Standard Length Article

Sociological Methodology
1–25
© American Sociological Association 2023
DOI: 10.1177/00811750231177026
http://sm.sagepub.com

From Sequences to Variables: Rethinking the Relationship between Sequences and Outcomes

Satu Helske[1,2], Jouni Helske[3], and Guilherme K. Chihaya[4,5]

Abstract

Sequence analysis is increasingly used in the social sciences for the holistic analysis of life-course and other longitudinal data. The usual approach is to construct sequences, calculate dissimilarities, group similar sequences with cluster analysis, and use cluster membership as a dependent or independent variable in a regression model. This approach may be problematic, as cluster memberships are assumed to be fixed known characteristics of the subjects in subsequent analyses. Furthermore, it is often more reasonable to assume that individual sequences are mixtures of multiple ideal types rather than equal members of some group. Failing to account for uncertain and mixed memberships may lead to wrong conclusions about the nature of the studied relationships. In this article, the authors bring forward and discuss the problems of the "traditional" use of sequence analysis clusters as variables and compare four approaches for creating explanatory variables from sequence dissimilarities using different types of data. The authors conduct simulation and empirical studies, demonstrating the importance of considering how sequences and outcomes are related and the need to adjust analyses accordingly. In many typical social science applications, the traditional approach is prone to result in wrong conclusions, and similarity-based approaches such as representativeness should be preferred.

Keywords

sequence analysis, cluster analysis, typology, representativeness, life-course

Over the past few decades, researchers have become more interested in sequence analysis (SA) for the holistic analysis of life-course and other longitudinal data. The usual approach is to construct sequences, calculate pairwise dissimilarities, and then use a clustering algorithm on the dissimilarities for finding groups of similar sequences. Typically, these clusters are then described and interpreted as typologies. Increasingly, researchers are interested in analyzing the relationships between sequences and other

[1] INVEST Research Flagship Center, University of Turku, Turku, Finland
[2] Department of Social Research, University of Turku, Turku, Finland
[3] Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland
[4] Faculty of Social Sciences, Nord University, Bodø, Norway
[5] Institute for Analytical Sociology, Linköping University, Linköping, Sweden

Corresponding Author:
Satu Helske, University of Turku, Department of Social Research, FI-20014 Turun yliopisto, Finland
Email: satu.helske@utu.fi

characteristics, usually by using cluster membership as a dependent or independent variable in a linear or nonlinear regression model.

Almost unanimously, the clustering methods used in the SA context have been hard or crisp clustering algorithms, such as Ward's method or partitioning around medoids (PAM). These algorithms find a partitioning where each sequence belongs to one cluster and one cluster only, which easily translates into a categorical variable with internally homogeneous and mutually exclusive groups. Applications using cluster membership as an observed characteristic of the units of analysis in regression models are also common (e.g., Chaparro et al. 2017; Fuller 2015). This approach is often problematic because the implicit assumption is that cluster membership is a fixed and known characteristic of an individual (or other subject), even though there is considerable uncertainty in clustering solutions because of various possibilities of choosing (dis)similarity measures, clustering algorithms, and the number of clusters. Furthermore, individual sequences might be mixtures of two or more ideal types or distant from all ideal types, making the whole concept of classification into clear or true clusters problematic. Failing to account for uncertain and mixed memberships may lead to wrong conclusions about the existence and nature of the studied relationships. Our aim is to bring forward and discuss the potential problems of the "traditional" approach of creating variables from SA clusters and to compare alternative options for creating explanatory variables using dissimilarities between sequences.

METHODS

Social scientists have increasingly called attention to how existing methods understate the certainty with which individual cases are allocated to sequence clusters and overstate within-cluster homogeneity, arguing for the need for methodological developments (e.g., Warren et al. 2015). Studer (2013) and Piccarreta and Studer (2019) discussed the problems with linking SA cluster membership and a covariate. By assigning the same cluster membership value to all sequences in the same cluster, we are neglecting the possible within-cluster variation of the sequences. This is not a problem if the structure of the clustering is strong, that is, there are clear subgroups in the data and we can be fairly certain of cluster memberships.

Furthermore, the relationship between the sequences and the outcome of interest should be sufficiently explained by the cluster memberships (we refer to this as a "class-dependent outcome"). This refers to type A in Figure 1: there are two clear clusters and the value of the outcome—indicated by the shade of the dot—depends on the class only, not on the subject's position within the class (all within-class variation is random). A simple example of this situation is when changes in childhood family structure explain educational outcomes, such as when parental separation would have the same kind of effect on all children. In this case, children's position in relation to the clusters (e.g., because of the timing of the separation and possible parental repartnering) would not matter for explaining the relationship between the pattern of childhood family structural changes and later educational outcomes.

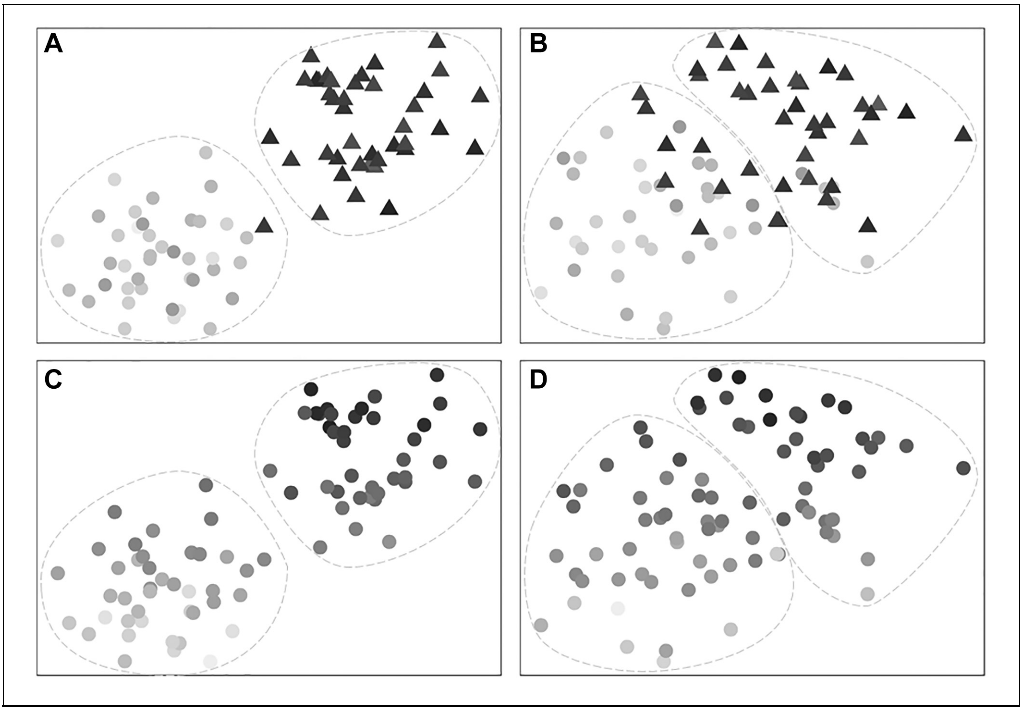

Figure 1. Illustration of four data types on the basis of the strength of the clustering tendency and the type of the sequence–outcome link: (A) strong clustering, class-dependent outcome; (B) weak clustering, class-dependent outcome; (C) strong clustering, similarity-based outcome; and (D) weak clustering, similarity-based outcome.

Note: The points refer to the relative positions of sequences in two-dimensional space. The shade of the points refers to the value of an outcome variable. In Panels A and B, the value of the outcome depends on the class membership (classes differentiated by shape) and the within-class variation is random; in Panels C and D, the value of the outcome depends on the relative positions of the sequences (here, increases along the vertical axis). The dashed lines show a partitioning suggested by a partitioning-around-medoids clustering algorithm.

In all other cases, however, the standard approach is potentially problematic. In a type B situation (Figure 1), the sequence–outcome link is similar to that of type A, but the clusters are overlapping. The weak clustering structure is a problem as it leads to misallocation of sequences. Even if this misallocation is random, this can bias the estimates, as in the analogous case of measurement error in covariates (cf. regression dilution/attenuation; e.g., Berglund 2012), and in some cases failing to account for this classification error can lead to too small standard errors and *p* values, increasing the risk for type 1 error (Bakk, Oberski, and Vermunt 2017; Bakk, Tekle, and Vermunt 2013).

In the social sciences, we argue, it is often unrealistic to assume that any true underlying clusters exist (contrary to, e.g., pattern recognition applications). However, even if true clusters existed, they are difficult to identify using existing methods (Warren et al. 2015) and thus the sequence–outcome link cannot be easily reduced to the relationship between fixed cluster memberships and an outcome. Typically, the sequence typology derived from clustering can be regarded as an imperfect assignment of sequences to categories that approximate different ideal types. In this situation, the outcome depends on how strongly the sequences resemble the ideal types, or how they relate to one another (their relative positions). Illustrations of such data with "similarity-based outcomes" are shown in Panels C and D of Figure 1. A simplified example is the relationship between employment trajectories and lifetime accumulated income. In such a case, accounting for other factors, such as education level, an individual 1 in a long, stable employment career would have, on average, higher accumulated income than individual 2, who never had a stable job. In such a situation, the accumulated income of individual 3, who entered the labor market at a later age and was consistently employed thereafter, would be somewhere in between those of individuals 1 and 2 (again accounting for educational level). Careers more similar to that of individual 1 would tend to have higher incomes, and careers more similar to that of individual 2 would tend to have lower incomes.

In a type C situation, we have a strong clustering structure from which we can easily name some representative or ideal-type sequences (e.g., normative school-to-work trajectories). In a type D situation, there is merely a weak clustering tendency or no clear structure at all, but different types of trajectories are nevertheless related to different levels of the outcome. In this situation, cluster analysis can be used as a tool for finding some representative sequences that help in assessing and interpreting the sequence–outcome relationship. For a general presentation on the differences of uncertain or mixed memberships in clustering crisp or fuzzy data, see, for example, D'Urso (2007).

To date, there are few proposals to account for the uncertainty of the clustering result. Studer (2018) first brought up the idea of using "fuzzy" or "soft" clustering methods to account for mixed cluster memberships of sequence data in cases where sequences are the outcome of interest. In terms of sequences as a predictor (the interest in this article), to account for classification error, Jalovaara and Fasang (2020) conducted robustness checks by excluding cases with poor silhouette values (reflecting a poor fit to their respective cluster; Rousseeuw 1987). In their study, excluding cases with low silhouette values led to relatively small deviations in estimates but a substantial loss of cases and a considerable increase in standard errors of the estimates. In the following sections, we propose and discuss three alternatives to the traditional hard classification approach.

Membership Probability and Representativeness

If we assume we have fixed cluster memberships and class-dependent outcomes, our main goal is to assign individuals to their correct clusters. *Fuzzy* or *soft clustering* or *soft classification* is a form of clustering whereby individuals belong to clusters with a certain probability or degree. Instead of assigning subjects to one cluster and one cluster only, which can effortlessly be turned into an easily interpretable categorical variable, soft classification leads to a membership matrix, which describes the uncertainty in cluster assignments, or the degree or strength of cluster memberships of hybrid members.

A membership matrix is less straightforward to use in a regression model. Studer (2018) proposed using the membership matrix as the outcome in a Dirichlet regression model, but to our knowledge, no one has yet proposed creating explanatory variables from a membership matrix of sequence data. If we look beyond the SA literature, some work in the latent class analysis (LCA) literature has suggested creating independent variables from latent classes where, similar to cluster analysis, true class memberships are unknown. The most interesting approach is the *multiple pseudoclass method*, whereby individuals are randomly assigned to clusters multiple times on the basis of their membership probabilities, and the final estimates are obtained in a similar way to multiple imputation (Bandeen-Roche et al. 1997; Bray, Lanza, and Tan 2015; Lanza, Tan, and Bray 2013; for an application to sociological data, see Ellwardt, Aartsen, and van Tilburg 2017). We apply a similar strategy in the context of SA.

Although likely an improvement over hard classification, typically when using soft classification and pseudoclass methods, researchers still assume that each subject belongs to a single cluster, but the methods account for the uncertainty in the cluster assignments. We argue that this dependence on specific clusters is often unrealistic in the social sciences, as many individual characteristics are continuous in nature and there are an unlimited number of different life-courses instead of fixed categories. If we do not believe in the existence of true clusters, but instead assume the relative positions of the sequences matter more, we need to focus on their (dis)similarities directly. Using pairwise dissimilarities in explaining an outcome is practically impossible, so we turn to the concept of *representativeness*.

In discussing representativeness of sequences, Gabadinho and Ritschard (2013) consider different options, including frequency, neighborhood density, and centrality. Here, centrality considers the distances or dissimilarities between sequences. Centrality can be calculated as the sum of dissimilarities between a subject and all (other) members in a group. The smaller the sum, the more central the subject; the most central subject is called the *medoid*. The closer a medoid is to a sequence or a group of sequences, the better representative it is to them.

As an example of how membership probability and representativeness differ, consider the situation depicted in Figure 2. Subjects M1, M2, and M3 are the medoids, that is, the most central members of their respective clusters. As such, they are the best single representatives to their clusters. We can be fairly certain they belong to their respective clusters; their membership probabilities are high regarding their own clusters and low regarding all other clusters.

Subjects S1 and S2, on the other hand, are distant from the closest medoid M2, so they are much less representative to cluster 2, and medoid M2 is much less representative of them than most of the other members. S1 and S2 are, however, different in their positioning. Subject S1 is of a mixed type, almost equally distant from medoids M2 and M3. Its membership probabilities for clusters 2 and 3 are thus similar, close to 0.5. Subject S2, however, is simply a distant subject: it is distant from medoid M2 but even further away from medoids M1 and M3. Even though it does not fit any cluster

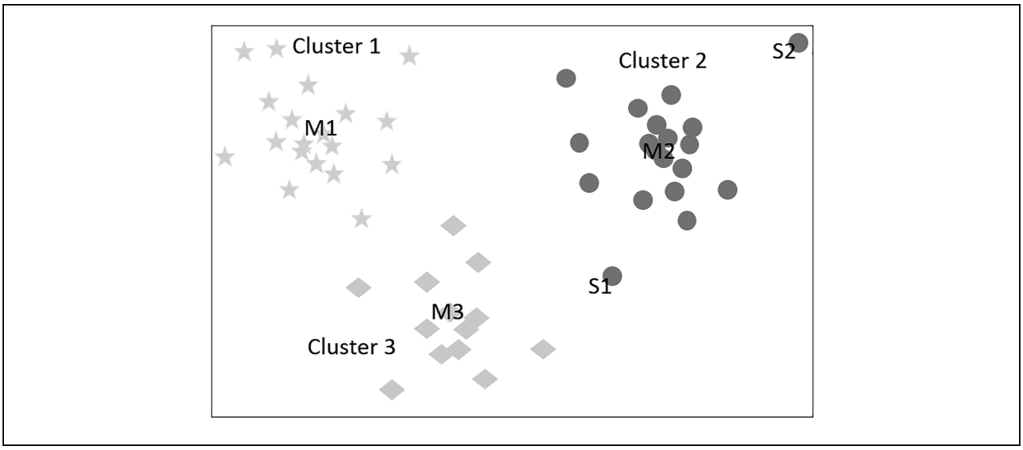

Figure 2. Example clusters with strong representatives (medoids M1, M2, and M3) and two types of weak representatives (S1 and S2).

particularly well, its membership probability to cluster 2 is high, corresponding to strong certainty of being a member of cluster 2. Hence, we see that membership probability itself is not always a good measure of representativeness.

If we are dealing with a type A or type B situation (class-dependent sequence–outcome relationship), the relative position within the cluster and thus subjects' representativeness is not an issue, unless we assume to find subjects that are not members of any clusters (outliers). However, in situations of types C and D, representativeness is arguably more important and often a theoretically more justified approach, as we must consider subjects' positions in relation to others, for example, by comparing them with some theoretical ideal types or medoids.

Creating Variables from Sequences

Table 1 presents different ways of constructing variables from sequences, two of which are based on a crisp clustering algorithm (in this case, the PAM algorithm) and two on a fuzzy clustering algorithm, here the fuzzy analysis (FANNY) algorithm (Kaufman and Rousseeuw 2009).

Let K be the number of clusters obtained from a clustering algorithm. We refer to the traditional approach of constructing a categorical variable with K categories from crisp cluster memberships as *hard classification. Soft classification* refers to using membership probabilities from fuzzy clustering as K continuous variables (which sum to 1 for each subject). In both types of variables, one cluster is typically chosen as a reference, and the respective (dummy or probability) variable is omitted from the model.

Pseudoclass is the equivalent of the multiple pseudoclass technique in probabilistic LCA, where we draw multiple samples of cluster memberships from fuzzy clustering, and for each sample estimate a model with a categorical membership variable the usual way. Finally, we combine the results across the models similarly to the multiple imputation technique (Rubin 2004). This type of an approach is fairly common in the

Table 1. Variable Construction for the Simulation and Empirical Studies Including Two Methods for Crisp Clustering (Using the PAM Algorithm) and Two for Fuzzy Clustering (Using the FANNY Algorithm)

Name	Clustering Method	Variable Construction	Variable Type
Hard classification	Crisp (PAM)	Cluster membership	Dummies
Soft classification	Fuzzy (FANNY)	Membership degree	Continuous
Pseudoclass	Fuzzy (FANNY)	Multiple pseudoclass technique	Dummies
Representativeness	Crisp (PAM)	(Modified) distance to medoids	Continuous

Note: FANNY = fuzzy analysis; PAM = partitioning around medoids.

LCA literature, despite some more recent studies (e.g., Lanza et al. 2013) showing this approach might not provide improvements over hard classification.

Finally, we construct a variable that takes into account *representativeness*. For this technique we need to define a set of representative sequences we can choose, on the basis of theory or from using a clustering algorithm. Here we use a crisp clustering algorithm for finding medoids and calculate the dissimilarity of each sequence to each cluster medoid on the basis of the same distance matrix as for the hard clustering (PAM). We transform these dissimilarities to representativeness values so that value 1 refers to perfect representation and 0 to poorest representation. More specifically, we define the representativeness value of representative *k* (here, the medoid of cluster *k*) to sequence *i* as

$$R_i^k = 1 - \frac{\text{distance of sequence } i \text{ to representative } k}{\text{maximum distance between two sequences}}.$$

This leads to *K* continuous variables for subsequent analysis (which do not sum to 1).

SIMULATION STUDY

In this section we illustrate how different approaches succeed in predicting the outcome when the sequence–outcome relationship is class dependent or similarity based. All analyses were done in the R environment (R Core Team 2021), using packages cluster (Maechler et al. 2021), seqHMM (Helske and Helske 2019), TraMineR (Gabadinho et al. 2011), ggplot2 (Wickham 2016), and dplyr (Wickham et al. 2021). The code to reproduce the simulation experiment and additional analyses can be found on GitHub (https://github.com/helske/seqs2vars).

We first generated sequence data by creating three mixture Markov models with varying clustering tendencies, each with four states and four mixture components ("clusters"). We simulated 10,000 sequences of length 20 from each of these models. We then calculated dissimilarities using optimal matching for spell sequences with constant substitution costs (Studer and Ritschard 2016). We chose this measure because it is sensitive to sequencing and thus is well suited for analyzing data generated with a Markovian model. We then clustered the sequences using PAM and FANNY. Assessed using the average silhouette width (ASW; based on PAM) as a measure of clustering tendency (Kaufman and Rousseeuw 2009), the first model generated sequences with strong clustering tendency (ASW of about 0.8), the second generated sequences with a reasonable clustering tendency with some overlap between sequences from different submodels (ASW of about 0.6), and the third generated sequences with a weak clustering tendency (ASW of about 0.3). Figure 3 shows samples of clustered sequences. Using the clustering solutions and the corresponding dissimilarity matrix, we then created several covariate matrices X on the basis of the methods outlined in the prior section and summarized in Table 1. For each generated matrix X, we then generated a response variable y by

$$y_i = x_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

with $\beta = (0, 1, 1, -1)'$ and $\sigma = 0.25$ for each X.

Using each of these data sets, we ran Monte Carlo simulations in which for each replication we sampled 1,000 of the original sequences and a corresponding y from the full data set and computed covariate matrix X using all the methods presented in Table 1. For FANNY, we fixed the membership exponent to 1.4, as larger values often led to complete fuzziness and convergence issues during the simulations. We then split the data in half and used the first half to estimate the regression model $y_i = x_i\beta + \epsilon_i$. Finally, we used this model to predict the values of the hold-out sample and computed the corresponding root mean squared error (RMSE) of the predictions. This allowed us to estimate the out-of-sample performance of our methods, also taking into account the potential bias and uncertainty stemming from the construction of X using the subsamples instead of the full population data.

In reality, sequence data are unlikely to be generated by such simple Markovian models, and the relationship between sequences and outcome variables is more complex. Thus, the following results reflect more of a best-case scenario; in practice, the differences between the methods and potential errors could be much larger than observed here.

Figure 4 shows the average RMSE and 95th percentile intervals from 10,000 replications for different data-generating models and estimation methods. We see that for classification-based data, the prediction improved (RMSE decreased) when the clustering tendency strengthened. Not surprisingly, the estimation based on hard classification performed best with strong, clear clusters. Soft classification performed, on average, slightly better in cases where we had classification error (data with a reasonable or weak clustering tendency). The hard classification method produced the widest percentile intervals: its performance was the most inconsistent. When the outcome was generated on the basis of membership probabilities, the clustering tendency did not have a strong effect on the average RMSE when using the estimation method that matched the data-generation process (soft clustering, the best-case scenario), whereas other methods performed best with stronger clustering tendency.

On the other hand, when the data were generated on the basis of representativeness (the case we argue is typically the most realistic in social sciences), the clustering tendency did not have a clear effect on the average RMSE for any of the methods, and all methods produced results not far from the theoretical value of 0.25 (the standard

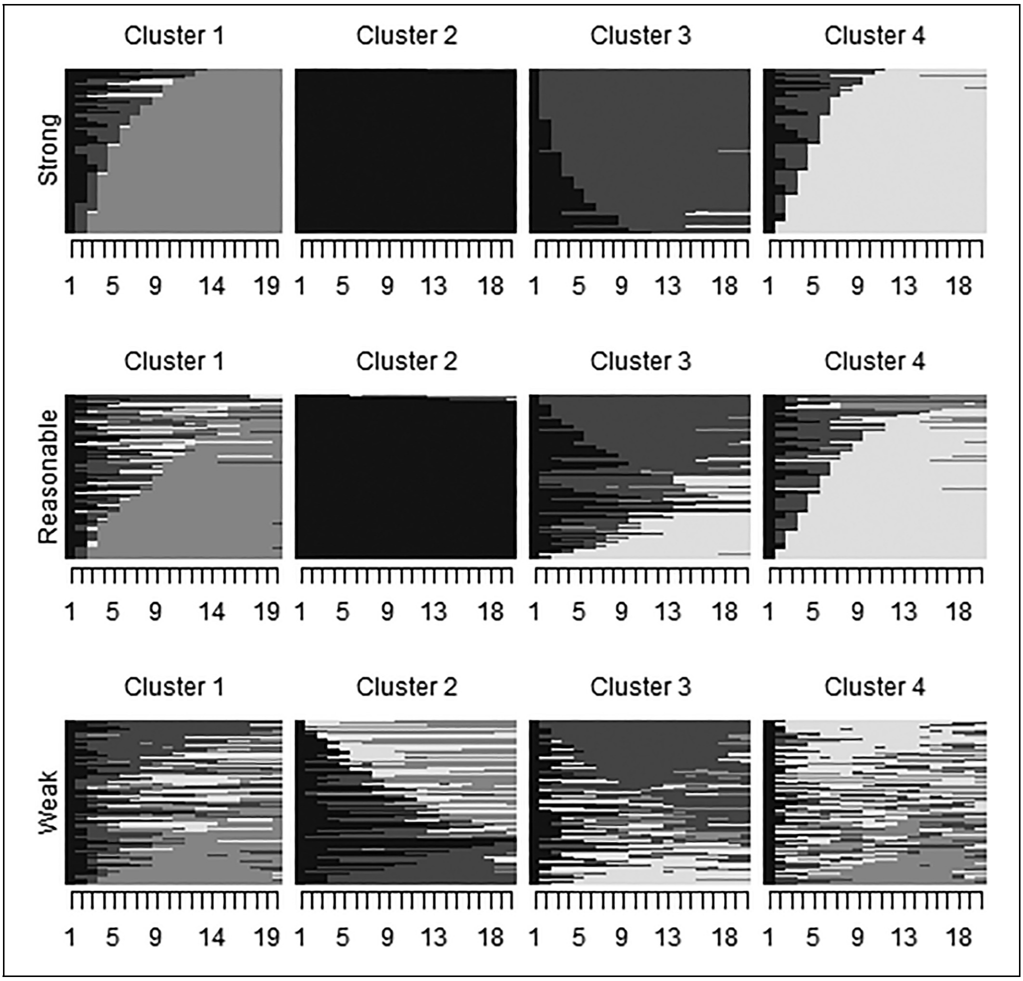

Figure 3. Clusters of sequences simulated from three types of mixture Markov models with varying clustering tendencies (weak, reasonable, and strong).
Note: The shades refer to four simulated states.

deviation of the noise term ϵ). Still, the predictions based on representativeness outperformed others, most notably the commonly used hard classification approach. The pseudoclass method, inspired by a similar approach from the LCA literature, did not perform particularly well in any setting. The soft classification was the most robust method over all scenarios and naturally performed best in the middle panel with a matching data-generating process.

We performed additional experiments where the original data-generation and covariate creation was done with FANNY-based hard classification and gravity centers, a potential alternative to our representativeness measure (Batagelj 1988). We also tested ranking the methods on the basis of the Bayesian information criterion instead of RMSE (excluding the pseudoclass method, for which the Bayesian information

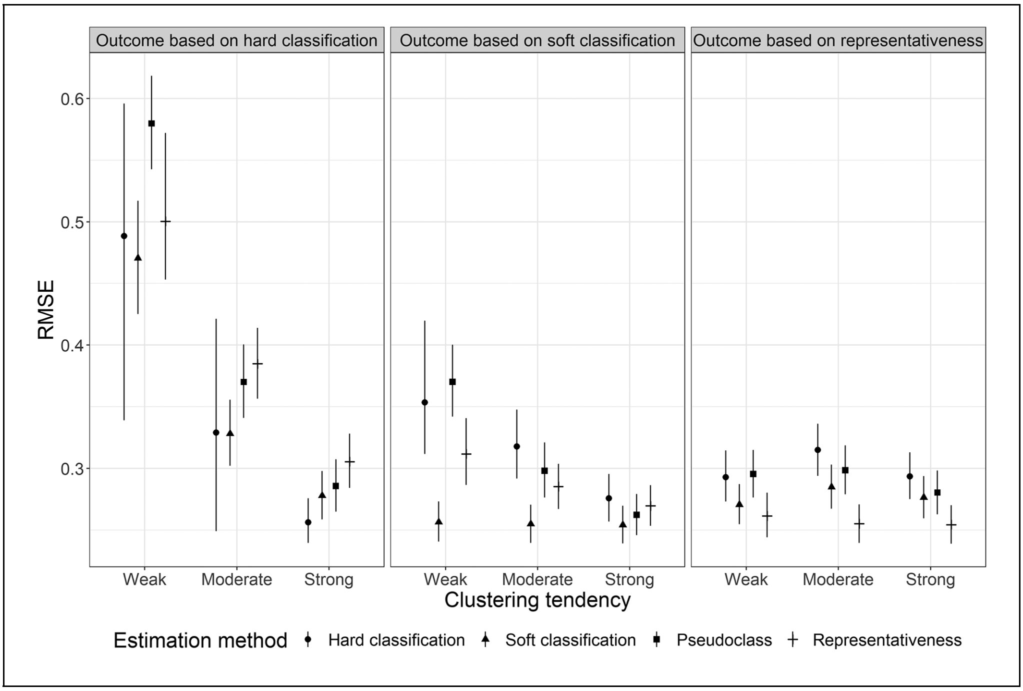

Figure 4. Average root mean squared errors (RMSEs) of predictions from 10,000 simulations with 95th percentile intervals.

criterion is not defined). These results are available in the supplementary material on GitHub https://github.com/helske/seqs2vars/tree/main/simulations. These additional simulations were in line with the conclusions of the main results, with FANNY-based hard classification performing similarly to the PAM-based hard classification and the gravity center method being similar to the representativeness method.

EMPIRICAL STUDY

We now illustrate the performance of the four methods with an empirical research problem: predicting a continuous earnings variable or a binary poverty variable with simple two-state sequences of employment trajectories. The timing, length, and frequency of employment and unemployment spells have a profound effect on earnings (Fuller 2015; Gangl 2006). These features of one’s occupational career determine the opportunities for on-the-job human capital accumulation, while also signaling a worker’s competence and unobservable qualities to potential employers (Gangl 2006). Over time, the cumulative effects on earnings can be substantial (Fuller 2015).

The data used in this example come from the Swedish population registers. The data set comprises a sample of all residents of Sweden who turned 18 years old in 1997 and who lived continuously in the country until 2017 (n = 10,000). In other words, we observe all subjects from age 18 to age 38. Yearly states are coded as “1 = working” and “2 = not working” on the basis of income and employment information from the

Longitudinal Integrated Database for Health Insurance and Labour Market Studies. Individuals who declared any income from employment in a given year are coded as employed. Other variables included in the multivariate analysis come from the Total Population Register.

We were interested in two outcome variables: (1) the probability of being in the lowest income quintile at the end of the sequence (a measure of poverty) and (2) the square root of cumulative income over the entire sequence (in 1,000 SEK). Income in this case is income from wages, business, and other economic activity, including social benefits related to economic activity (e.g., parental leave and sick leave compensations). We also had measurements of characteristics of the individual and their family background at the start of the sequence: region of residence (metropolitan areas, smaller cities, countryside), mother's education, father's education, mother's employment status, father's employment status, and sex.

We estimated the models for poverty using logistic regression and the models for income using ordinary least squares regression with four different methods to predict the outcome with employment histories. In both cases, we controlled for characteristics of the individual and their family at the start of the sequence.

For the clustering of sequences, we used a dissimilarity measure that is sensitive to the duration of (un)employment spells, namely, optimal matching with a substitution cost of 2 and an indel cost of 1 (Studer and Ritschard 2016). We chose a solution with five clusters for our example, with clusters differing in timing, prevalence, and continuity of employment. Figure 5 shows the medoids and the index plots for the sequences within each cluster of the hard classification (PAM) solution. The first cluster, *mainly no work*, is characterized by the prevalence of spells without work, being the cluster with longer and more prevalent states outside of employment. The second cluster, *unstable work*, is characterized by frequent transitions in and out of work, which become more common toward the end of the observation period. The cluster *late entry* is characterized by very late entry into the labor market, mostly after age 28, and often followed by frequent transitions in and out of work. *Delayed entry* is characterized by spells of unemployment or inactivity in the first seven years (i.e., between ages 18 and 25) and by a transition to mostly stable employment afterward. Last, *early entry* is characterized by a long spell in employment, starting early in the observation window.

A classification assigning cluster membership on the basis of the highest membership probability obtained by the FANNY algorithm showed similar qualitative patterns, with the majority of all sequences in each of the five classifications being allocated to the same cluster. There were, however, minor differences in allocation. First, the lowest degree of overlap between the two classifications was 57 percent for the category *unstable work*, which had 35.8 percent of its sequences allocated to *late entry* in the soft classification. Second, 33.7 percent and 20.2 percent of the *early entry* and *delayed entry*, respectively, were allocated to *unstable work*. These differences result in a higher share of the *working* state at later positions in the cluster *unstable work* of the soft classification than in the hard classification, but otherwise, the other clusters have remarkably similar patterns (see Tables A2 and A3 and Figure A1 in the Appendix).

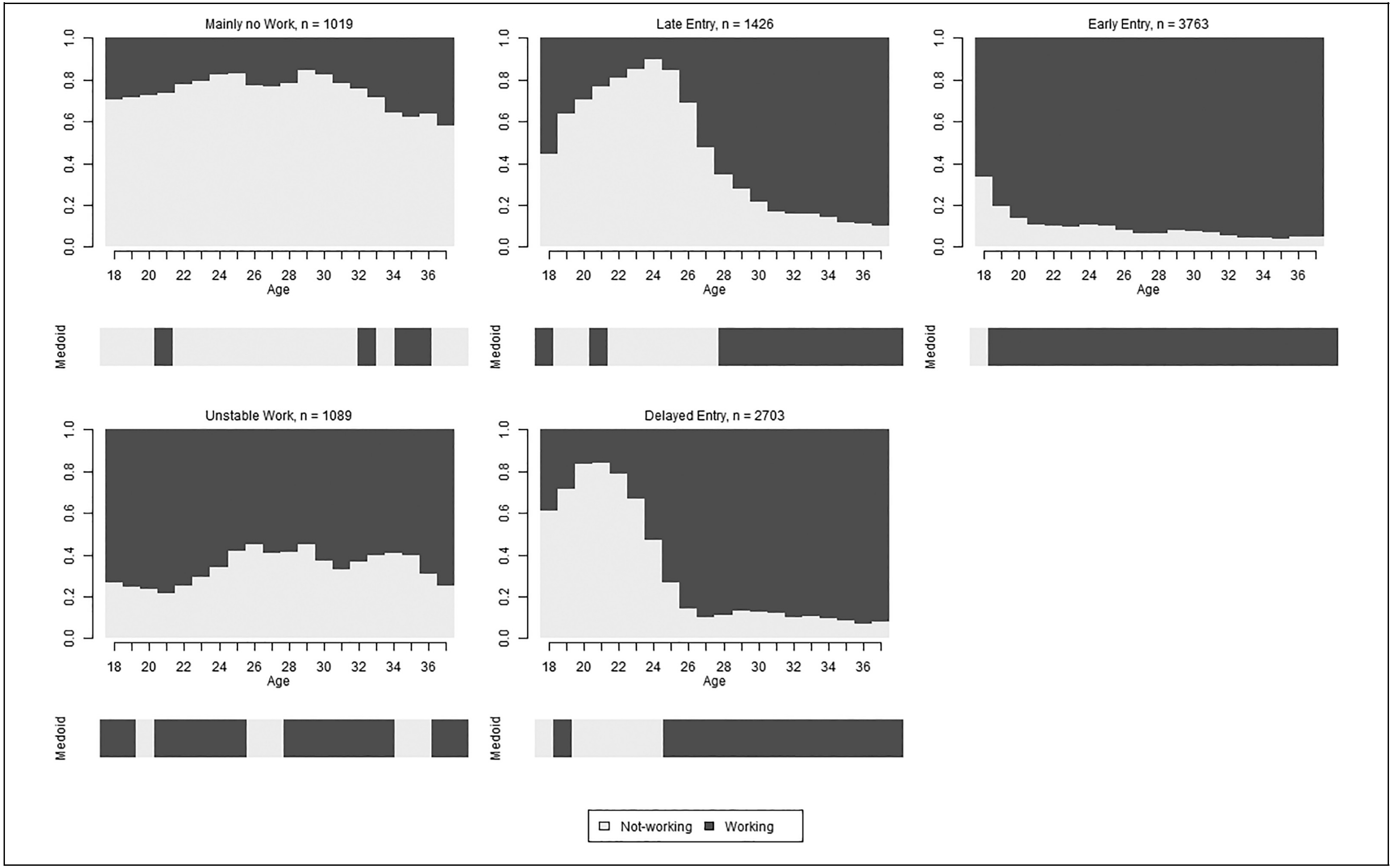

Figure 5. Yearly employment state distribution and medoids for five-cluster partitioning-around-medoids solution.
Note: The panels show a random sample of the Swedish cohort entering working age in 1998. *N* = 10,000.

Earlier research suggests the varying degrees of attachment to employment and the different lengths of employment spells found in each cluster would have distinct outcomes in terms of poverty and cumulative earnings. This can be studied using the cluster variable as a predictor of these two outcomes, in a similar way to the study by Fuller (2015). In addition, we repeated the analysis using the other approaches described in the simulation study, namely pseudoclass, soft classification, and representativeness.

In this case, we did not believe any true employment clusters exist or that the outcomes would be class dependent. Instead, we assumed the relationship between the work trajectory and the outcome (income or poverty) is similarity based and expected that representativeness would be the most appropriate measure to use. Our analysis highlights the substantial differences in how the different types of sequence variables perform as predictors. Before showing the full results, we illustrate the differences in predicted values with a simple example.

When using hard cluster memberships and setting *early entry* (cluster 5) as the reference category, the expected cumulative income for an individual would be

$$\mathbb{E}(I_i) = \beta_0 + \beta_1 C_i^1 + \beta_2 C_i^2 + \beta_3 C_i^3 + \beta_4 C_i^4 + Z_i \gamma, \tag{1}$$

where $\mathbb{E}(I_i)$ is the expected (square root of) cumulative income for individual i, β_0 the intercept, β_k the regression coefficient, C_i^k membership in cluster k of sequence i (0 or 1 for hard classification), and Z_i represents all other covariates in the model and γ the vector of their coefficients. Now consider three sequences from the *mainly no work* cluster (cluster 1), (M), (A), and (B), consisting of employment (E) and unemployment/inactivity (U):

(**M**) U–U–U–E–U–U–U–U–U–U–U–U–U–U–E–U–E–E–U–U
(**A**) U–U–U–U–U–U–U–U–U–U–U–U–U–U–U–U–U–U–U–U
(**B**) E–E–E–E–E–U–E–E–E–U–U–U–U–U–U–U–U–U–E–U

A hard classification method (PAM) assigns these three sequences to the same cluster, which is characterized by long unemployment spells. Here, sequence (M) is the medoid of the cluster and shows a pattern of mostly unemployment, (A) consists solely of unemployment spells, and (B) is an outlier with a long spell of nearly continuous employment that ends halfway through the period. For the case of hard cluster memberships as predictors, the square root of expected 20-year cumulative earnings for all these sequences is reduced to

$$\mathbb{E}(I_i) = \beta_0 + \beta_1 + Z_i \gamma = 64.69 - 33.53 + Z_i \gamma.$$

Note that cluster membership is reflected in the equation as a single parameter referring to the cluster assigned to all three sequences (in this case the first cluster). For simplicity, if we assume the individuals in question belonged to the baseline category for all other covariates, the predicted value of the square root of the 20-year cumulative earnings (in thousands of Swedish kronor) is 31.16, which translates approximately to SEK 970,000 for the three cases of (M), (A), and (B).

Likewise, for the pseudoclass approach, the equation is

$$\mathbb{E}(I_i) = \beta_0^* + \beta_1^* + Z_i\gamma^* = 63.28 - 27.49 + Z_i\gamma^*,$$

where the coefficients are averages over multiple pseudoclass samples (the estimates are different compared with those from the hard classification method, as reflected by the asterisks). The equivalent square root of predicted earnings is 35.79, translating into approximately SEK 1,280,000 for all of (M), (A), and (B).

As discussed earlier, a key difference between hard clustering and pseudoclass is that pseudoclass assigns cluster memberships on the basis of the estimated membership probabilities from a fuzzy cluster solution. The coefficients represent the averaged cluster membership effect over all the replications, and the standard errors are adjusted to reflect the uncertainty deriving from the probabilistic cluster allocation. In this way, pseudoclass deals with the problem of treating group assignment as certain by adjusting the estimated parameters and standard errors so they reflect the uncertainty in cluster allocation. Yet pseudoclass is similar to hard classification in that it attributes a uniform effect to all members of the same cluster, as our example shows. Also note the difference in estimates between the methods: pseudoclass tends to shrink estimates toward the average (Bray et al. 2015; Lanza et al. 2013), which makes it the most conservative of all methods in terms of finding differences between the groups.

In contrast, the equations for the soft classification and representativeness methods reflect within-cluster variability by incorporating more parameters and changing the predictors into continuous measures. For soft classification, the equation includes $k - 1$ parameters representing membership probabilities, leading to

$$\mathbb{E}(I_i) = 69.31 - 56.31P_i^1 - 12.36P_i^2 - 19.67P_i^3 + 9.39P_i^4 + Z_i\gamma^\dagger,$$

where P_i^k is the probability that sequence i belongs to the soft cluster k (similar to the hard cluster approach, one of the cluster membership probabilities is left out of the equation because the probabilities sum to 1). Instead of sampling-based adjustment to uncertainty in the pseudoclass method, here the uncertainty of classification is incorporated into the covariates P_i^k themselves. Imputing the membership probabilities for each sequence yields

$$\mathbb{E}(I_M) = 69.31 - 56.31\times0.83 - 12.36\times0.03 - 19.67\times0.08 + 9.39\times0.05 + Z_i\gamma^\dagger.$$
$$\mathbb{E}(I_A) = 69.31 - 56.31\times0.86 - 12.36\times0.03 - 19.67\times0.05 + 9.39\times0.04 + Z_i\gamma^\dagger.$$
$$\mathbb{E}(I_B) = 69.31 - 56.31\times0.43 - 12.36\times0.16 - 19.67\times0.20 + 9.39\times0.13 + Z_i\gamma^\dagger.$$

For (M), the predicted value of square root earnings (in 1,000 SEK) is 21.10, translating into about SEK 445,000; for (A), the same predicted value is 19.9, translating into about SEK 396,000; and for (B) the value is 40.41, which translates into about SEK 1,633,000. Thus, the estimation based on soft classification captures the considerable earnings difference that results from differences in the presence of unemployment spells within the three sequences.

In a similar vein, the equation using representativeness incorporates multiple parameters (which do not have to sum to 1):

$$\mathbb{E}(I_i) = -4.58 - 1.12R_i^1 - 11.68R_i^2 + 26.41R_i^3 + 25.70R_i^4 + 48.25R_i^5 + Z_i\gamma^\dagger,$$

where R_i^k is the representativeness value of the representative sequence k (here, the medoid of cluster k) to sequence i. The resulting equations from imputing the representativeness values of each example sequence are

$$\begin{aligned}
\mathbb{E}(I_M) &= -4.58 - 1.12\times1.00 - 11.68\times0.45 + 26.41\times0.60 + 25.70\times0.50 + 48.25\times0.25 + Z_i\gamma^\dagger,\\
\mathbb{E}(I_A) &= -4.58 - 1.12\times0.80 - 11.68\times0.25 + 26.41\times0.40 + 25.70\times0.30 + 48.25\times0.05 + Z_i\gamma^\dagger,\\
\mathbb{E}(I_B) &= -4.58 - 1.12\times0.65 - 11.68\times0.60 + 26.41\times0.50 + 25.70\times0.45 + 48.25\times + .45 + Z_i\gamma^\dagger,
\end{aligned}$$

The predicted square root earnings (in 1,000 SEK) for sequence (M) using the representativeness method is 29.80, translating into approximately SEK 888,000. For sequence (A) it is 12.29, which translates to approximately SEK 151,000. For sequence (B), the value is 34.17, translating into about SEK 1,167,000. As in the case of soft classification, representativeness also captures the differences in earnings between the three sequences even when the clustering algorithm has assigned them to the same group.

As illustrated, the four approaches differ in terms of how predictions are calculated, which means they also differ in terms of interpreting the estimated modeling results. Interpretation is most straightforward for hard clustering, as it is interpreted as any categorical variable: parameter coefficients β_k correspond to the average differences between members of cluster k and members of the reference category. The estimates from pseudoclass, even though calculated in a different manner, are interpreted in the same way. For soft classification, the coefficients are interpreted as the difference in the outcome (or the probability of the outcome, in terms of the logistic model) between a fully certain member of cluster k (whose membership probability is 1 for that cluster and 0 for others) and a fully certain member of the reference category. For representativeness, the parameter estimates are difficult to interpret as such, as they correspond to dissimilarities/distances to all the medoids (or other representative sequences). When interpreting the parameter estimates, one cannot simply interpret one of them but must consider all at the same time. However, average marginal effects (AMEs) or average marginal predictions (AMPs) can be used for easier interpretation, and they are comparable across all four approaches.

Here we show model results as AMPs for all four approaches. AMPs and AMEs are similar concepts, except that instead of comparisons with a reference case as in the more typical AMEs, the AMPs, also known as average adjusted predictions, show marginal predictions under some interesting configurations, in our case, at the medoids obtained from the hard classification. Specifically, separately for each medoid, we predicted the outcome for each individual by replacing their observed representativeness values with the representativeness values of the medoid (while keeping other covariates at their observed values) and then calculated the average of the predictions over all individuals. Similarly, for soft classification, we replaced the observed membership

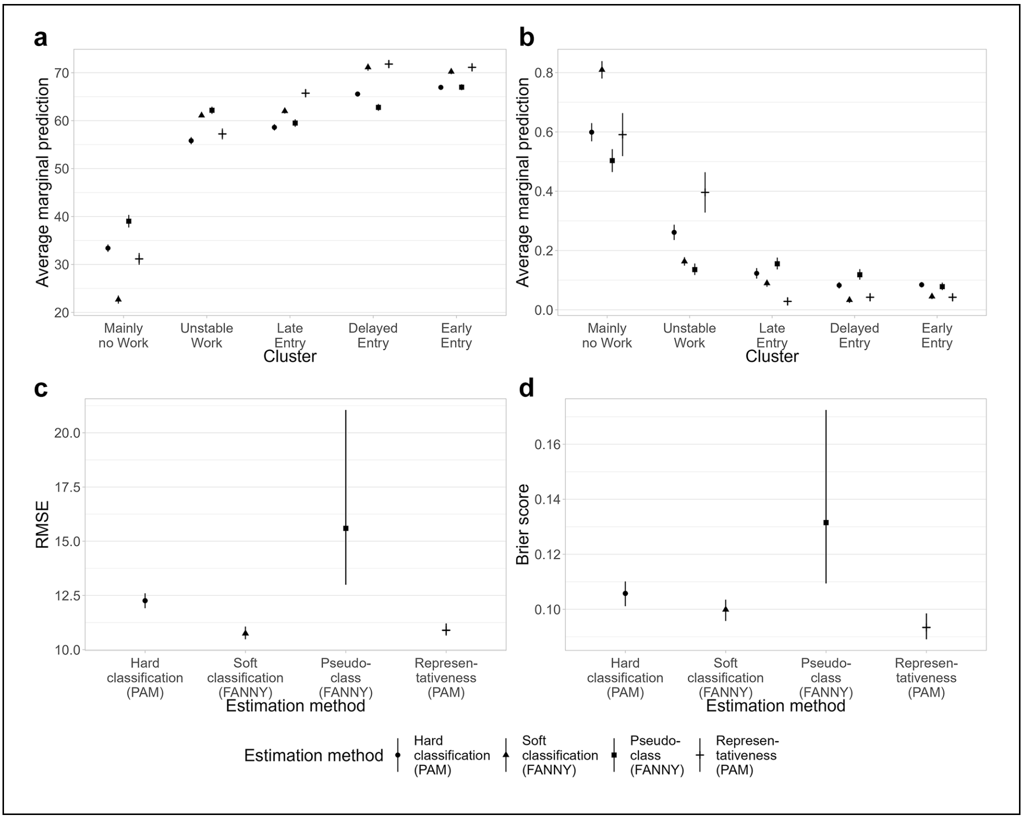

Figure 6. Average marginal predictions, root mean squared errors (RMSEs), and Brier scores by estimation method and outcome: (a) average marginal predictions (income), (b) average marginal predictions (poverty), (c) RMSEs (income), and (d) Brier scores (poverty).
Note: FANNY = fuzzy analysis; PAM = partitioning around medoids.

probabilities of each individual with those of the medoids, and with hard classification, AMPs are calculated by replacing the observed cluster memberships. Finally, the pseudoclass AMPs are calculated for each pseudoclass replication as with hard classification, and the set of AMPs obtained from all pseudoclass replications are then combined using Rubin's rules.

The top two panels (a and b) in Figure 6 display the AMPs for the clusters (hard and soft classification and pseudoclass) or medoids of each cluster (representativeness) by outcome and estimation approach. The estimates largely agree with each other, predicting worst outcomes for the *mainly no work* cluster concerning both discrete (poverty) and continuous (income) outcomes. The differences between the methods are the largest in the heterogeneous *mainly no work* cluster and the smallest for the more homogeneous *early entry* cluster. Here, all confidence intervals are narrow with the exception of the *mainly no work* and *unstable work* clusters for representativeness (and even there the estimates are clearly statistically significantly different). Overall, representativeness- and soft classification–based estimates show larger differences between the clusters in comparison with hard classification and pseudoclass.

The lower panels (c and d) in Figure 6 show the RMSEs and Brier scores that we used to assess the accuracy of the predictions. We computed them using a leave-one-out cross-validation method over 100 folds and estimated confidence intervals by using bootstrapping with 1,000 replications. As expected, representativeness produced more accurate estimates in both cases than did the hard classification and pseudoclass methods; soft classification was close to the performance of representativeness, especially in the continuous case. The Appendix provides further results for the empirical study, such as descriptive statistics, parameter estimates, and information criteria from each model.

DISCUSSION

In this article, we aimed to bring forward and discuss the problems of the traditional approach of creating variables from SA clusters and to propose some alternative approaches. Our simulation study demonstrated how the type of data-generating process affects the performance of the different methods. In cases with true but unknown clusters, hard classification worked well on data with strong clustering tendency, whereas soft classification was consistently better on data with weaker clustering tendencies (i.e., when classification error is an issue). However, when there were no true clusters to begin with but the sequence–outcome relationship was assumed to be similarity based, representativeness clearly outperformed other methods.

We also studied the performance of the methods on empirical data, where we predicted two types of income-related variables (a continuous cumulative income variable and a binary poverty measure) with simple employment trajectories and control variables. In this case, we assumed the relationship between the sequences and the outcome would be closest to the similarity-based setup and expected that the representativeness measure would result in better predictions than the other methods. This was confirmed by our analyses using cross-validation, but the advantage of using representativeness was not as evident in the empirical case as it was in the simulations. Soft classification was equally good for the continuous outcome, but it performed less well when the outcome was binary.

We argue that in the social sciences, subjects are typically more or less hybrids of multiple ideal types, and the outcome variable of interest is affected by multiple factors with varying magnitudes, which is not properly captured by hard classification into clusters. Earlier LCA literature hypothesized that the pseudoclass method can better account for uncertainty due to clustering. The benefit of pseudoclass method over other proposed alternatives is that it tries to adjust for the uncertainty in the classification without altering the interpretation of the model in terms of the corresponding predictors. However, its performance in our simulation and empirical studies was less than convincing, which is in line with recent LCA literature (Bray et al. 2015; Lanza et al. 2013). The pseudoclass method is also computationally the most demanding of the considered methods. Although our pseudoclass approach is based on fuzzy clustering of sequence dissimilarities, not latent class models, on the basis of all these findings we cannot recommend the pseudoclass method as an alternative to the traditional hard classification technique.

Soft classification with mixed memberships account for uncertainty of membership allocation, and as such it is a clear improvement over the traditional hard classification with fixed memberships. The potential problem with soft classification is its inability to deal with cases that are not well represented by any of the ideal types (outliers). Similarity-based approaches such as representativeness take into account the closeness of the sequence to the ideal types while also distinguishing between mixed types and outliers. Other similarity-based measures may also work, such as measures based on multidimensional scaling, especially when the data show clear and easily interpretable principal components or when the goal is to construct a control variable (where interpretation of effects of the sequence variables is not relevant). If outliers are not a big issue, soft classification and representativeness measures are expected to lead to relatively similar results. In this case, soft classification could be favored because of simpler interpretation. In theory, the use of representativeness and membership probabilities can induce some level of multicollinearity to the modeling, but we do not see this as a major issue, as multicollinearity affects only the interpretation of individual predictors, and in these cases, the effects of these sequence-related variables are best considered as a whole (as in our examples).

Related to multicollinearity, we fixed the number of clusters and medoids to be the same across the different approaches for comparability. In practice, it may be advisable to use a smaller number of clusters/medoids for representativeness and possibly also for soft classification in comparison with hard classification, because of the continuous nature of these measures. For example, in our simple empirical example, dissimilarities to sequences of “always working” and “never working” capture the same information (causing multicollinearity), so in practice adding only one of these as a representativeness predictor would be sufficient.

To conclude, we demonstrated the importance of considering how sequences and the outcome variable of interest are related, and the need to adjust the analysis accordingly. If true underlying clusters are expected to exist, then hard or soft classification methods should be preferred (depending on how big an issue classification error is expected to be). In social sciences, the whole idea of the existence of any “true clusters” is often implausible. Often the main purpose of cluster analysis is to reduce the complexity of the sequence data, in which case similarity-based approaches or soft classification should be considered. On the basis of our analyses, the representativeness method shows promising results, and perhaps other alternatives will emerge in future work. We hope this article will encourage further discussion and research on combining SA and subsequent modeling.

APPENDIX: EMPIRICAL STUDY

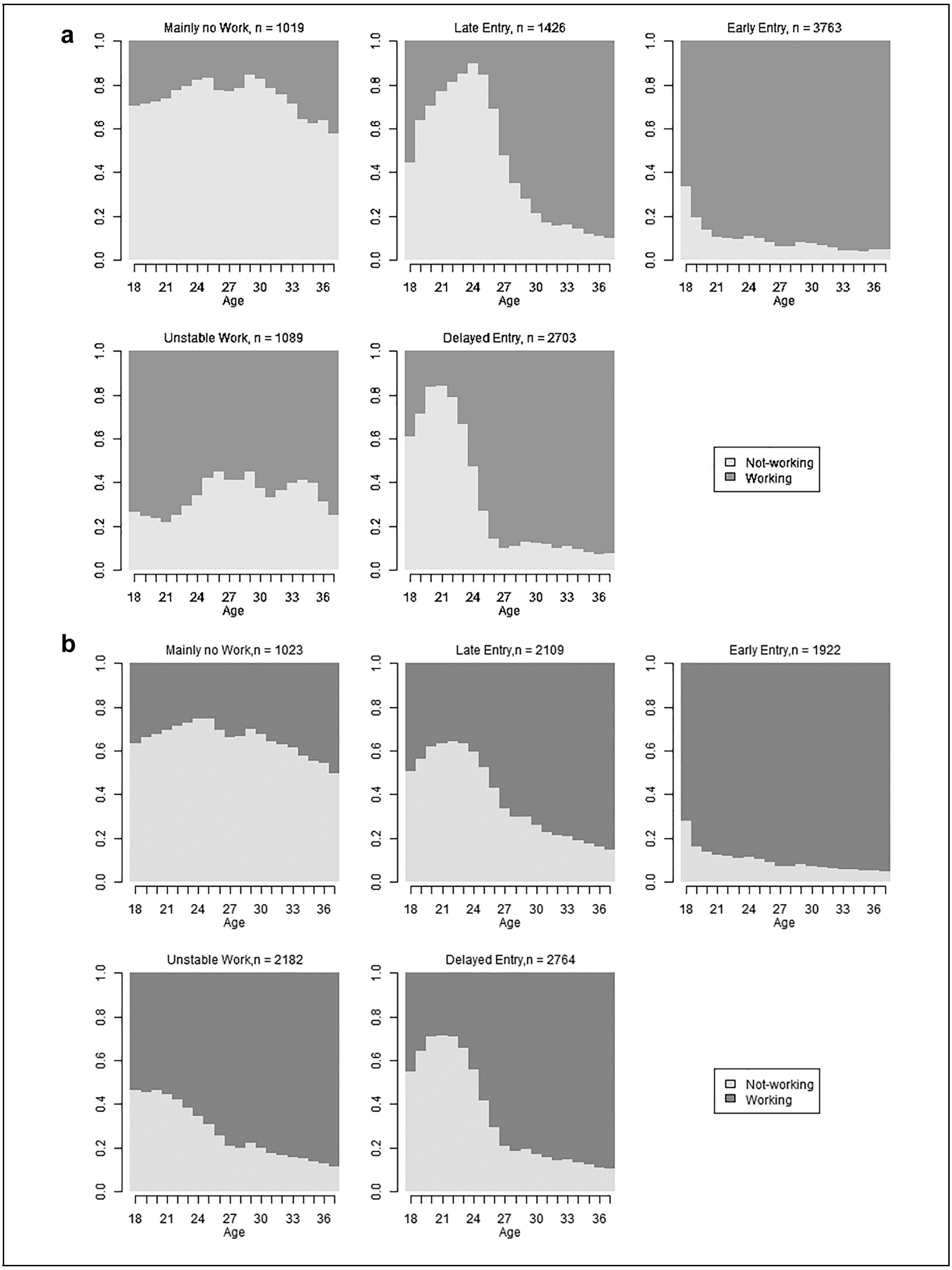

Figure A1. Comparison between hard and soft cluster classifications: (a) hard classification and (b) soft classification.

Note: Soft classification is presented using weighted sequence plots, as suggested by Studer (2018).

Table A1. Outcome Variables by Cluster

Cumulative Income over 20 Years

Cluster	Mean	S.D.	*n*
Mainly no work	1,520.71	1,502.40	1,019
Unstable work	3,076.23	1,186.19	1,089
Late entry	3,606.48	1,544.39	1,426
Delayed entry	4,503.90	1,894.54	2,703
Early entry	4,632.64	1,420.23	3,763

Percent in Poverty at Year 20

Cluster	%	Count	*n*
Mainly no work	61.14	623	1,019
Unstable work	28.10	306	1,089
Late entry	11.64	166	1,426
Delayed entry	7.73	209	2,703
Early entry	8.64	325	3,763

Table A2. Cross-Tabulation between Cluster Assignments by Hard Classification (PAM) and Soft Classification (Based on the Maximum Membership Probabilities from FANNY)

	Soft Classification				
Hard Classification	Early Entry	Delayed Entry	Unstable Work	Mainly No Work	Late Entry
Early entry	66.3% (2,494)	.0% (0)	33.7% (1,269)	.0% (0)	.0% (0)
Delayed entry	.0% (0)	65.4% (1,769)	20.2% (546)	.1% (3)	14.2% (385)
Unstable work	.2% (2)	.0% (0)	57.0% (621)	7.0% (76)	35.8% (390)
Mainly no work	.0% (0)	.0% (0)	.0% (0)	91.9% (936)	8.1% (83)
Late entry	.0% (0)	15.1% (215)	.6% (9)	1.5% (21)	82.8% (1,181)

Note: FANNY = fuzzy analysis; PAM = partitioning around medoids.

Table A3. Average Soft Classification Probabilities by Hard Classification Cluster

	Soft Classification				
Hard Classification	Early Entry	Delayed Entry	Unstable Work	Mainly No Work	Late Entry
Early entry	63.9%	7.0%	22.3%	.9%	5.9%
Delayed entry	5.3%	43.5%	23.7%	3.0%	24.5%
Unstable work	12.3%	18.6%	32.4%	10.3%	26.5%
Mainly no work	2.7%	9.6%	7.1%	64.5%	16.1%
Late entry	3.9%	30.9%	14.2%	9.9%	41.1%

Table A4. Estimated Coefficients and Standard Errors for Ordinary Least Squares Models with the Square Root of Cumulative Income (in 1,000 SEK) as a Dependent Variable

	Hard Classification		Soft Classification		Pseudoclass		Representativeness	
	β	s.e.	β	s.e.	β	s.e.	β	s.e.
Sex (reference: male)								
Female	–5.293	(.258)	–3.967	(.240)	–6.007	(.321)	–3.678	(.237)
Region (age 18; reference: Stockholm)								
Gothenburg	–.650	(.494)	–.443	(.442)	–.890	(.601)	–.340	(.436)
Malmö	–2.194	(.554)	–1.779	(.496)	–3.235	(.684)	–1.750	(.489)
Rest of Sweden	–1.549	(.317)	–1.607	(.284)	–1.432	(.382)	–1.639	(.280)
Mother education (age 18; reference: primary)								
Secondary	1.326	(.321)	1.038	(.287)	1.628	(.378)	.933	(.283)
Postsecondary	3.289	(.384)	2.941	(.345)	3.314	(.459)	2.646	(.340)
Father education (age 18; reference primary)								
Secondary	1.326	(.3)	1.278	(.269)	1.473	(.353)	1.163	(.265)
Postsecondary	2.816	(.372)	2.763	(.334)	2.768	(.446)	2.501	(.330)
Mother employment (age 18; reference: did not work)								
Mother worked	1.667	(.32)	1.094	(.286)	2.317	(.386)	1.124	(.282)
Father employment (age 18; reference: did not work)								
Father worked	2.081	(.355)	1.333	(.318)	2.960	(.426)	1.254	(.314)
Cluster								
Mainly no work	–33.534	(.451)	–56.308	(.616)	–27.954	(.751)	–1.123	(2.837)
Unstable work	–11.120	(.437)	–12.362	(.797)	–4.815	(.522)	–11.678	(2.059)
Late entry	–8.338	(.405)	–19.672	(1.018)	–7.479	(.534)	26.409	(2.224)
Delayed entry	–1.388	(.328)	9.391	(.804)	–4.214	(.487)	25.696	(2.095)
Early entry	Reference	Reference	Reference	Reference	Reference	Reference	48.253	(2.836)
Intercept	64.686	(.539)	69.306	(.521)	63.617	(.662)	–4.584	(2.494)
AIC	78,713.483		76,485.934		NA		76,196.274	
BIC	78,828.848		76,601.3		NA		76,318.85	

Note: AIC = Akaike information criterion; Bayesian information criterion = BIC; NA = not applicable.

Table A5. Estimated Coefficients and Standard Errors for Logistic Regression Models with Poverty Status at Year 20 as a Dependent Variable

	Hard Classification		Soft Classification		Pseudoclass		Representativeness	
	β	s.e.	β	s.e.	β	s.e.	β	s.e.
Sex (reference: male)								
Female	.205	(.063)	.123	(.067)	.271	(.064)	.094	(.07)
Region (age 18; reference: Stockholm)								
Gothenburg	–.138	(.118)	–.170	(.123)	–.092	(.117)	–.220	(.127)
Malmö	.01	(.127)	–.054	(.133)	.130	(.125)	–.074	(.139)
Rest of Sweden	–.245	(.077)	–.263	(.079)	–.245	(.076)	–.273	(.082)
Mother education (age 18; reference: primary)								
Secondary	–.178	(.075)	–.139	(.078)	–.204	(.074)	–.137	(.08)
Postsecondary	–.478	(.095)	–.467	(.098)	–.480	(.093)	–.406	(.101)
Father education (age 18; reference: primary)								
Secondary	–.157	(.072)	–.162	(.074)	–.176	(.070)	–.148	(.076)
Postsecondary	–.278	(.092)	–.315	(.095)	–.315	(.090)	–.223	(.099)
Mother employment (age 18; reference: did not work)								
Mother worked	–.248	(.074)	–.212	(.076)	–.309	(.072)	–.226	(.079)
Father employment (age 18; reference: did not work)								
Father worked	–.214	(.080)	–.141	(.084)	–.290	(.079)	–.128	(.087)
Cluster								
Mainly no work	2.837	(.092)	5.435	(.173)	2.480	(.123)	–3.016	(.795)
Unstable work	1.362	(.093)	2.783	(.259)	.615	(.128)	4.931	(.553)
Late entry	.422	(.106)	2.100	(.310)	.773	(.122)	–6.989	(.68)
Delayed entry	–.026	(.096)	–1.602	(.324)	.458	(.130)	.242	(.653)
Early entry	Reference	Reference	Reference	Reference	Reference	Reference	–9.769	(.815)
Intercept	–1.609	(.125)	–2.436	(.160)	–1.551	(.136)	8.437	(.763)
AIC	72,82.581		68,78.545		NA		65,34.489	
BIC	73,90.737		6,986.7		NA		66,49.855	

Note: AIC = Akaike information criterion; Bayesian information criterion = BIC; NA = not applicable.

Table A6. Correlation Coefficients between Soft Classification Probabilities

	Early Entry	Delayed Entry	Unstable Work	Mainly No Work	Late Entry
Early entry	1.00	–.67	–.33	–.34	–.75
Delayed entry	–.67	1.00	.11	–.22	.57
Unstable work	–.33	.11	1.00	–.32	.07
Mainly no work	–.34	–.22	–.32	1.00	.01
Late entry	–.75	.57	.07	.01	1.00

Table A7. Correlation Coefficients between Cluster Representativeness

	Early Entry	Delayed Entry	Unstable Work	Mainly No Work	Late Entry
Early entry	1.00	.50	.82	–.95	–.06
Delayed entry	.50	1.00	.57	–.32	.73
Unstable work	.82	.57	1.00	–.71	.16
Mainly no work	–.95	–.32	–.71	1.00	.26
Late entry	–.06	.73	.16	.26	1.00

Acknowledgments

We thank the anonymous reviewers for their valuable comments, Mara Nelson Grynaviski for copyediting, and the CSC – IT Center for Science, Finland, for computational resources.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Academy of Finland (decision numbers 331816, 331817, and 311877), the Academy of Finland Flagship Programme (decision number 320162), the Swedish Research Council for Health, Working Life and Welfare (decision number 2016-07105), and the Swedish Research Council (decision number 445-2013-7681). The project was financially supported by the funding organizations AKA, ANR, CAS, DFF, DFG, ESRC, FCT, IRC, NWO, RANNÍS, RCN, SNSF, and VR, involved in the NORFACE Joint Research Programme on Dynamics of Inequality Across the Life-Course, which is cofunded by the European Commission through Horizon 2020 under grant agreement 724363.

ORCID iDs

Satu Helske https://orcid.org/0000-0003-0532-0153
Jouni Helske https://orcid.org/0000-0001-7130-793X
Guilherme K. Chihaya https://orcid.org/0000-0002-5219-6784

Supplemental Material

Supplemental material for this article is available online.

References

Bakk, Zsuzsa, Daniel L. Oberski, and Jeroen K. Vermunt. 2017. “Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference.” *Political Analysis* 22(4):520–40.

Bakk, Zsuzsa, Fetene B. Tekle, and Jeroen K. Vermunt. 2013. “Estimating the Association between Latent Class Membership and External Variables Using Bias-Adjusted Three-Step Approaches.” *Sociological Methodology* 43(1):272–311.

Bandeen-Roche, Karen, Diana L. Miglioretti, Scott L. Zeger, and Paul J. Rathouz. 1997. “Latent Variable Regression for Multiple Discrete Outcomes.” *Journal of the American Statistical Association* 92(440): 1375–86.

Batagelj, Vladimir. 1988. “Generalized Ward and Related Clustering Problems.” Pp. 67–74 in *Classification and Related Methods of Data Analysis*, edited by H. H. Bock. Amsterdam, the Netherlands: North-Holland.

Berglund, Lars. 2012. “Regression Dilution Bias: Tools for Correction Methods and Sample Size Calculation.” *Upsala Journal of Medical Sciences* 117(3):279–83.

Bray, Bethany C., Stephanie T. Lanza, and Xianming Tan. 2015. “Eliminating Bias in Classify-Analyze Approaches for Latent Class Analysis.” *Structural Equation Modeling: A Multidisciplinary Journal* 22(1):1–11.

Chaparro, M. Pia, Xavier de Luna, Jenny Häggström, Anneli Ivarsson, Urban Lindgren, Karina Nilsson, and Ilona Koupil. 2017. “Childhood Family Structure and Women’s Adult Overweight Risk: A Longitudinal Study.” *Scandinavian Journal of Public Health* 45(5):511–19.

D’Urso, Pierpaolo. 2007. *Fuzzy Clustering of Fuzzy Data*. Hoboken, NJ: John Wiley.

Ellwardt, Lea, Marja Aartsen, and Theo van Tilburg. 2017. “Types of Non-kin Networks and Their Association with Survival in Late Adulthood: A Latent Class Approach.” *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 72(4):694–705.

Fuller, Sylvia. 2015. “Do Pathways Matter? Linking Early Immigrant Employment Sequences and Later Economic Outcomes: Evidence from Canada.” *International Migration Review* 49(2):355–405.

Gabadinho, Alexis, and Gilbert Ritschard. 2013. “Searching for Typical Life Trajectories Applied to Childbirth Histories.” Pp. 287–312 in *Gendered Life Courses between Standardization and Individualization: A European Approach Applied to Switzerland*, edited by R. Levy and E. D. Widmer. Vienna, Austria: LIT.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S. Müller, and Matthias Studer. 2011. “Analyzing and Visualizing State Sequences in R with TraMineR.” *Journal of Statistical Software* 40(4):1–37.

Gangl, Markus. 2006. “Scar Effects of Unemployment: An Assessment of Institutional Complementarities.” *American Sociological Review* 71(6):986–1013.

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

Jalovaara, Marika, and Anette Eva Fasang. 2020. “Family Life Courses, Gender, and Mid-life Earnings.” *European Sociological Review* 36(2):159–78.

Kaufman, Leonard, and Peter J. Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Lanza, Stephanie T., Xianming Tan, and Bethany C. Bray. 2013. “Latent Class Analysis with Distal Outcomes: A Flexible Model-Based Approach.” *Structural Equation Modeling: A Multidisciplinary Journal* 20(1):1–26.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2021. “cluster: Cluster Analysis Basics and Extensions.” https://CRAN.R-project.org/package=cluster.

Piccarreta, Raffaella, and Matthias Studer. 2019. “Holistic Analysis of the Life Course: Methodological Challenges and New Perspectives.” *Advances in Life Course Research* 41(1):100–251.

R Core Team. 2021. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing.

Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20(1):53–65.

Rubin, Donald B. 2004. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley.

Studer, Matthias. 2013. “WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R.” *LIVES Working Papers* 24:1–32.

Studer, Matthias. 2018. “Divisive Property-Based and Fuzzy Clustering for Sequence Analysis.” Pp. 223–39 in *Sequence Analysis and Related Approaches*, edited by G. Ritschard and M. Studer. Berlin, Germany: Springer.

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 (Statistics in Society)* 179(2):481–511.

Warren, John Robert, Liying Luo, Andrew Halpern-Manners, James M. Raymo, and Alberto Palloni. 2015. “Do Different Methods for Modeling Age-Graded Trajectories Yield Consistent and Valid Results?” *American Journal of Sociology* 120(6):1809–56.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. “dplyr: A Grammar of Data Manipulation.” R package version 1.0.7. Retrieved May 19, 2023. https://cloud.r-project.org/web/packages/dplyr/index.html.

Author Biographies

Satu Helske is a senior research fellow in the INVEST Research Centre and the Department of Social Research, University of Turku in Finland. She works at the crossroads of sociology and statistics, studying topics related to the life-course, inequality, and intergenerational processes.

Jouni Helske is a senior researcher in statistics at the University of Jyväskylä in Finland. His current research focuses on Bayesian methods, causal inference, and statistical software development, especially in the context of time-series and panel data.

Guilherme K. Chihaya is an associate professor of sociology on the Faculty of Social Sciences, Nord University, and an affiliated researcher at the Institute for Analytical Sociology, Linköping University. His research focuses on the residential and labor market outcomes of migrants.