable share of sequences that are characterized by states that are not identical to the respective medoid states. As a result, the distribution of states derived from a relative frequency plot should be considered reliable only if the plot has a high goodness of fit and if it corresponds with the actual distribution in the full data. Note, however, that the observed discrepancy of the state distribution does not indicate that the relative frequency sequence plot is wrong. It merely reminds us of the fact that this plot is a technique that is reducing the complexity of sequence data by plotting only a selection of representative sequences. According to Fasang and Liao (2014, p. 658), it performs best "when there is strong but fuzzy patterning in the data, that is, when there is patterning into similar sequences but there are few identical sequences." In most applications, the degree of fuzzy patterning and thus the quality of the plot can and should be increased by plotting more homogeneous subgroups rather than the full sample. Rendering relative frequency plots of family trajectories by the level of education or gender, for instance, would produce more accurate and insightful representations of the data.

2.5 Description of Sequences II: Assessing Sequence Complexity and Quality

Except for representative sequences, the descriptive tools introduced in Section 2.3 are providing aggregated summary measures or cross-sectional snapshots of sequence data. In contrast, this section explores indicators for summarizing the longitudinal characteristics of individual sequences. The discussed indicators differ in their capabilities of accounting for the order of states (sequencing), the duration of states, and the quality of the states that constitute a sequence.

2.5.1 Unidimensional Measures

Sequencing—counting Transitions and Subsequences

Unidimensional indicators focus on describing one aspect of a sequence. The sequencing of states, for instance, could be captured by counting either the number of transitions or the number of subsequences of a given DSS sequence, with the latter being the more nuanced measure. That is because the number of transitions is not affected by the states between which the transitions occur, while the number of subsequences increases if more distinct states are involved; every additional distinct state adds one extra subsequence of length $k=1$. The two sequences $x=(S,LAT,COH,MAR)$ and $y=(S,LAT,COH,S)$, for example, both contain four out of four possible transitions, but they differ in their number of subsequences with $\phi(x)=16$—the maximum for a sequence of length $k=4$ and an equally sized alphabet—and $\phi(y)=15$ (see Table 2.9). Given that sequence x contains every state only once, whereas S appears twice in sequence y, the subsequence indicator is performing better than the number of transitions in retaining the visual impression that x is more complex than y.

That said, the SA literature rarely reports the raw subsequence indicator because an increasing length of the examined sequences leads to a rapid inflation of the number of subsequences. For this reason, Elzinga (2010) proposed to use $\log_2 \phi(x)$ rather than a raw count of subsequences to measure the degree of sequencing.[3]

[3] Elzinga named this index *turbulence*. He used the same term for a second index that also accounts for the time spent in each state. In accordance with the literature, and in order to avoid confusion, we reserve the term *turbulence* exclusively for the second index.

Table 2.9 The subsequences of sequences *x* and *y*

Sequence x = (S,LAT,COH,MAR); $\phi(x)$=16				
***k* = 0**	***k* = 1**	***k* = 2**	***k* = 3**	***k* = 4**
λ	*S*	*S, LAT*	*S, LAT, COH*	*S, LAT, COH, MAR*
	LAT	*S, COH*	*S, LAT, MAR*	
	COH	*S, MAR*	*S, COH, MAR*	
	MAR	*LAT, COH*	*LAT, COH, MAR*	
		LAT, MAR		
		COH, MAR		
Sequence y = (S,LAT,COH,MAR); $\phi(y)$=15				
***k* = 0**	***k* = 1**	***k* = 2**	***k* = 3**	***k* = 4**
λ	*S*	*S, LAT*	*S, LAT, COH*	*S, LAT, COH, S*
	LAT	*S, COH*	*S, LAT, S*	
	COH	*S, S*	*S, COH, S*	
		LAT, COH	*LAT, COH, S*	
		LAT, S		
		COH, S		

Both the number of transitions and the logarithmic subsequence indicator can be normalized to have values between 0 and 1 by dividing them by their theoretical maximum. In the case of transitions, the maximum equals $k-1$. The theoretical maximum of subsequences is obtained by counting the number of subsequences of a sequence that is constructed by repeating the elements of the alphabet until the length of the examined sequence is reached. Accordingly, the normalized subsequence measure for x equals 1, while it is $\frac{\log_2 15 - 1}{\log_2 16 - 1} = 0.97$ for sequence y. Note that 1 is subtracted from the logarithmic number of subsequences in the numerator and denominator to ensure a minimum of 0 for the normalized index. Otherwise, only an empty sequence could reach the minimum.

Duration—Longitudinal Shannon Entropy

Earlier, we introduced entropy as an aggregate measure of the dispersion of states at different positions of the sequences. However, the entropy index can also describe how the duration of the time spent in each state of the alphabet is distributed within individual sequences. The normalized longitudinal entropy is defined as

$$S = \frac{-\sum_{i=1}^{a} \pi_i \times \log\left(\pi_i\right)}{\log a}$$

with π_i depicting the relative frequency of time spent in state i and a indicating the size of the alphabet A. The maximum entropy value of a sequence is reached only if the same amount of time is spent in each state of the alphabet. In our earlier toy example, the entropy value for sequence x is identical with the two normalized sequencing indicators. The differences between the measures, however, become evident if we expand sequence x by attaching the same states once again—$x_2 = \left(S, LAT, COH, MAR, S, LAT, COH, MAR\right)$—and compare it with the sequence $y_2 = \left(S, S, LAT, LAT, COH, COH, MAR, MAR\right)$. In this example, we still obtain a Shannon entropy of 1 for both sequences, whereas only sequence x_2 reaches the maximum number of subsequences ($\phi = 224$) and transitions ($k - 1 = 7$). Sequence y_2 comprises only three transitions and 16 distinct subsequences, which translates into a normalized subsequence index of 0 and a normalized index of $\frac{\log_2 16 - 1}{\log_2 224 - 1} = 0.44$ transitions of $3/7 = 0.43$.

2.5.2 Composite Indices

In what follows, we introduce well-established measures that consider sequencing and duration simultaneously. These indices are complemented by more recent approaches that factor in a feature that is even more difficult to measure, sequence quality. Table 2.10 provides a comparison of the different indices for a set of constructed partnership sequences using the same alphabet as for our pairfam partnership sequences.

The table illustrates how the description and comparison of sequences hinge on the chosen indicator. Sequences 5 and 6, for instance, are identical in terms of sequencing. However, they obviously differ in terms of the distribution of times spent in different states (see entropy values). In contrast, the distribution of state durations is identical in sequences 11 and 12 (resulting in the same entropy of 0.68) while the sequences differ with regard to sequencing, with one additional transition observed in sequence

Table 2.10 Comparison of longitudinal sequence indices

ID	Sequence	Transitions	Entropy	Turbulence	Complexity	Precarity	Quality
1	*(S,20)*	0.00	0.00	0.00	0.00	0.20	0.00
2	*(MAR,20)*	0.00	0.00	0.00	0.00	0.00	1.00
3	*(MAR,5)-(COH,5)-(LAT,5)-(S,5)*	0.16	1.00	0.47	0.40	0.73	0.07
4	*(S,5)-(LAT,5)-(COH,5)-(MAR,5)*	0.16	1.00	0.47	0.40	0.20	0.43
5	*(S,3)-(LAT,1)-(COH,6)-(MAR,10)*	0.16	0.82	0.27	0.36	0.20	0.74
6	*(S,4)-(LAT,4)-(COH,6)-(MAR,6)*	0.16	0.99	0.42	0.39	0.20	0.50
7	*(MAR,6)-(S,4)-(LAT,4)-(COH,6)*	0.16	0.99	0.42	0.39	0.29	0.10
8	*(S,10)-(MAR,10)*	0.05	0.50	0.40	0.16	0.20	0.74
9	*(S,2)-(LAT,5)-(S,3)-(COH,5)-(MAR,5)*	0.21	1.00	0.42	0.46	0.42	0.43
10	*(S,2)-(LAT,5)-(COH,5)-(MAR,5)-(S,3)*	0.21	1.00	0.43	0.46	0.50	0.36
11	*(S,2)-(MAR,10)-(COH,8)*	0.11	0.68	0.24	0.27	0.35	0.36
12	*(S,2)-(MAR,2)-(COH,8)-(MAR,8)*	0.16	0.68	0.28	0.33	0.35	0.66

12. Often, researchers are interested in both aspects and want to consider duration and sequencing at the same time. This can be achieved by the turbulence and complexity indices.

Turbulence

The turbulence index proposed by Elzinga (2010) combines the number of subsequences with the variation in the time spent in each *episode* of the sequence, whereas the entropy index considers only the total time spent in each *state*. Turbulence is defined as

$$T(x) = \log_2\left(\phi(x)\frac{s_{t,max}^2(x)+1}{s_t^2(x)+1}\right)$$

with $s_t^2(x)$ denoting the sequence's x variance of the state durations t and

$$s_{t,max}^2(x) = (k(x)-1)(1-\overline{t})^2$$

indicating the maximum of that variance given the total duration of the sequence, with $k(x)$ denoting the length of the DSS sequence and $\overline{t}$ the average of state durations. The obtained index can be normalized by $T(x)-1/T_{max}(x)-1$, where T_{max} is defined as the turbulence value of a sequence, which is as long as the longest sequence (STS format) in the examined set of sequences and constructed by repeating the states of the alphabet until this length is reached. In our example set, this would be a sequence that repeats the four partnership states five times. The turbulence index is increasing when either (a) the variance of the time spent in each state is decreasing or (b) the number of subsequences is increasing. As a result, sequences with the same number of subsequences—such as sequences 5 and 6—will have different turbulence values when they differ in terms of the variance of state durations.

Complexity

The complexity index is another composite measure simultaneously considering sequencing and duration. The index introduced by Gabadinho and colleagues (2010) combines the number of transitions and within-sequence entropy. It reads,

$$C(x) = \sqrt{\frac{(k(x)-1)}{(k_{STS}(x)-1)} * S(x)}$$

with $S(x)$ denoting the normalized entropy, $k(x)$ the length of the DSS sequence, and $k_{STS}(x)$ the length of the same sequence in STS notation. Sequences 3 to 7, for example, all have a DSS sequence length of 4 and an SPS length of 20. The quotient of the equation thus equals $(4-1)/(20-1)=0.16$, which happens to be the normalized number of transitions. Combining two normalized indices, the complexity measure equals 0 for sequences without transitions and thus without variation in the state duration. The maximum of 1 can be reached only if the same amount of time is spent in each state and if the DSS sequence is of the same length as the STS sequence.

Although turbulence and complexity take different approaches to generate a composite measure, they are usually highly correlated. In our set of constructed sequences, the correlation is 0.87. In the pairfam data of yearly partnership trajectories, we obtain a correlation of 0.89. Both indicators provide information on the unpredictability or instability of sequences and are of most use if they are analyzed in tandem with other variables to compare different subgroups. While most SA applications still focus on the comparison and clustering of sequences using dissimilarity measures (Chapters 3 and 4), the analytical potential of composite indices is increasingly acknowledged in more recent applications (Biemann et al., 2011; Van Winkle, 2018; Van Winkle & Fasang, 2017). Table 2.11 illustrates one possible application using the turbulence and complexity of combined fertility and partnership sequences as a dependent variable in a simple linear regression model. The results indicate that family biographies are less predictable among highly educated persons, whereas the sequences of first-generation migrants are less complex or turbulent than those of the autochthonous majority. Further analysis revealed that the migrants' comparatively low complexity and turbulence scores are driven by their earlier onset of family formation and a high prevalence of the status "married, at least two children," which is their sequences' modal state for 12 out of 22 years (compared to 8 for the persons without a migration background). Compared to most standard dependent variables in social science applications, composite indices are rather complex outcome measures that summarize multiple aspects of sequence data, such as sequencing and duration of states. Moreover, substantively different sequences can exhibit the same degree of complexity or turbulence. As a result, single indicators, such as education or gender, tend to capture only a small share of the variation of these outcome measures, which is also the case in our example regression shown in Table 2.11.

A New Wave of Composite Measures Assessing Sequence Quality

Although they provide a good summary of sequence volatility, turbulence and complexity do not consider potential qualitative differences

Table 2.11 Composite indices used dependent variables in linear regressions

	Turbulence	**Complexity**
Intercept	0.29^{***}	0.32^{***}
	(0.00)	(0.00)
Woman (ref.: man)	$–0.01^{**}$	–0.01
	(0.00)	(0.01)
Education: high school	0.04^{***}	0.04^{***}
	(0.00)	(0.01)
Migration status (ref.: no migration background)		
1st generation	$–0.04^{***}$	$–0.04^{***}$
	(0.01)	(0.01)
2nd generation	0.01	0.01
	(0.01)	(0.01)
Observations	1,809	1,809
R^2	0.06	0.05

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; standard errors in parentheses.

between individual states and transitions. Sequences 3 and 4 from Table 2.10, for instance, comprise the same states but in reverse order. For each of the first four indicators displayed in the table (transitions, entropy, turbulence, and complexity), these two sequences produce identical results. Many people, however, would argue that there is a qualitative difference in family biographies of persons who see their marriage ended and those of persons who end up being married.

Against this background, several approaches have been suggested to measure sequence quality or weight existing composite indices by a factor grasping sequence quality (Brzinsky-Fay, 2007; Van Winkle & Fasang, 2017). Among the most elaborated approaches in this domain of the SA literature are the indices that have been suggested by Ritschard et al. (2018) and Manzoni and Mooi-Reci (2018). Computing these and related indices requires that the researcher specify some sort of qualitative hierarchy between the states of an alphabet. That is, some states or transitions between states, such as from unemployed to employed or from divorced to married, have to be declared as more positive or desirable than others. In most social science applications, establishing a quality hierarchy involves normative and controversial decisions that might be difficult to defend and thus should be made explicit when they are imposed.

The precarity index is based on the complexity index $C(x)$ and combines it with a notable set of other parameters that will be described only briefly here. For a more detailed introduction to the index, we refer to Ritschard et al. (2018). The index is defined as

$$prec(x) = \lambda a(x_1) + (1-\lambda)C(x)^{\alpha}\left(1+q(x)\right)^{\beta}$$

The basic idea is to weight the well-established complexity index by a correction factor measuring whether positive or negative transitions mainly characterize a sequence. In focusing on transitions, the correction factor does not consider the time spent in high- or low-quality states. The precarity index increases with the number of negative transitions and the degree of complexity. The default correction factor is obtained by subtracting the proportion of positive transitions from the proportion of negative transitions: $q(x) = q(x)^{-} - q(x)^{+}$. The index allows for several different approaches to determine $q(x)$, which include the exclusive consideration of positive or negative transitions or the assignment of weights for different transitions. The parameter $a(x_1)$ denotes the precarity of sequence x 's starting state and the parameters control the relevance of the different components of the index: λ controls the relative importance of the sequence's starting state vis-à-vis the weighted complexity index, whereas the exponents α and β specify the relevance of the complexity index as opposed to the correction factor.

Apart from those parameters, and most importantly, the researcher has to assign a rank order to the states of the alphabet, which allows distinguishing precarious from positive transitions. Initially, the index was developed to study employment trajectories, for which it is easier to establish a hierarchy between states than for family trajectories. That said, even for employment trajectories, it might not be straightforward to establish a rank order between all states of an alphabet. While it might seem reasonable to justify that being full-time employed is better than being unemployed, things become much more complicated if the alphabet also comprises states such as part-time employment or higher education. Given the common lack of clear hierarchies, the index allows the researcher to assign the same rank to multiple states (class of equivalent states) or declare them to be noncomparable states. Transitions that involve noncomparable states or that occur between equivalent states do not contribute to the correction factor.

In sum, the precarity index is a very flexible tool to jointly consider a sequence's order of states, state duration, and quality. The flexibility, however, comes at a high price: The results very much hinge on the specification of various complex parameters and thus provide an easy target for critique. The developers of the index argue that the "index provides most often sensible results with default parameter values and automatic methods for setting transition weights and starting precarity degrees" (Ritschard et al., 2018, p. 293). That said, we recommend recalculating the index with slightly adjusted parameters as a robustness check until the statistical properties of the index have been further evaluated by additional research.

Although in our view, the precarity index does not lend itself particularly well to studying partnership trajectories, we calculated it for the example sequences in Table 2.10 using the default parameters. Accordingly, we only had to specify a rank order of states. Taking a traditionalist's perspective, we came up with the following rank order of partnership states:

$$MAR > COH > LAT > S$$

Because the correction factor of the precarity index is not affected by the time spent in states of different quality, sequences with the same order of transitions end up with the same $q(x)$. If a sequence includes only positive transitions, $q(x)$ equals 0 and the precarity index boils down to the weighted precarity of the sequence's starting state. The precarity vector for the states in our alphabet is $S = 1, LAT = 0.67, COH = 0.33, MAR = 0$. With the default λ of 0.2, the precarity index for the sequences 4, 5, 6, and 8, which all include only upward transitions, therefore equals 0.2. The same is true for sequence 1 with a complexity of 0 and single as the starting state. These results illustrate that quite different sequences can have the same score on the precarity index either because they experience only positively evaluated transitions or no transitions at all. As already indicated by the term *precarity*, the index's main aim is to grasp negatively rated or precarious transitions. This becomes obvious in sequences 9 and 10, which are characterized by the same entropy but show different precarity scores. Both sequences contain one downward transition. While sequence 9 transitions from a LAT relationship into the single state, sequence 10 moves from marriage to singlehood. Given our hierarchy of states, the latter is a much more precarious transition, yielding a higher precarity score for sequence 10.[4]

The sequence quality index proposed by Manzoni and Mooi-Reci (2018) is another recent effort to go beyond the traditional composite measures.

[4] For a more refined application proposing a weighted partnership complexity index based on the precarity index, we refer to Hiekel and Vidal (2020).

The index is re-introducing an idea which was originally developed under the label integrative potential or capability in a research paper studying school-to-work transitions by Brzinsky-Fay (2007). Given that Manzoni and Mooi-Reci (2018) provide a more formal and general discussion of the index, we refer to it as quality index instead of using the term integrative capability that seems to be coined specifically for the study of employment trajectories. Different from the precarity index, this approach is interested in the quality of states rather than the quality of transitions. Quality is assessed by dividing the alphabet into states of success and states of failure. The resulting dichotomous sequences are evaluated to create a quality index that increases with the number of states indicating success. In addition, the index embraces the principle that more recent successes should contribute more to sequence quality than successes from the past. The index reads

$$\gamma^w\left(x^k\right)=\frac{\sum_{i=1}^{k} p_i^w}{\sum_{i=1}^{k} i^w}, \text{ with } p_i \begin{cases} i \; if \;\; x_i = S \\ 0 \;\; otherwise \end{cases}$$

where i indicates the position within the sequence, $x_i = S$ denotes a state of success at position i, and w is a weighting factor that affects how strong and fast the index reacts to and recovers from failure states.

Before we turn to the example sequences shown in Table 2.10, we illustrate the impact of different weighting factors using the following little sequence: *(S,2-LAT,1-COH,1-MAR,2)*. *MAR* is considered a success, while all other states are deemed to be failures.

$$\gamma^0 = \frac{0+0+0+0+1+1}{1+1+1+1+1+1} = 0.33;\; \gamma^1 = \frac{0+0+0+0+5+6}{1+2+3+4+5+6} = 0.52;$$

$$\gamma^2 = \frac{0+0+0+0+5^2+6^2}{1^2+2^2+3^2+4^2+5^2+6^2} = 0.67$$

If the weight equals 0, the quality index is just indicating the proportion of successful states. A recent success counts just as much as earlier successes. Increasing the size of the weight changes this behavior and emphasizes the more recent events. As a result, the example sequence's quality index is increasing with the size of the weighting factor.

The values for the quality index shown in Table 2.10 use a weight of $w = 1$. The first two indices depict the two extremes of the distribution consisting exclusively of either failure or success states. The differences in the quality scores of the two sequence pairs 9 and 10 or sequences 11 and 12 illustrate how the quality index emphasizes more recent states of success. In both pairs, the sequences comprise the same number of states spent in marriage. However, these states appear in different positions. As a result, the quality index is higher for sequences 9 and 11, in which the marriage spell occurs at the end of the sequences. The recency of the 5 marital states in sequence 9 even outweighs the 10 marital states occurring early in sequence 11 ($\gamma_9 = 0.43$; $\gamma_{11} = 0.36$). Reducing w would change this behavior by attenuating the recency effect.

In sum, the index provides a straightforward measure of sequence quality, which avoids the necessity of specifying a large number of parameters. The index can also be conceptualized as a time-varying variable by calculating it repeatedly, initially considering only a sequence's starting state and then incrementing it by one additional sequence position until the total length of the sequence is reached. In its current form, however, the calculation of the quality index requires boiling down the alphabet to two states indicating either success or failure. This arguably poses a problem for many social science applications that often call for incorporating a more nuanced hierarchy of success, as suggested by the precarity index.[5]

The computation of composite indices trying to grasp qualitative features of a sequence is an area of research that only recently gained momentum in the SA literature. Thus, the indices shown in this chapter should be considered only a small snapshot of an actively evolving field of research. A specific evaluation of composite indices in the context of life course research has been conducted by Pelletier et al. (2020). For a very recent, more detailed, and comprehensive review of this field, we refer to the excellent overview article by Ritschard (2021).

[5] On the companion page, we present a generalized version of the sequence quality index that allows the researcher to specify a quality hierarchy containing more than two states.