Typologies in Sequence Analysis:

Practical Guidelines for Identifying Robust Cluster Solutions

Stefan Bastholm Andrade

VIVE - the Danish Center for Social Science Research

Department of Psychology, Health and Technology, University of Twente

Anette Eva Fasang

Department of Social Science, Humboldt University Berlin

Satu Helske

INVEST Research Flagship Center, University of Turku

Department of Social Research, University of Turku

Aleksi Karhula

Ecosystems and Environment Research Programme, University of Helsinki

Acknowledgements:

We gratefully acknowledge funding from the project EQUALLIVES, which is financially supported by the NORFACE Joint Research Programme on Dynamics of Inequality Across the Life-course, which is co-funded by the European Commission through Horizon 2020 under grant agreement No 724363.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 724363.

Abstract

Sequence analysis in the social sciences heavily relies on cluster techniques to identify typologies. Clustering techniques and statistical cluster cut-off criteria for selecting the optimal number of clusters have greatly improved. In contrast, we lack a systematic assessment of how data features, such as the sequence sample size, the number of time points in the sequences, and the number of distinct states in the sequence alphabet might systematically impact the identification of sequence typologies. Drawing on both simulated data from mixture Markov models and real data from the German Family Panel survey, we provide best-practice guidelines for applied researchers to gauge whether their data is sufficient for extracting robust sequence typologies, if they empirically exist. Sequence typologies are most robust for samples with at least 500 sequences, sequence lengths greater than 10 time points, and state alphabets that have at least as many states as the "true" number of clusters.

1. Introduction

Sequence analysis is widely used in social and demographic research to analyze life trajectories of categorical states (Cornwell 2015; Aisenbrey and Fasang 2010; Brzinsky-Fay and Kohler 2010). A key features of sequence analysis is that it offers a holistic approach to analyze detailed descriptions of people's life trajectories as they unfold over time as units of analysis (Abbott and Tsay 2000). To reduce complexity, researchers typically apply cluster techniques to create ideal typical groups that summarize the most characteristic differences between groups of life trajectories (Ritschard and Studer 2018; Barban and Billari 2012). For example, Aisenbrey and Fasang (2017) applied a sequence typology approach to classify typical parallel work-family trajectories in Germany and the United States. Compared to Germany, typical work-family life courses in the United States are less gendered but more class-stratified.

A contested issue in sequence analysis is to what degree life course typologies are statistically robust (Studer 2021). Scholars have developed a number of effective methods to measure differences between sequences (Ritschard 2021; Studer 2013) and improve cluster techniques for identifying ideal types in the data (Studer and Ritschard 2016; Studer et al. 2010). Yet, we lack a basic understanding of the importance of data limitations for the identification of sequence typologies (for an early remark, see Gabadinho et al. 2009b). For example, it is unknown to what degree small sequence sample sizes, short sequence lengths, or too many categories with few observations in the sequences compromise the identification of robust sequence typologies. These questions are of crucial importance for applied researchers to prepare their data for sequence analysis and evaluate the soundness of the resulting typology.

Typically, researchers explore the robustness of sequence typologies by applying different cost specifications when calculating dissimilarities between sequences (Aisenbrey and Fasang 2017), or evaluating cluster solutions with different numbers of clusters (Piccarreta and

Struffolino 2019). Recently, studies started to exclude poorly classified sequences from clusters based on low silhouette values when clusters are used as dependent or independent variables in further regression-based analyses (Jalovaara and Fasang 2020). Although such explorations are useful for evaluating sequence typologies and limiting the influence of poorly classified sequences in the clusters for further analyses, they do not provide any guidance on the sensitivity of sequence typologies to data features. Formal guidelines on the influence of data features are particularly important in comparative analyses that rely on different types of data, such as, prospective data, retrospective survey data, and administrative registers (Fasang et al. 2022). When data come from different sources, we have even less guidelines to evaluate whether results are artifacts of different characteristic of data sets, or substantially meaningful.

To address this gap, we develop a set of formal guidelines to deal with the uncertainty in identifying the optimal number of types in a cluster-based typology for sequence data. Inspired by previous work on how to construct typologies (Studer 2013; Shalizi 2009), we argue that a valid typology should be generalizable to other observations and thus not dependent on the sampling. Our analyses demonstrate how three types of data challenges systematically affect the identification of ideal types in sequence analysis, that is, 1) size of the sequence sample (i.e., number of sequences), 2) the number of time points (sequence length), and 3) the number of distinct states in the sequence alphabet (alphabet length). These three data challenges correspond to issues that every sequence analysis study has to address and researchers have to make choices about. Based on a simulation approach with mixture Markov models, we analyze 540,000 samples from simulations that allow us to specify a "true" number of clusters under different scenarios for the data structure. In addition, we analyze 135,000 samples drawn from real data of the German Family Panel survey (Pairfam) (Huinink et al. 2011). Findings show similar patterns across the simulated and the real data based on

which we specify formal guidelines for how to evaluate cluster solutions in view of data features, and how to make key analytical choices in applied sequence analysis.

2. Challenges for identifying sequence typologies

There are both methodological and data-related challenges to identifying the optimal or "true" number of types in a cluster-based sequence typology, assuming that the empirical distributions in the data indeed reflect a meaningful typology. To identify the "true number of groups" cluster analyses should inform 1) whether there is any meaningful group structure in the data or not, and 2) what the optimal number of groups is, that is, the cluster typology that is most discriminant with internally homogeneous groups that are distinct from the other groups and are theoretically and substantively meaningful. Both of these goals can be compromised by methodological choices related to the sequence distance measure and clustering algorithm or by data limitations. While a large and rich literature has discussed the methodological challenges of evaluating robustness of typologies to using different distance measures (Wu 2000; Studer and Ritschard 2016) and different clustering algorithms (Studer 2013, 2021), surprisingly few studies have discussed data-related challenges, such as the minimum necessary sequence sample size, number of time points, or number of distinct states in the sequence alphabet. As a result, researchers tend to approach sequence analysis with few systematic guidelines on how data availability and methodological choices might jointly and systematically affect their results.

To systematically evaluate three common data-related challenges in identifying robust cluster-based sequence typologies, we use life course research as a case and specifically focus on family life courses. Life course researchers typically apply sequence analysis to identify clusters (often understood as ideal types) that describe representative trajectories for individuals belonging

to different social groups, for example, between cohorts (Buchmann and Kriesi 2011; Billari 2001a), social classes (Andrade 2016; Bühlmann 2010), and countries (Aisenbrey and Fasang 2017; Elzinga and Liefbroer 2007). Even studies focusing on dyadic trajectories, for example between parents and children or between siblings tend to use cluster typologies to highlight important differences between subgroups (Karhula et al. 2019; Raab et al. 2014). More recently, the process-oriented perspectives of sequence analysis to create proliferated from social and demographic research into other fields, such as history and political science (Ritschard and Studer 2018; Barban and Billari 2012). An overview and introduction to the basics of social sequence analysis can be found in Cornwell (2015), Raab and Struffolino (2022), and Liao et al. (2022).

To group sequences into a typology, researchers typically follow a two-step approach (Raab and Struffolino 2022; Liao et al. 2022; Cornwell 2015; Aisenbrey and Fasang 2010). First, optimal matching (OM) techniques calculate a pairwise distance matrix between all possible pairs of sequences. Second, this distance matrix enters clustering techniques to identifying groups of ideal typical trajectories (Abbott and Tsay, 2000). Since Abbott (1990) introduced sequence analysis into the social sciences, OM techniques have been the subject of much criticism (Warren et al. 2015; Aisenbrey and Fasang 2010; Brzinsky-Fay and Kohler 2010). While some of the criticisms relate to theoretical matters about a lack of sociological meaning when implementing a method from biology and computer science (Levenshtein 1965) in the social sciences (Levine 2000; Wu 2000), others address mainly methodological and to a much lesser extent data issues that could compromise the robustness of sequence typologies (Warren et al. 2015; Barban and Billari 2012). In response to the criticisms, methods for calculating sequence distances and clustering algorithms have greatly improved (for a comprehensive review see Studer 2021 and Studer and Ritschard, 2018). However other issues, especially the sensitivity of typologies to data features, remain largely

unaddressed but might be just as consequential for the resulting typologies. In the following, we therefore focus on three key data challenges in the identification of typologies from sequence data:

a) Too few sequences. As stated by Lieberson (1991:306), no empirical study should use any type of statistical methods in small N cases without "rigorous justifications of heroic assumptions and a guard against possible distortions". Yet, studies using sequences analysis rarely discuss issues of small sample size that is, having only few sequences. The units of analysis are usually individual sequences but the sequences can also pertain to dyads, households or other aggregate units, such as countries. Typical sample sizes in studies that extract cluster-based typologies range from just 400 to 500 sequences using survey data on individuals (Struffolino et al. 2020; Fasang and Raab 2014) to more than 10,000 sequences extracted from register data (Karhula et al. 2019, Raab et al. 2014, Jalovaara and Fasang 2020). Studies using sequences from household panel surveys, a very common data source, often analyze between 500 and 1,500 individual sequences. Due to panel attrition, sample sizes dwindle quickly in analyses that follow individuals over more than a decade or two. Researchers commonly include missing value states in the sequence alphabet or broaden sample inclusion criteria to maximize sample size of the sequences but it is unclear which target sample sizes would have to be achieved to enable extracting a reliable typology. In addition, due to various technical limitations for implementing survey weights, sequence studies often do not use weights to compensate for small sample size, which can also compromise the representativeness of the resulting descriptive typologies (Mikolai and Lyons-Amos 2017). Some studies present weighted and unweighted typologies, usually as robustness checks (Aisenbrey and Fasang 2017; Liao et al. 2022).

b) Too few time points. A related challenge of insufficient data is a too short sequence length to capture the social process of interest. For example, to analyze employment sequences around a job loss, longer time periods after first re-employment have to be observed to assess, whether re-employment was enduring or individuals continue to cycle in and out of low paid jobs with recurrent intermittent unemployment (on the relationship between time and social context see Abbott 2001). Two issues are at play. First, the length of the observation window, that is whether a process needs to be observed for only a few days or for decades. Second, the number of time points in the sequence, which depend on the observation window and the time intervals observed: days, months, or years for example. Which observation window and time intervals are necessary to capture the relevant patterns is a substantive and theoretical question. In the following, we focus on the number of time points in the sequences, which results from both the observation window and the time intervals considered. For example, for family formation sequences, relatively long time windows (ages 15 to 40) are sensible to capture the entire reproductive phase but a yearly or at most half-yearly interval will suffice as individuals usually do not change their marital, cohabitation and parenthood status on a monthly basis. Annual observations would result in a sequence length of 25 time points between ages 15 and 40, half-yearly intervals would imply 50 time points. Most sequence analyses rely on yearly panel or register data, or monthly retrospective life course data. Notable exceptions are 15-minute intervals in the course of a day in time use research using sequence analysis (Lesnard, 2006, Vagni, 2019). Yearly sequences start at five year-long sequences with monthly sequences going up to several hundred time points (Aisenbrey and Fasang, 2017). There has been practically no discussion in the literature about how and whether the number of time points in a sequence, a key potential source of insufficient data, systematically affect the identification of sequence typologies.

c) Too few distinct states in the sequence alphabet. As with other types of quantitative analyses, too few categories of a variable or a sequence alphabet can obscure the social phenomenon in focus. Too few distinct states in specifying the sequence alphabet could falsely let one state become dominant or simply be overlooked in the identification of cluster typologies (Emery and Berchtold 2022). Common rules of thumb in social sequence analysis are not to include sequence states that occur very infrequently, for example below five percent of the sequence states. Including infrequent states will likely not aid the identification of the main patterns. Yet this could be misleading if there are theoretically very important states that only occur for short time spans. Generally, the specification of sequence states has to be guided by theoretical considerations. Similar to the logic of degrees of freedom, the ability of sequence and cluster analyses to reliably identify typologies might depend on the number of distinct sequence states – in conjunction with the number of sequences and the number of time points. To date there are no guidelines to inform how the number of sequence states included systematically affects the identification of sequence typologies. Typically, sequence studies include between as little as two or three and up to 15 states with most studies specifying between 6 and 12 states (Liao et al. 2022; Aisenbrey and Fasang 2017; Abbott and Tsay 2000).

3. Methodological Design to Assess Sensitivity to Data Challenges

Our methodological design to assess how the three data challenges affect the identification of cluster typologies (i.e., too few sequences, too few time points, and having too few distinct states) draws on two data sources, both of which contain a clustered data structure. Studer (2021) recently proposed a parametric bootstrapping method to assess, whether any meaningful cluster structure exists in the data or not. In contrast, we assess how data limitations affect the identification of the correct number of groups, in data that contains a true cluster structure. First, we simulate data with a

mixture Markov model designed specifically for sequence data (Helske and Helske 2019). Second, we use survey data from 1,866 respondents who participated in the German Family Panel survey (Brüderl et al. 2019; Huinink et al. 2011). Simulations will never be an identical representation of the complexity in observed sequences but they exemplify how different data issues affect sequence typologies in 'a sandbox environment'. Importantly, the simulated data based on the mixture Markov models allows us to create a true number of clusters as a benchmark, which always remains unknown in real data.

3.1 Scenarios

Corresponding to the three data challenges, we define a set of scenarios combining different specifications of: the number of (a) sequences, (b) number of time points, and (c) number of distinct states in the alphabet. We use Mixture Markov models to create the simulated data, which means that we can vary the true number of clusters (d). The scenarios for the simulated data and real data are summarized in Table 1. For the simulated data, we assess 360 different scenarios (six different sequence sample sizes × five different sequence lengths × four different sizes of state alphabet schemes × three different true cluster solutions, see Table 1). For the real data, we assess 90 different scenarios (six different sequence sample sizes × five different sequence lengths × three sizes of state alphabet schemes). We distinguish between six sequence sample sizes, ranging from only 100 sequences to larger samples with thousands of sequences. For the simulated data, we construct scenarios with up to 2,000 sequences. We have a maximum of 1,500 sequences in the real data due to limitations in the original sample, which only include 1,866 sequences in total. Each scenario is drawn 1,500 times, resulting in 540,000 samples in total for the simulated data with 360 scenarios (1,500 × 360) and 135,000 samples for the real data with 90 scenarios (1,500 × 90).

[TABLE 1 ABOUT HERE]

For the sequence length (i.e., number of time points), the scenarios range from five to 120 time points (Table 1), which corresponds to a real-life case of observations across ten years measured in monthly intervals (120 time points), every three months (40 time points), every six months (20 time points), annually (10 time points), or every two years (5 time points). One of the most important methodological decisions in sequence analysis concerns selecting the number and types of states in the alphabet. As noted above, ultimately this decision is a theoretical one: the researcher should include states that are theoretically relevant for the case study, however, the reliability of the cluster typology might also depend on the number of states. Typically, the literature stresses that the number of states should fit the data and researchers therefore should avoid states that occur very infrequently, as they will not aid the main pattern search targeted by sequence analysis (Billari 2001b). The degree of detail in the sequence alphabet (i.e., the number of states in the alphabet) is to a large extent dictated by the data, for example, when information on cohabiting or living apart together relationships are simply not recorded in the data. However, for some research questions, such as studies about vulnerable young people in complex life situations (Bühlmann 2018), also infrequent states are substantively very important, and it can be problematic to subsume them under larger categories. To address this concern, we vary the number of sequence states in our simulations and real data analyses from simple scenarios with only three distinct states to more complex scenarios that include 12 distinct states for the simulations and 9 distinct states for the real data.

3.2 Cluster Identification in Different Data Scenarios

For each scenario, we follow standard procedures and begin by measuring the distances between the sequences (Aisenbrey and Fasang 2010). Researchers typically measure distances in sequence

analysis by calculating the cost of making two sequences identical. Following recent research (Liao 2021; Hart 2019), we use an algorithm called optimal matching of spell sequences (OMspell) (Studer and Ritschard 2016) that resembles the logic of how a mixed Markov model simulates sequences. Based on a probability distribution, a mixed Markov model randomly chooses time point by time point whether to stay in the current state (extending the spell) or to switch to a different state (starting a new spell), making the differences between the simulated sequences most sensitive to ordering of spells and least sensitive to timing. Compared to other distance measures for sequence analysis, the OMspell algorithm emphasizes the ordering and the duration of the spells in the sequences, which means that we treat each spell duration as a distinct state. Thus, five time points in a given state is different from six time points in the same state. To measure the cost of transforming one sequence into another, we define a substitution costs matrix based on substantive considerations about the difference between different states (see Appendix Table A1). We used the OMspell dissimilarity measure, because it is theoretically the most reasonable choice that emulates the rules followed in the Markov models to construct the artificial sequences in our case. We do not expect the conclusions on the sensitivity of sequence typologies to data issues to differ dramatically for other dissimilarity measures or substitution cost specifications, as long as they are sensitive to sequencing.[i]

One critique relates to the subjective interpretation of sequence typologies. Clustering techniques to identify typologies range from simple (hierarchical) algorithms to more complex algorithms that use partitions around medoids (Kaufman and Rousseeuw 1990; Ward 1963) but they ultimately all require the researcher to select the optimal number of clusters. The literature offers several statistical measures to inform this decision. In a review of available cluster cut-off criteria, Studer (2013) highlights the Average Silhouette Width (ASW) measure, which has become the standard indicator in life course research (see also Studer 2021). The ASW indicates, whether

trajectories are more similar within the clusters than between the clusters. A key advantage is that the ASW indicates the internal coherence of each cluster. Silhouette methods compute coefficients for each observation (i.e., individual sequence) that measure how closely each sequence resembles the sequences within the same cluster relative to the neighboring clusters.

In principle, ASW values above 0.5 are considered to indicate reasonable patterns in data, values between 0.26 and 0.50 indicate a weaker data structure, and values below 0.26 indicate no structure (Kaufman and Rousseeuw 1990 in Studer 2013). However, as noted by Studer (2013), the thresholds for the ASW values were developed in cluster analysis of simple random variables, whereas sequences can potentially include much more variation (Liao and Fasang 2020). Therefore, in applied sequence analysis, ASW values above 0.20 are often treated as acceptable. Moreover, individual silhouette values identify poorly classified cases that can be excluded from further analyses when using clusters as dependent or independent variables (Jalovaara and Fasang 2020). Note that the ASW does not consider individual sequences that are outliers to all clusters but only those that are equally similar to several of the identified clusters, and therefore might overstate the true cluster structure (Struffolino and Piccarreta 2019).

For most cluster cut-off criteria, small variations in absolute values for different numbers of clusters should not be interpreted (Studer 2021). In general, support by several cluster cut-off criteria for one cluster solution can be considered relatively strong (Studer 2013). Early on, Aisenbrey and Fasang (2010) concluded that selecting the "best number of clusters" is not a decision that can only be guided by (imperfect) statistical criteria, but also by the criterion of construct validity – if a resulting typology conforms to prior theoretically informed expectations. In our evaluation of the cluster solutions, we supplement the ASW measure with the Point Biserial Correlation (PBC) and Hubert's Gamma (HG), which both provide alternative measures to evaluate the clusters. Results using these measures yield similar results (available from authors).

Recently, Studer (2021) demonstrated how parametric bootstrapping can be used to compare the cluster quality of an observed typology with the quality of one obtained by clustering similar but non-clustered data. This approach relies on the assumption that the initial sequence data used for the bootstrap procedure *is* in principle sufficient and appropriate to identify a sequence typology. The goal of this paper is to add guidelines that can help the researcher to determine whether data fulfill the pre-conditions for reliably identifying the "true" number of clusters in a typology, given that a true number of groups exists in the data at hand.

4. Creating the samples

We first present how we created the 540,000 simulated samples and then the 135,000 samples from real data of the German Family Panel. We draw on a novel approach (Helske and Helske 2019) using mixture Markov models (MMM) to create simulated sequence data. The assumptions of our MMM do not directly transfer to typical life course data from surveys. The main advantage of this simulation is that we fully control the content of the data and can create a true number of groups as a benchmark to test our data scenarios.

A simple (first-order) Markov model estimates the probabilities for starting in each state of the alphabet and transitioning between them. It relies on a number of assumptions. First the model assumes that the observation at time *t* depends on the observation at time *t-1* only, not on any prior history. Second, the transition probabilities are assumed to be constant across the observation window. For our simulations, we use an extension of the basic Markov model, the mixture Markov model (MMM). This model expects that the population consists of clusters (latent classes) with varying sequence patterns and allows for different specifications of initial and transition probabilities between these clusters. The MMM for sequence data can be described with the following notations and probabilities:

– y_{it}: observation of individual *i* at time *t*

– *s, r:* states from the alphabet

– $\pi_k(s)$: Probability, that a sequence starts in state *s* (initial probability) when in cluster *k*

– $a_k(s,r)$: Probability to transition from state *s* to state *r* (transition probability) when in cluster *k*

– w_k: membership probability for cluster *k*

The log-likelihood of the model is calculated as:

$$\log L = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} w_k \pi_k(y_{i1}) \prod_{t=2}^{T} a_k(y_{i(t-1)}, y_{it})\right),$$

where *N* is the number of individual sequences, *K* is the number of clusters, and *T* is the number of time points. When simulating sequences from MMM, we can (and must) determine the numbers of sequences, time points, distinct states, and clusters as well as initial and transition probabilities $\pi_k(s)$ and $a_k(s,r)$ for each cluster *k*, which gives us full control over the contents of the data.

[TABLE 2 ABOUT HERE]

For the simulations we vary the sequence states from 12 states to just three states (Table 2). Resonating with our real data example below and to ease readability, we label the different states in the simulated sequences as family formation states, i.e., partnership (married, cohabitation, and single), and number of children. The resulting data are not designed to be realistic as such but the aim is to create comparable datasets in scenarios of different complexities (see the following guidelines), so we could give the states any other names. Figure 1 shows an overview of all clusters and state alphabets included in the scenarios.

[FIGURE 1 ABOUT HERE]

The MMM simulations begin by constructing the most complex scenario with 12 states, 9 clusters, and 120 time points. We simulated data for all the clusters separately and followed these guidelines within each cluster:

1. All states should be present in data.
2. There should be clusters with simpler and more complex sequences.
3. The clusters should be distinct from one another even if the number of states, time points, and/or clusters are reduced.

To reduce the number of clusters, we picked a subset of the original 9 clusters so that each of the 12 states were present in the 6-cluster and 3-cluster scenarios. To reduce the number of time points, we simulated sequences that were like the original sequences of duration 120 but were observed in coarser intervals. In practice, this can be done by creating new transition matrices in the MMM by multiplying original transition matrices for as many times as the length between the observations (for example, for observations for every three months we multiply the monthly transition probability matrices A with itself three times, i.e., the three-month transition matrix is AAA). See Appendix Table A2 shows the modal sequence state order for each cluster in different scenarios by true cluster number and number of distinct states. A full R script for our approach is available as supplementary material.

In simulated contrast to the simulated data in the real data from German Family Panel, we no longer know the true number of clusters and the data structure becomes more complex. Yet,

we selected real data on family formation processes that are known to be strongly clustered in Germany, which is also supported by statistical cluster-cut-off criteria. The sequences we use are publicly available as example sequences accompanying the introductory book on sequence analysis by Raab and Struffolino (2022) and can be downloaded here: https://sa-book.github.io/. The German Family panel has two additional categories for individuals living alone (Table 3): "Living apart together without children" (L0) and "Living apart together with children" (L1).[ii] Each of the 135,000 samples drawn from the real data is based on a random subsample with replacement of the real data (bootstrap samples) guided by parameters of the scenarios, which determine the sequence sample size, sequence length, and number of states. Figure 2 shows how sequences in the survey data are represented by the scenarios with 3, 6, or 9 states and with varying sequence length (5, 20 or 120 time points).

[TABLE 3 ABOUT HERE]

[FIGURE 2 ABOUT HERE]

5. Results

For the simulated data we know the true number of clusters and can easily gauge how combinations of the three types of data challenges affect the identification of the true number of clusters. For the real data, the true number of clusters is unknown. We therefore focus on the optimal number of clusters, which we define as the number of clusters that are identified under the most optimal data conditions, i.e., data without any of the three types of data challenges present. Based on our results

with the simulated data this is likely to match the true number of clusters. The optimal number of groups in the real data according to this procedure is five clusters that also are substantively and theoretically meaningful comprising: Late married with children (35 pct.), cohabitation with children (10 pct.), extended singlehood with no children (22 pct.), early married one child (11 pct.), and early married with two or more children (24 pct.). For graphic illustration see Appendix Figure A1.

We first classify the 540,000 samples from the simulated data and the 135,000 samples from the real data according to the three data challenges: Too few sequences, too few time points, and too few distinct states (overview in Table 4). Concerning the challenge of small sequence sample size, we specify three variations: 0 = none/samples with at least 1000 sequences; 1 = modest/samples with only 500 sequences; and 2 = severe/samples with only 250 sequences. In terms of sequence length, we consider: 0 = none/more than 20 time points; 1 = modest/ten time points; and 2 = severe/only five time points. Regarding the number of distinct states, we distinguish three challenges: 0 = none/samples that include at least nine distinct states, 1 = modest/samples with six distinct states; and 2 = severe/samples with just three states. Taken together, this results in a total of ($3 \times 3 \times 3 =$) 27 combinations of data challenges.

[TABLE 4 ABOUT HERE]

4.1 How the data challenges affect the identification of cluster solutions

We start by determining how many clusters are found under the most optimal circumstances, that is, when there are no data challenges, i.e., a high number of sequences (at least 1,000 sequences), long sequence chains (at least 20 time points), and a high number of distinct sequence states (at least 9

different states). For each of the 27 combinations of data challenges, we calculate the average mean of the cluster solution with the highest ASW values. Figure 3 shows how the data challenges affect the identification of the optimal number of clusters.

[FIGURE 3 ABOUT HERE]

The y-axis in Figure 3 indicates the average number of clusters found in the samples categorized under the 27 combinations of data challenges. The x-axis indicates which specific data challenge combinations are given in the respective samples. The data challenges are represented with three digits (first digit refers to limitations in the number of sequences, second digit refers to limitations in the sequence length, and third digit refers to limitations in the number of distinct sequence states). To ease the interpretation of the graphs, the data challenges on the x-axis are order in terms of their average suggested number of clusters going from the lowest to the highest average. The red lines in Figure 3 indicate the optimal number of clusters, which is identified by number of clusters when there are no data challenges, i.e., at 0-0-0. The blue dashed lines illustrate cluster solutions that are at most one cluster apart from the optimum, which indicate only a minor deviation from the optimal number of clusters. See Appendix Table A3 for full information on the cluster fits for all combinations of the data challenges for the simulated and the real data.

Figure 3 shows that for the simulated data with three true clusters 18 out of 27 data scenarios (67 percent) suggest cluster numbers in an acceptable range with only one cluster deviation from the true cluster number. For simulated data with six true clusters only 12 out of 27 scenarios (44 percent) are within and acceptable range. For the most complex simulated data structure with nine true clusters a mere 6 out of 27 data scenarios (22 percent) are within an

acceptable range. This resembles the results for the real data where 7 out of 27 data scenarios (25 percent) are within an acceptable range. In all simulated scenarios most samples that only have one modest data challenge are within an acceptable range, that is, either identify the true number of clusters or only one cluster more or less. However, if one of the data challenges is severe (e.g., sample size with only 250 sequences or sequences with only 5 time points), or if a modest data challenge combines with another modest data challenge, deviation from the optimal number of clusters substantially increase.

Consequently, when researchers theoretically assume that a higher number of clusters exist, the required data conditions to correctly identify this cluster structure increase. Even the combination fo modest data challenges severely compromises the correct identification of clusters with a higher likelihood to overestimate rather than underestimate the "true" number of groups. In both the simulated and real data, if the data challenge only concerns the number of sequences, the suggested cluster solutions are still within an acceptable range. Thus, having many sequences is not necessarily enough to compensate for short sequence length or too few distinct states. Findings thereby highlight that there are no simple rules of thumb on the minimum number of sequences, time points or sequence states. Instead the combination of all three data challenges matter jointly with the true number of groups present in the data, which can only be approximated based on theoretical considerations in real world sequence analysis applications.

4.2 Which data challenges are most consequential?

Simple linear probability models based on ordinary least squares (OLS) estimate how much each of the data challenges contributes to deviations from the true or optimal number of clusters (Table 5). For both the simulated and the real data, we estimate an 1) *acceptable range model*, where the

outcome is, whether the average suggested cluster solutions are within +/- one cluster from the optimal solution, and 2) and *exact model* in which the outcome is the probability to identify the exact optimal number of clusters. For both models, key independent variables are the number of sequences, number of time points, and number of distinct states in the alphabet in the data scenario samples. Table 5 shows results with standardized coefficients where coefficient sizes can be interpreted as differences in percentage points.

[TABLE 5 ABOUT HERE]

Across the two models, large sample sizes are not sufficient to avoid deviation from the optimal number of clusters. For example, even for the simplest scenario for the simulated data, the coefficient in the *acceptable range model (1)* only changes from 0.152 to 0.155 as we move from 1,000 sequences in the sample to 2,000 sequences. For the *acceptable range model (1)* based on real data, we find only a modest increase once sequence samples contain at least 500 sequences. As we more from 250 sequences to the 500 sequences, the coefficient only changes from 0.146 to 0.216. Regarding the number of time points, longer sequence length is associated with a more precise identification of the optimal number of clusters. Effect sizes for sequence length in both the *acceptable range model (1)* and the *exact model (2)* are substantially larger than for the number of sequences. Consequently, a high number of time points is more conducive to reliably identifying a "true" cluster typology than a high number of sequences, at least beyond a threshold of at least 500 sequences. Researchers should therefore choose long observation windows and narrower time intervals with smaller sequence sample sizes over datasets with shorter time spans, fewer time intervals, and higher sequence sample size.

For the number of distinct states in the alphabet, we find different patterns for the two models and for the different data types. For the *acceptable range model (1)* and the simulated data, the number of distinct states aids the identification of the optimal number of clusters most when the number of states exceeds the optimal number of clusters. However, for the real data, effects for the number of states are all negative in the *acceptable range model (1),* indicating that a higher number of distinct states decreases the probability of identifying cluster numbers that only deviate by +/- 1 cluster from the optimal number. As discussed above (see Table 3), this discrepancy between simulated and real data might be due to the introduction of states with few observations in the alphabet for the real data (specifically, living apart together with children). Consequently, not only the number of states, but their empirical distribution likely matters for the identification of cluster typologies. Yet, for the real data in the *exact model (2*), we find a similar pattern to the simulated data in which larger alphabets significantly improved the identification of the optimal number of clusters. For example, even in the simple simulated data with just 3 clusters, having an alphabet of 12 relative to 9 states increases the probability of identifying the exact optimal number of clusters by 17 percent (model 2, 0.619 vs. 0.527). Note that in the simulated data all states were empirically fairly equally distributed across the entire sample and important in the mixture Markov modeling. Thus, our results do not imply that adding theoretically meaningless or highly infrequent states would improve the cluster identification, most probably the opposite would be the case. A higher number of distinct states thus generally seems to aid the identification of the optimal number of clusters only if these states are reasonably frequent and theoretically relevant.

For both simulated and real data, we conclude that the most reliable identification of a "true" or optimal sequence typology that empirically exists in a given data structure is possible with samples that have at least 500 sequences, at least a length of 10 time points, and a state alphabet that has at least as many distinct states as the optimal number of clusters. Since the optimal number of

clusters is always unknown and can only be theoretically approximated in real life applications, our analysis emphasizes the need to work iteratively—even for inductive methods as sequence analysis, and underline the significance of theoretical reasoning about the expected number of groups for specifying the state alphabet.

5. Discussion

Cluster-based sequence typologies have proliferated in life course research and social demography. In the past two decades, statistical cluster cut-off criteria and clustering procedures have greatly improved (Studer 2013, Piccarreta and Studer 2018, Studer 2021), along with a much better understanding of the difference that different distance measures in sequence analysis make (Ritschard and Studer 2016). In contrast, to date we have very little systematic understanding how features of the data, that any applied sequence analyst has to make decisions about, systematically affect the identification of cluster-based sequence typologies.

In this paper, we systematically assessed the ability of a given clustering procedure to identify the true or optimal number of clusters in a range of scenarios combining three types of data challenges: the number of sequences, number of time points, and the number of distinct states in the alphabet. To this end, we compared data scenarios drawn from simulated sequence data from mixture Markov models and to real family sequence data from the German Family Panel Survey.

We conclude that researchers can be relatively confident about their sequence typology, if they have at least 500 sequences, and at least 10 time points. However, this finding should be read with some caution. The number of time points in sequence analysis is very much context dependent and the three data challenges act in combination with each other. Nevertheless,

life course researchers should be careful in making conclusions about sequence typologies, if their sequences are shorter than 10 time points, regardless of whether they have a large sample of sequences and only a limited number of sequence states. There might be applications with extreme distributions of states within sequences, where reliable typologies with fewer time points can still be identified. But social sequence data that tends to have a similar structure to the simulated and real data presented in this paper do not seem to enable the identification of "true" or optimal typologies with only five time points, even if these clearly exist in the data. Notably, DNA sequence in biology for which optimal matching algorithms were originally developed (Abbott and Tsay 2000; Levenshtein 1965) tend to be much longer with hundreds and thousands of sequence elements, stressing that having long sequences might be an important feature for data to be appropriate for meaningful sequence analysis. We also find that the state alphabet needs to have at least as many unique states as the optimal number of clusters. As the optimal number of clusters is always unknown in real data, researchers using sequence analysis should employ an iterative approach comparing different numbers of states and deriving precise theoretical expectations about the expected typology, when possible.

Our analysis is not without limitations. First of all, simulated data are simplistic in comparison to the complexity of most real-life sequence data. Nevertheless, using real data, we find a number of similarities to the simulated data in how the different types of data challenges are associated with deviations from identifying the optimal number of clusters. Second, in smaller sequence samples, statistical measures such as ASW cannot give reliable answers to which number of clusters best captures the underlying structure. This resonates with the fact that the ASW alone cannot capture how meaningful a given cluster solution is theoretically (as argued by Studer 2013). Third, we disregarded variation in the empirical prevalence of different states in the sequences, i.e., entropy. Subsequent studies could benefit from assessing how not only the number of distinct states

but especially their empirical distribution impacts the identification of cluster-based sequence typologies.

Our findings highlight that for identifying meaningful and reliable sequence typologies, basic data features might be just as or even more important than the choice of a distance measure and clustering algorithm—even though the latter two have received disproportionately more attention in both the critique and further development of sequence analysis (Studer and Ritschard 2016; Wu 2000; Levine 2000). At the same time, our study shows that under favorable data scenarios, sequence and cluster analysis identify a correct number of clusters with a very high reliability. We contribute to the literature with to the best of our knowledge the first systematic assessment of the sensitivity of cluster-based sequence typologies to key data issues: the number of sequences, number of time points, and number of distinct sequence states.

REFERENCES

Abbott, Andrew 1990. "A Primer on Sequence Methods. *Organization Science* 1(4): 375-392.

Abbott, Andrew 2001. *Time matters: On Theory and Method*. Chicago: University of Chicago Press.

Abbott, Andrew and Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods & Research* 29(1):3-33.

Aisenbrey, Silke and Anette E. Fasang. 2010. "New Life for Old Ideas: The Second Wave of Sequence Analysis. Bringing the Course back into the Life Course." *Sociological Methods & Research* 38(3): 420-462.

Aisenbrey, Silke and Anette E. Fasang. 2017. "The Interplay of Work and Family Trajectories over the Life Course: Germany and the United States in Comparison." *American Journal of Sociology* 122(5):1448-1484.

Andrade, Stefan B. 2016. "Transition and Adaptation: An Analysis of Adaption Strategies amongst Danish Farm Families from 1980–2008." *Sociologia Ruralis* 56(3):371-390.

Barban, Nicola and Francesco C. Billari. 2012. "Classifying Life Course Trajectories: A Comparison of Latent Class and Sequence Analysis." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61(5):765-784.

Billari, Francesco C. 2001a. "The Analysis of Early Life Courses: Complex Descriptions of the Transition to Adulthood." *Journal of Population Research*, *18*(2):119-142.

Billari, Francesco C. 2001b. "Sequence Analysis in Demographic Research." *Canadian Studies in Population*, 439-458.

Brzinsky-Fay, Christian and Ulrich Kohler. 2010. “New Developments in Sequence Analysis.” *Sociological Methods & Research* 38(3):359–364.

Brüderl, Josef, Sonja Drobnič, Karsten Hank, Bernhard Nauck, Franz J. Neyer, Sabine Walper and Philipp Alt. 2019. *The German Family Panel (Pairfam). Za5678 Data File Version 10.0.0*. GESIS Data Archive. https://doi.org/10.4232/pairfam.5678.10.0.0.

Buchmann, Marlies C. and Irene Kriesi. 2011. “Transition to Adulthood in Europe.” *Annual Review of Sociology* 37:481-503.

Bühlmann, Felix. 2010. “Routes into the British Service Class: Feeder Logics According to Gender and Occupational Groups.” *Sociology* 44(2):195-212.

Bühlmann, Felix. 2018. “Trajectories of Vulnerability: A Sequence-analytical Approach.” In: *Social Dynamics in Swiss Society:*129-144. Cham: Springer.

Cornwell, Benjamin. 2015. *Social Sequence Analysis: Methods and Applications*. Cambridge University Press.

Elzinga, Cees. H. and Aart C. Liefbroer. 2007. “De-standardization of Family-life Trajectories of Young Adults: A Cross-national Comparison using Sequence Analysis.” *European Journal of Population/Revue européenne de Démographie 23*(3):225-250.

Emery, Kevin and André Berchtold. 2022. “Comparison of Two Approaches in Multichannel Sequence Analysis using the Swiss Household Panel.” *Longitudinal and Life Course Studies:*1-32.

Fasang, Anette E., Stefan Bastholm Andrade, Selçük Bedük, Zafer Büyükkeçeci and Aleksi Karhula. 2022. ”Lives in Welfare States: Life Courses and Accumulated Earnings at Mid-life in four European Countries.” *Working Paper.*

Gabadinho, Alexis, Gilbert Ritschard, Matthias Studer and Nicolas S. Müller. 2009a. "*Mining Sequence Data in R with the TraMineR Package: A Users Guide for Version 1.2*." Geneva: University of Geneva.

Gabadinho, Alexis, Gilbert Ritschard, Matthias Studer and Nicolas S. Müller. 2009b. "Extracting and Rendering Representative Sequences." In: *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*:94-106. Berlin, Heidelberg: Springer.

Hart, Rannveig K. 2019. "Union Histories of Dissolution: What can they say about Childlessness?" *European Journal of Population. Revue Europeenne de Demographie* 35(1):101.

Helske, Satu and Joni Helske. 2019. "Mixture Hidden Markov Models for Sequence Data: The seqHMM Package in R." *Journal of Statistical Software* 88(3):1-32.

Helske, Satu, Joni Helske and Guikherme K. Chihaya. 2021. "From Sequences to Variables – Rethinking the Relationship between Sequences and Outcomes. https://doi.org/10.31235/osf.io/srxag

Huinink, Johannes, Josef Brüderl, Bernhard Nauck, Sabine Walper, Laura Castiglioni and Michael Feldhaus. 2011. "Panel Analysis of Intimate Relationships and Family Dynamics (Pairfam): Conceptual Framework and Design." *Zeitschrift Für Familienforschung* 23(1):77–101.

Jalovaara, Marika and Anette E. Fasang. 2020. „Family Life Courses, Gender, and Mid-life Earnings." *European Sociological Review* 36(2):159-178.

Kaufman, Leonard and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* New Jersey: John Wiley & Sons Inc.

Karhula, Aleksi, Jani Erola, Marcel Raab and Anette E. Fasang. 2019. Destination as a process: Sibling similarity in early socioeconomic trajectories. *Advances in Life Course Research*, *40*, 85-98.

Lesnard, Laurent. 2006. "Optimal Matching and Social Sciences." *Manuscript. Observatoire Sociologique du Changement*, Paris.

Levine, John H. 2000. "But What Have You Done for us Lately? Commentary on Abbott and Tsay." *Sociological Methods & Research* 29(1): 34-40.

Levenshtein, Vladimir I. 1965. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady,* 10(8):707-710. [Translated from *Doklady Akademii Nauk SSSR* 163(4):845-848, August 1965]

Liao, Tim F. 2021. "Using Sequence Analysis to Quantify How Strongly Life Courses Are Linked." *Sociological Science* 8:48-72.

Liao, Tim F., Danilo Bolano, Christian Brzinsky-Fay, Benjamin Cornwell, Anette E. Fasang, Satu Helske, Raffaella Piccarreta, Marcel Raab and Emanuela Struffolino. 2022. "Sequence Analysis: Its Past, Present, and Future." *Social Science Research* 107: 102772.

Lieberson, Stanley. 1991. "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases." *Social Forces* 70(2):307-320.

Mikolai, Julia and Mark Lyons-Amos. 2017. "Longitudinal Methods for Life Course Research: A Comparison of Sequence Analysis, Latent Class Growth Models, and Multi-state Event History Models for Studying Partnership Transitions." *Longitudinal and Life Course Studies* 8(2):191-208.

Piccarreta, Raffaella and Francesco C. Billari. 2007. “Clustering Work and Family Trajectories by using a Divisive Algorithm.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170(4):1061-1078.

Piccarreta, Raffaella and Emanuela Struffolino. 2019. “An Integrated Heuristic for Validation in Sequence Analysis.” preprint. *SocArXiv*.

Raab, Marcel, Anette Eva Fasang, Aleksi Karhula and Jani Erola. 2014. “Sibling Similarity in Family Formation.” *Demography* 51(6):2127-2154.

Raab, Marcel and Emanuela Struffolino. 2022. *Sequence Analysis*. SAGE Publications.

Ritschard, Gilbert. 2021. “Measuring the Nature of Individual Sequences.” *Sociological Methods & Research* 27*:* 00491241211036156.

Ritschard, Gilbert and Matthias Studer. 2018. “Sequence Analysis: Where Are We, Where Are We Going?” In: *Sequence Analysis and Related Approaches:*1-11. Springer.

Rousseeuw, Peter. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20:53-65.

Shalizi, Cosma. 2009. “Distances between Clustering, Hierarchical Clustering.” *Lectures notes*, Carnegie Mellon University.

Studer, Matthias. 2013. “Weighted Cluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R.” *LIVES Working Papers* 24. Geneva, Switzerland: University of Geneva Institute for Demographic and Life Course Studies.

Studer, Matthias. 2021. “Validating Sequence Analysis Typologies Using Parametric Bootstrap. *Sociological Methodology* 51(2):290-318.

Studer, Matthias and Gilbert Ritschard. 2016. “What Matters in Differences between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures.” *Journal of the Royal Statistical Society Series A* 179(2):481–511.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho and Nicolas S. Müller. 2010. „Discrepancy Analysis of Complex Objects using Dissimilarities.” In F. Guillet, G. Ritschard, D. A. Zighed, and H. Briand (Eds.), *Advances in Knowledge Discovery and Management, Studies in Computational Intelligence* (292):3-19.

Struffolino, Emanuela, Laura Bernardi and Ornella Larenza. 2020. “Lone Mothers’ Employment Trajectories: A Longitudinal Mixed-method Study.” *Comparative Population Studies* 45:265-198.

Ward, Joe H. 1963. “Hierarchical grouping to optimize an objective function.” *Journal of American Statistical Association* 58:236-244.

Warren, John R., Liying Luo, Andrew Halpern-Manners, Jim M. Raymo and Albert Palloni. 2015. “Do Different Methods for Modeling Age-graded Trajectories Yield Consistent and Valid Results?” *American Journal of Sociology* 120(6):1809-1856.

Wu, Lawrence L. 2000. “Some Comments on “Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect”. *Sociological Methods & Research* 29(1):41-64.

TABLES AND FIGURES

Table 1 Overview of the 360 scenarios for the simulated data and the 90 scenarios of the real data

	Values	**Simulated data**		**Real data**
a. Number of sequences	100	✔		✔
	250	✔		✔
	500	✔		✔
	750	✔		✔
	1000	✔		✔
	1500	-		✔
	2000	✔		-
b. Sequences length (time points)	5	✔		✔
	10	✔		✔
	20	✔		✔
	40	✔		✔
	120	✔		✔
c. Number of states in the alphabet	3	✔		✔
	6	✔		✔
	9	✔		✔
	12	✔		-
d. Number of true clusters	3	✔		*Unknown*
	6	✔		*Unknown*
	9	✔		*Unknown*
Scenarios in total		360		90
Total samples of scenarios		540,000		135,000

Table 2. Number of distinct states in simulated data from Mixture Markov Models

<table>
<tr><th colspan="4">Number of distinct states in the alphabet</th></tr>
<tr><th>3</th><th>6</th><th>9</th><th>12</th></tr>
<tr><td rowspan="4">Single
(S)</td><td>Single without children
(S0)</td><td>Single without children
(S0)</td><td>Single without children
(S0)</td></tr>
<tr><td rowspan="3">Single with children
(S1)</td><td>Single with one child
(S1)</td><td>Single with one child
(S1)</td></tr>
<tr><td rowspan="2">Single with more than one child
(S2)</td><td>Single with two children
(S2)</td></tr>
<tr><td>Single with three or more children
(S3)</td></tr>
<tr><td rowspan="4">Cohabitation (C)</td><td>Cohabitation without children
(C0)</td><td>Cohabitation without children
(C0)</td><td>Cohabitation without children
(C0)</td></tr>
<tr><td rowspan="3">Cohabitation with children (C1)</td><td>Cohabitation with one child
(C1)</td><td>Cohabitation with one child
(C1)</td></tr>
<tr><td rowspan="2">Cohabitation with more than one child
(C2)</td><td>Cohabitation with two children
(C2)</td></tr>
<tr><td>Cohabitation with three or more children
(C3)</td></tr>
<tr><td rowspan="4">Married
(M)</td><td>Married without children
(M0)</td><td>Married without children
(M0)</td><td>Married without children
(M0)</td></tr>
<tr><td rowspan="3">Married with children
(M1)</td><td>Married with one child (M1)</td><td>Married with one child (M1)</td></tr>
<tr><td rowspan="2">Married with more than one child
(M2)</td><td>Married with two children
(M2)</td></tr>
<tr><td>Married with three or more children
(M3)</td></tr>
</table>

Table 3. Number of distinct states in the real data from the German Family Panel Survey

Number of distinct states in the alphabet		
3	**6**	**9**
Single (S)	Single without children (S0)	Single without children (S0)
		Living together apart without children (L0)
	Single with children (S1)	Single with children (S1)
		Living together apart with children (L1)
Cohabitation (C)	Cohabitation without children (C0)	Cohabitation without children (C0)
	Cohabitation with children (C1)	Cohabitation with one or more children (C1)
Married (M)	Married without children (M0)	Married without children (M0)
	Married with children (M1)	Married with one child (M1)
		Married with two or more children (M2)

Table 4. Classification of the 360 scenarios from the 540,000 samples of the simulated data and the 90 scenarios from the 135,000 samples of the real data from German Family Panel survey into combinations of three types of data challenges (3 × 3 × 3 = 27 data challenges in total).

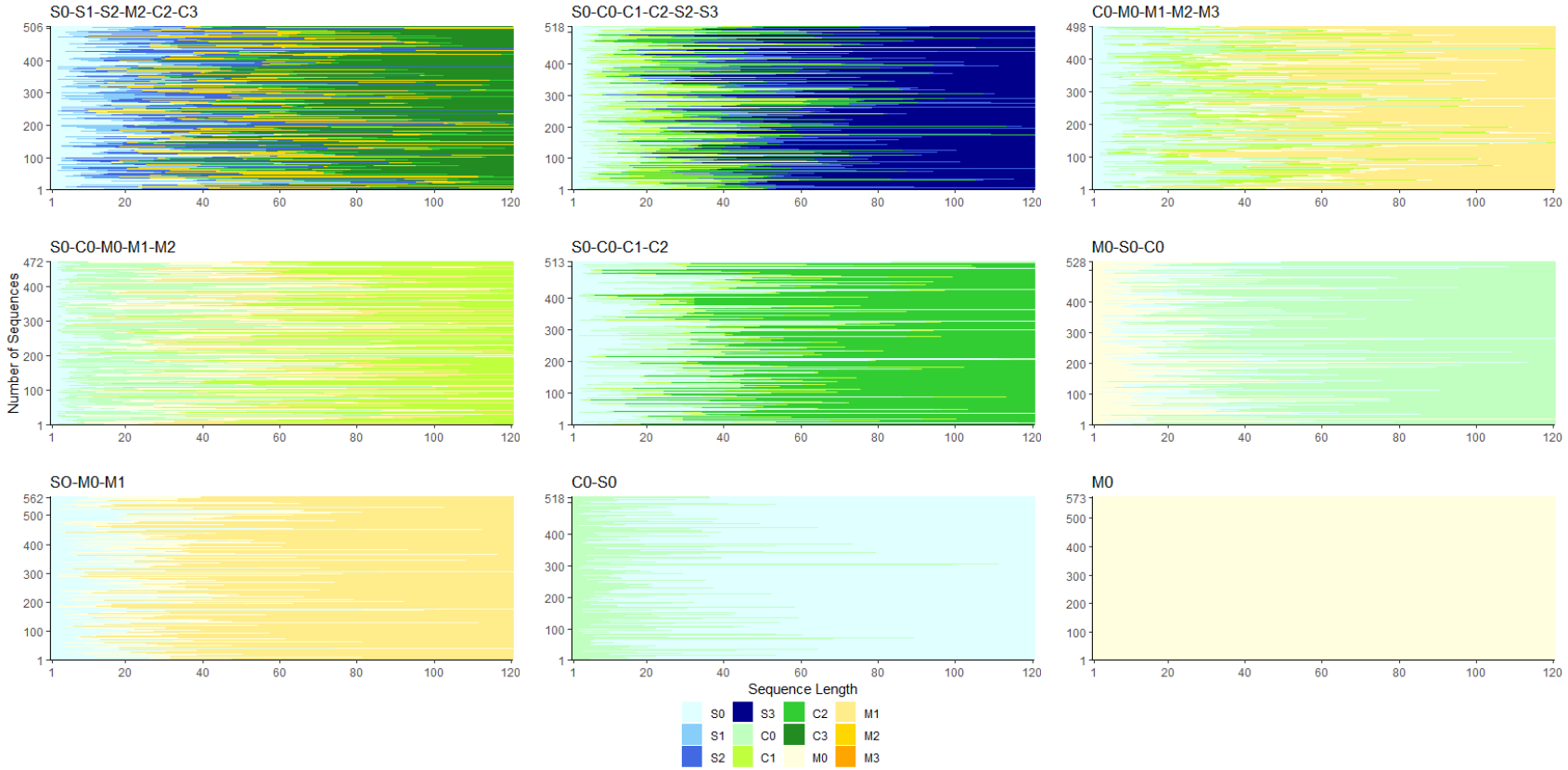
Code	**Data challenge**		
	Number of sequences	**Number of distinct states in the alphabet**	**Number of time points**
0 (none)	At least 1000 sequences	At least 9 distinct states	More than 20 time points
1 (modest)	500 sequences	6 states	10 time points
2 (severe)	250 sequences	3 states	5 time points

Table 5. Result of two linear probability models. *Acceptable Range Model 1*: average cluster solution for samples are within an acceptable range, i.e., at most one cluster more or less than the optimal number. *Exact Model 2*: average cluster solution for samples are exactly the optimal number of clusters. Standardized coefficients.

	SIMULATED						**REAL DATA**	
	3 clusters (simple)		**6 clusters**		**9 clusters (complex)**			
Observations	**Model 1: Fit within +/-**	**Model 2: Perfect fit**	**Model 1: Fit within +/-**	**Model 2: Perfect fit**	**Model 1: Fit within +/-**	**Model 2: Perfect fit**	**Model 1: Fit within +/-**	**Model 2: Perfect fit**
100	Ref.		Ref.		Ref.	Ref.	Ref.	Ref.
250	0.120***	0.102***	0.107***	0.136***	0.060***	0.052***	0.146***	0.052***
500	0.142***	0.116***	0.137***	0.153***	0.081***	0.087***	0.216***	0.074***
750	0.148***	0.119***	0.146***	0.152***	0.085***	0.098***	0.238***	0.085***
1000	0.152***	0.120***	0.152***	0.149***	0.086***	0.106***	0.256***	0.097***
1500/2000	0.155***	0.121***	0.159***	0.146***	0.088***	0.122***	0.272***	0.127***
Sequence length								
5	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
10	0.379***	0.476***	0.296***	0.538	0.272***	0.180***	0.213***	0.116***
20	0.612***	0.630***	0.298***	0.540	0.265***	0.186***	0.374***	0.107***
40	0.620***	0.153***	0.295***	0.536	0.263***	0.190***	0.384***	0.138***
120	0.615***	0.478***	0.291***	0.530	0.269***	0.200***	0.414***	0.231***
States in alphabet								
3	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
6	0.416***	0.503***	0.713***	0.642***	0.016***	0.001**	-0.254***	-0.000
9	0.247***	0.527***	0.909***	0.752***	0.560***	0.077***	-0.039***	0.255***
12	0.375***	0.619***	0.823***	0.636***	0.672***	0.656***	-	-
R^2	0.524	0.620	0.741	0.740	0.622	0.453	0.269	0.111
Samples	180.000		180.000		180.000		135,000	

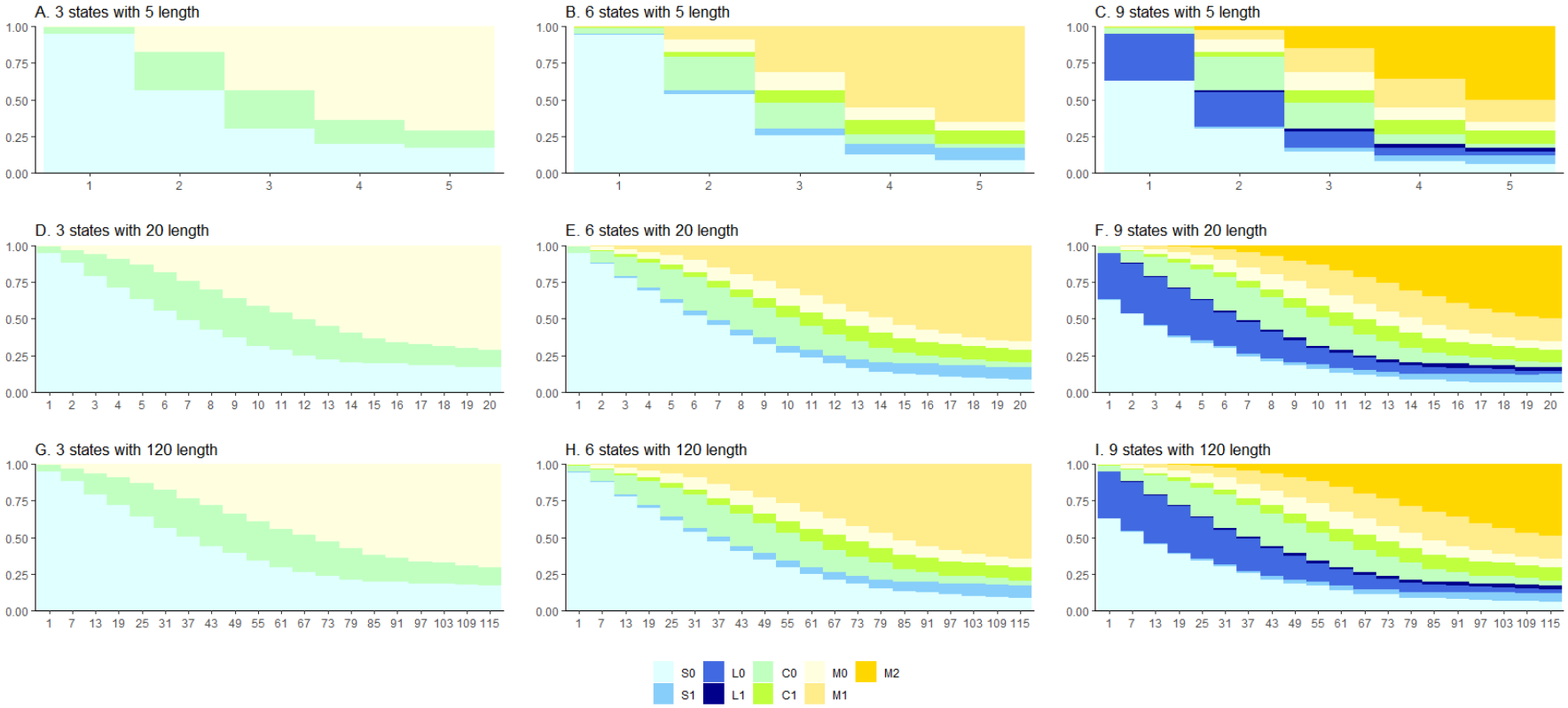
*** p < 0.0001; ** p < 0.001; * p < 0.01

Figure 1. Mixture Markov Model simulated sequences for the most complex scenario with 12 states, 9 clusters, and sequences of length 120. The groups represent 9 clusters.



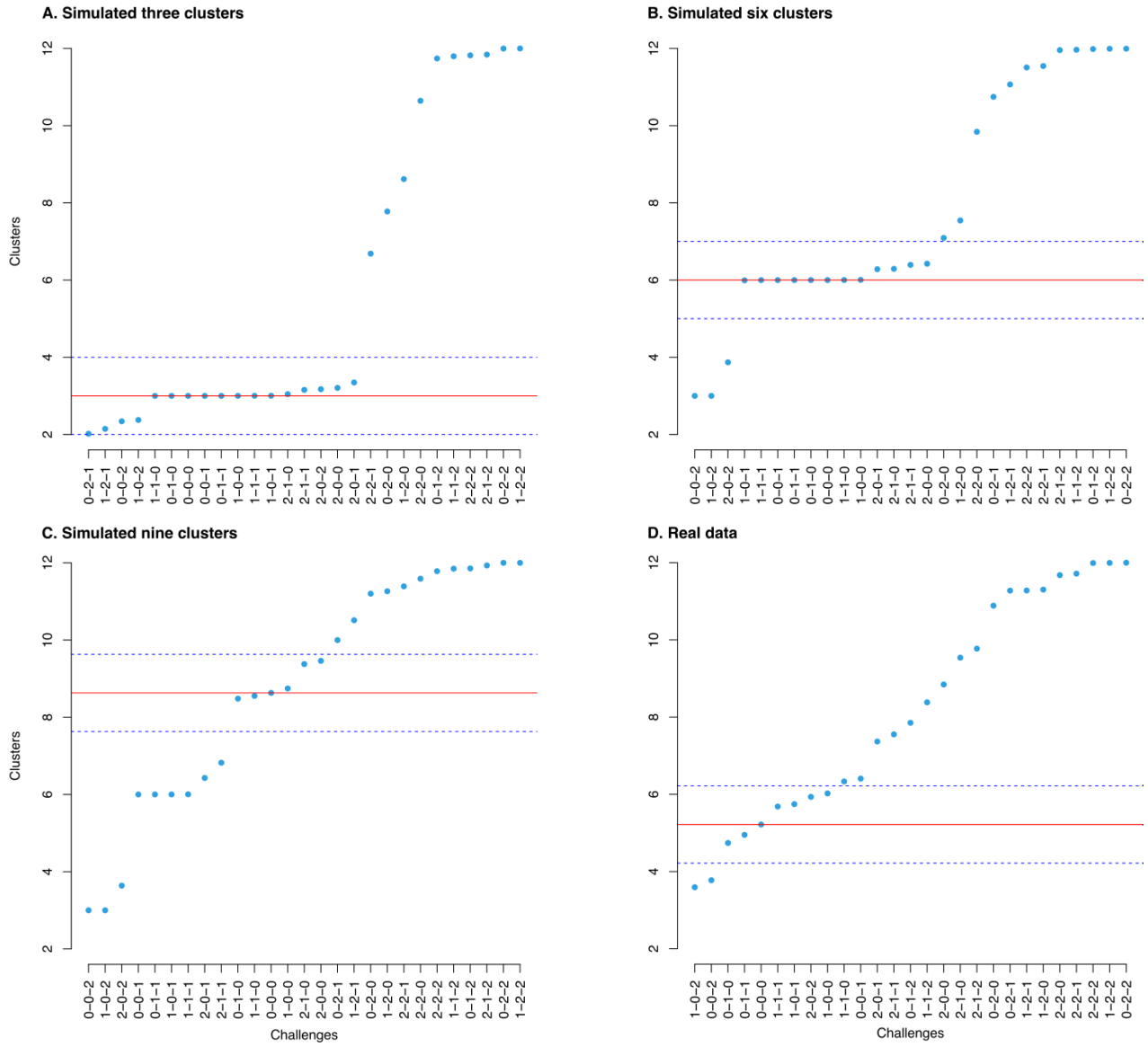
Note: S0 = Single, 0 children; S1 = Single, 1 child; S2 = Single, 2 children; S3 = Single, 3 children; C0 = Cohabitation, 0 children; C1 = Cohabitation, 1 child; C2 = Cohabitation, 2 children; C3 = Cohabitation, 3 children; M0 = Marriage, 0 children; M1 = Marriage, 1 child; M2 = Marriage, 2 children; M3 = Marriage, 3 children.

Figure 2. State distribution plots of 9 data scenarios of the full sample of the German Family Panel with varying number of distinct states and sequence length. The groups each represent the full sample in a different data specification.



Note: S0 = Single, no children; S1 = Single with children; C0 = Cohabitation, no children; C1 = Cohabitation with children; M0 = Marriage, no children; M1 = Marriage, one child; M2 = Marriage with two or more children.

Figure 3. Variance in cluster identification across the 27 scenarios of data challenges.



Note: On the x-axis, the digits read as follows: *The first digit* refers to limitations in the number of sequences: 0 = no limitations/at least 1,000 sequences; 1 = modest limitations/500 sequences; 2 = severe limitations/only 250 sequences. *The second digit* refers to limitations in the sequence length: 0 = no limitations in sequence length/at least 20 states long; 1 = modest limitations/10 states long; 2 = severe limitations/only 5 states long. *The third digit* refers to limitations in the number of unique states: 0 = no limitations/at least 9 unique states; 1 = modest limitations/6 unique states; 2 = severe limitations/only 3 unique states.

Red horizontal lines indicated the optimal cluster solution. Dotted blue lines indicate +/- 1 cluster from the optimal cluster solution, which we consider an acceptable range of deviation.

Appendix

Appendix Table A1. Transition cost matrix for calculating OM Spell distance between sequences

	M	**C**	**0 -> 1 child**	**each additional child**
Single (S)	2.0	2.5	+1	+0.5
Married (M)	0	1.0	+1	+0.5
Cohabiting (C)	2.5	0	+1	+0.5

Note: Even though here the states are artificial in the simulated data, we label them as family formation states and the costs are inspired by theoretical considerations on family formation: the cost is higher to move from S ("single") to M ("marriage") than from S ("single") to C ("cohabitation"). While the first increase in the number (first child) costs 1, subsequent costs for increasing numbers (additional children) are set to 0.5, which corresponds to more significant difference between couples with and without children than between couples with, e.g., two children and three children.

Appendix Table A2. Characteristic order of **s**tates in the clusters in the scenarios of the simulated data.

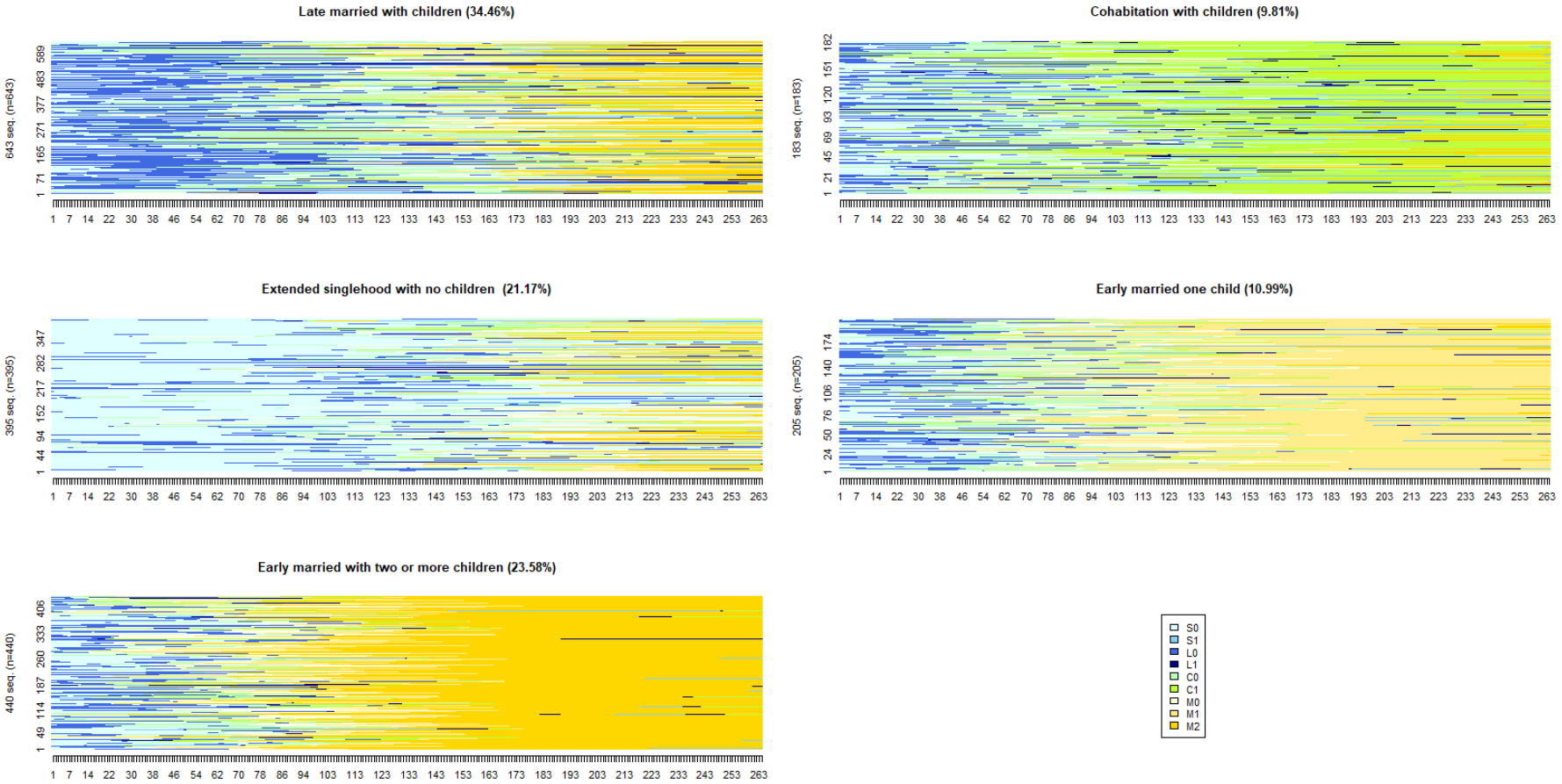
States	Cluster	3 Clusters	6 Clusters	9 Clusters
12	1	S0-S1-S2-M2-C2-C3	S0-S1-S2-M2-C2-C3	S0-S1-S2-M2-C2-C3
	2	S0-C0-C1-C2-S2-S3	S0-C0-C1-C2-S2-S3	S0-C0-C1-C2-S2-S3
	3	C0-M0-M1-M2-M3	C0-M0-M1-M2-M3	C0-M0-M1-M2-M3
	4	-	S0-C0-C1-C2	S0-C0-M0-M1-C1
	5	-	M0-S0- C0	S0-C0-C1-C2
	6	-	C0-S0	M0-S0-C0
	7	-	-	S0-M0-M1
	8	-	-	C0-S0
	9	-	-	M0
9	**Cluster**	**3 Clusters**	**6 Clusters**	**9 Clusters**
	1	S0-S1-S2-M2-C2	S0-S1-S2-M2-C2	S0-S1-S2-M2-C2
	2	S0-C0-C1-C2-S2	S0-C0-C1-C2-S2	S0-C0-C1-C2-S2
	3	C0-M0-M1-M2	C0-M0-M1-M2	C0-M0-M1-M2
	4	-	S0-C0-C1-C2	S0-C0-M0-M1-C1
	5	-	M0-S0- C0	S0-C0-C1-C2
	6	-	C0-S0	M0-S0- C0
	7	-	-	S0-M0-M1
	8	-	-	C0-S0
	9	-	-	M0
6	**Cluster**	**3 Clusters**	**6 Clusters**	**9 Clusters**
	1	S0-S1-M1-C1	S0-S1-M1-C1	S0-S1-M1-C1
	2	S0-C0-C1-S1	S0-C0-C1-S1	S0-C0-C1-S1
	3	C0-M0-M1	C0-M0-M1	C0-M0-M1
	4	-	S0-C0-C1	S0-C0-M0-M1-C1
	5	-	M0-S0- C0	S0-C0-C1
	6	-	C0-S0	M0-S0- C0
	7	-	-	S0-M0-M1
	8	-	-	C0-S0
	9	-	-	M0
3	**Cluster**	**3 Clusters**	**6 Clusters**	**9 Clusters**
	1	S-M-C	S-M-C	S-M-C
	2	S-C-S	S-C-S	S-C-S
	3	C-M	C-M	C-M
	4	-	S-C	S-C-M-C
	5	-	M-S	S-C
	6	-	C-S	M-S
	7	-	-	S-M
	8	-	-	C-S
	9	-	-	M

Appendix Table A3. Overview of the data challenges [*Observations – Number of states – Number of distinctive states*]

Data challenge	Simulations									Real data	
	Three clusters			Six clusters			Nine clusters			Fit	Variance
	Fit	Variance	Success	Fit	Variance	Success	Fit	Variance	Success		
0-0-0	1.000	0.000	1.000	1.000	0.000	1.000	0.939	-0.370	0.554	0.684	0.218
0-0-1	1.000	0.000	1.000	1.000	0.000	1.000	0.000	-3.000	0.000	0.200	1.407
0-0-2	1.000	-0.658	0.342	1.000	-0.658	0.342	0.000	-6.000	0.000	0.817	-1.226
0-1-0	1.000	0.000	1.000	1.000	0.000	1.000	0.943	-0.518	0.520	0.401	-0.260
0-1-1	1.000	0.000	1.000	1.000	0.000	1.000	0.000	-2.999	0.000	0.324	-0.051
0-1-2	0.001	8.740	0.001	0.001	8.740	0.001	0.000	2.855	0.000	0.309	2.852
0-2-0	0.464	4.777	0.458	0.464	4.777	0.458	0.154	2.202	0.068	0.000	5.886
0-2-1	1.000	-0.980	0.013	1.000	-0.980	0.013	0.013	0.999	0.005	0.000	6.276
0-2-2	0.000	8.997	0.000	0.000	8.997	0.000	0.000	3.000	0.000	0.000	6.999
1-0-0	1.000	0.001	0.999	1.000	0.001	0.999	0.914	-0.255	0.495	0.618	1.024
1-0-1	1.000	0.004	0.996	1.000	0.004	0.996	0.000	-2.999	0.000	0.254	0.746
1-0-2	1.000	-0.623	0.375	1.000	-0.623	0.375	0.000	-6.000	0.000	0.664	-1.407
1-1-0	1.000	0.000	1.000	1.000	0.000	1.000	0.924	-0.447	0.459	0.385	1.333
1-1-1	1.000	0.003	0.997	1.000	0.003	0.997	0.001	-2.996	0.000	0.235	0.683
1-1-2	0.002	8.797	0.002	0.002	8.797	0.002	0.001	2.849	0.000	0.161	3.383
1-2-0	0.365	5.616	0.348	0.365	5.616	0.348	0.154	2.264	0.058	0.003	6.303
1-2-1	0.991	-0.853	0.055	0.991	-0.853	0.055	0.049	1.513	0.016	0.000	6.279
1-2-2	0.000	8.999	0.000	0.000	8.999	0.000	0.000	3.000	0.000	0.000	6.994
2-0-0	0.961	0.208	0.897	0.961	0.208	0.897	0.692	0.461	0.305	0.266	3.845
2-0-1	0.935	0.348	0.837	0.935	0.348	0.837	0.045	-2.572	0.016	0.183	2.367
2-0-2	0.893	0.173	0.395	0.893	0.173	0.395	0.019	-5.361	0.007	0.348	0.934
2-1-0	0.993	0.046	0.965	0.993	0.046	0.965	0.700	0.379	0.295	0.134	4.539
2-1-1	0.981	0.156	0.879	0.981	0.156	0.879	0.110	-2.177	0.032	0.109	2.553
2-1-2	0.004	8.841	0.004	0.004	8.841	0.004	0.002	2.932	0.000	0.066	4.774
2-2-0	0.132	7.644	0.122	0.132	7.644	0.122	0.090	2.589	0.026	0.001	6.678
2-2-1	0.526	3.684	0.072	0.526	3.684	0.072	0.068	2.391	0.020	0.000	6.716
2-2-2	0.000	8.821	0.000	0.000	8.821	0.000	0.050	2.784	0.012	0.000	6.990

Fit: Share of samples within the combination of data challenges that is within +/1 cluster of the optimal solution; **Variance:** *Difference between average cluster solution in the* combination of data challenges and the average number in the optimal solution; **Success**: Share of simulated samples within the combination of data challenges that identified the correct number of clusters that was used to create the simulations.

Appendix Figure A1. Sequences for the five clusters in the real data from the German Family Panel (1,866 sequences, 264 time points and 9 states)



Note: S0 = Single, no children; S1 = Single with children; C0 = Cohabitation, no children; C1 = Cohabitation with children; M0 = Marriage, no children; M1 = Marriage, one child; M2 = Marriage with two or more children.

FOOTNOTES

[i] Comparisons to considerably different dissimilarity measures (such as those highly sensitive to timing) would not be meaningful in a similar research design, as they would lead to different types of clusters. Such clusters would not be likely to correspond to the true (simulated) clusters well enough for any direct comparisons to be meaningful.

[ii] Adding the two single categories of living together apart with or without children adds another data challenge of sequence states with few observations. More detailed sequence alphabets can make it more difficult for the cluster algorithm to provide robust identification of clusters. Living together apart without children (L0) represents a significant share of the single category (especially in the early years). Living together apart with children (L1) is at no point in the period larger than a few percent of the sample.