Comparing methods of classifying life courses: Sequence Analysis and Latent Class Analysis

Sapphire Y. Han Netherlands Interdisciplinary Demographic Institute (NIDI/KNAW), The Hague & University of Groningen, Groningen, The Netherlands

han@nidi.nl

Aart C. Liefbroer Netherlands Interdisciplinary Demographic Institute (NIDI/KNAW), The Hague, University of Groningen, Groningen & Vrije Universiteit, The Netherlands

Cees H. Elzinga Netherlands Interdisciplinary Demographic Institute (NIDI/KNAW), The Hague & Vrije Universiteit, The Netherlands

(Received February 2016 Revised May 2017) http://dx.doi.org/10.14301/llcs.v8i4.409

Abstract

We compare life course typology solutions generated by sequence analysis (SA) and latent class analysis (LCA). First, we construct an analytic protocol to arrive at typology solutions for both methodologies and present methods to compare the empirical quality of alternative typologies. We apply this protocol to develop and compare SA- and LCA-derived family-life typologies for women born between 1960 and 1964 in 15 European countries, using data from the Family and Fertility Survey. This paper contributes to the use of these classification techniques in four different ways. First, we present guidelines on how to establish the number of classes or clusters to use. Second, we show how to evaluate the stability of these clusters. Third, we provide a way to evaluate the validity of these clusters and finally, we provide for a formal heuristic to relate the stochastically defined latent classes to the distance-based clusters found with SA.

Keywords

Life course, sequence analysis, latent class analysis, typology comparison

Introduction

A prominent approach within life course research is to analyse life courses as sequences of states or state-transitions (see e.g. Buchmann & Kriesi, 2011). In this approach, two main methodological paradigms have been widely used: Event History Analysis (see e.g. Mills, 2011) focuses on describing or explaining the time to occurrence of specific events. The second approach takes a holistic perspective and utilises the life course itself as the unit of analysis and usually aims for a typology of life course trajectories. The typology itself may reveal substantive patterns and the resulting class-membership is often used as a dependent or independent variable in further analyses. Brzinsky-Fay and Kohler (2010) argue that these two types of approaches can be viewed as complementary rather than as competing.

In this paper, we compare two strategies to construct such holistic life course typologies. The first strategy, called Sequence Analysis (SA) (Abbott & Forrest, 1986; Cornwell, 2015), starts by calculating a distance measure over the set of sequences and then tries to partition the resulting distance matrix into clusters of trajectories. Sequence analysis and its related typology techniques have been widely applied in studying life course trajectories in the social sciences (e.g. Kleinepier, de Valk & Gaalen, (2015); Helske Steele, Kokko, Räikkönen & Eerola, (2014)). The second strategy uses a probabilistic model that describes an observed life course sequence of categorical values as resulting from the conditional probabilities that define membership of a latent class and is called Latent Class Analysis (LCA) (Hagenaars & McCutcheon, 2002). Barban and Billari (2012) suggested that LCA can be used as an alternative to SA to derive meaningful classifications of life course patterns. Barban and Billari (2012) demonstrated (see their Table 1) that SA and LCA could generate quite different typologies from the same data. This does not come as a surprise as SA and LCA use very different methodologies. Using SA implies selecting a distance measure followed by a clustering method to partition the distance matrix. Using LCA implies that a category or class is defined by a probability distribution function over a set of categorical observations like ’living single’ or ’getting married’: different classes are defined by different probability distribution functions over the same states. Thus, both methods imply quite different steps to generate a life course typology.

The main aim of this paper is to discuss how typologies derived from SA and LCA can be compared and their quality assessed. The tools introduced to allow this assessment are useful in a more general sense as well. They can also be used to decide between typologies generated by different distance metrics or different clustering methods in SA, and thus are of interest to all users of holistic life course methods. This paper also offers guidance to researchers who want to use SA and/or LCA in their research, by outlining the steps to be taken and the decisions to be made in performing an SA and/or LCA analysis and by discussing practical bottlenecks that often pop up.

The paper is structured as follows. First, we offer a description of the main steps to be taken in developing a typology using SA and LCA. Next, we discuss methods to compare the SA and LCA typology solutions and to decide on which particular solution is to be preferred. We illustrate these procedures by analysing data from the Family and Fertility Survey as presented in the Methods and Data section. Next, we present the results of our illustrative example and in our final section we draw conclusions about the more general implications of the suggested procedures. Three appendices are added. In Appendix 1, practical issues are discussed. In Appendix 2, we present a heuristic explanation of sequence generation that bridges the gap between SA and LCA. In Appendix 3, R-based commands are provided that can be used as a code-model for the analyses presented.

Sequence Analysis

Sequence analysis (SA) has become the key holistic method to study life course trajectories since Abbott (1983) introduced it in the social sciences. This section briefly outlines the necessary steps and decisions to arrive at an SA-based typology. A sequence dataset has to be constructed from life course data. In this paper, we organised the sequence data as a state-sequence dataset. Other methods have been discussed in Ritschard, Gabadinho, Studer and Müller (2009). The main idea behind SA is to express the dissimilarity between pairs of sequences as a distance. The larger the distance between two sequences, the more dissimilar they are (but see Elzinga & Studer, forthcoming). Therefore, the first decision in sequence analysis is about choosing an appropriate distance metric. The two main classes of metrics available are edit-based metrics and subsequence-based metrics. Edit-based metrics measure the distance between two sequences by counting the minimum number of (weighted) edit-operations required to turn one sequence into a perfect copy of the other. In the social sciences, these metrics (and their numerous variants) are known as ’Optimal Matching’ (OM) (Abbott & Forrest, 1986).

Edit-based metrics are rather insensitive to differences in the ordering of states (Elzinga & Studer, 2015). This motivated the development of so-called subsequence-based metrics (Elzinga, 2005; Elzinga & Wang, 2013). These metrics measure the distance between sequences by counting the number of (weighted) common subsequences. For a detailed review of distance metrics for SA, we refer to Robette and Bry (2012), Studer (2012) and Studer and Ritschard (2016). In our illustration, we only present the SA approach with an OM-metric. The reason for this choice is twofold. First, family-formation patterns in modern Western societies vary relatively little in the ordering of events (Billari & Liefbroer, 2010). This makes OM a quite natural choice. Today, OM is the most commonly used metric in studies on the transition to adulthood (Aassve, Billari & Piccarreta, 2007; Brzinsky-Fay, 2007; Robette, 2010). Second, preliminary analyses showed that in this particular illustration, the OM metric clearly outperformed other metrics. The code in Appendix 3 is easily adaptable to any of the other metrics offered through R-based software.

The computation of distances between all sequences results in a distance matrix. The second step in SA uses this distance matrix to partition sequences into more or less homogeneous groups. Various clustering methods are suitable for this purpose, including hierarchical clustering (Maimon & Rokach, 2005), partitioning around medoids (PAM) (Kaufman & Rousseeuw, 2009), and self-organising maps (SOM) (Massoni, Olteanu & Rousset, 2009). Among them, Ward's method is most widely used (Aassve et al. 2007; Billari & Piccarreta, 2005; Pailhé, Robette & Solaz, 2013).[i]

Ward's method (Ward Jr, 1963) iteratively merges ever-bigger clusters of sequences such that, in each iteration, the increase of the total within-cluster distance is minimised. Critical in this step is to determine at what level of agglomeration, i.e. at which number of clusters, to stop the merging process, as this number is not determined by the method itself. The number of clusters has to be decided upon by applying a combination of substantive theory and measures of statistical cluster quality. Substantive theory alone may not adequately summarise the observed heterogeneity of life course patterns and the clustering algorithm may not lead to a parsimonious set of internally homogeneous and well-separated clusters.

We use three statistics (e.g. Table 2 in Studer, 2013) to empirically determine cluster quality: Average Silhouette Width (ASW), Hubert's C index and the Point Bi-serial Correlation (PBC). ASW (Rousseeuw, 1987), compares the average packing of points within clusters to the average distance of points to the closest cluster to which these points do not belong. A high ASW-value implies that clusters are homogeneous and well separated from each other. The HC index (Hubert & Levin, 1976) shows the gap between the partition obtained and the best partition theoretically possible with this number of groups. A low value of HC indicates good clustering. Finally, PBC (Milligan & Cooper 1985) measures the capacity of the cluster solution to reproduce the original distance matrix. A high PBC value is preferred.

Ideally, one would not only want to know what the optimal number of clusters is, but also their stability. To evaluate stability, most statistical analyses involve not only the estimation of model parameters, but also the estimation of their standard errors. In SA, this is not possible. However, once a theoretically and numerically acceptable cluster solution is obtained, one can examine its stability to data sampling fluctuations by using bootstrap methods. Such bootstrapping (e.g. James, Witten & Tibshirani, 2013) allows examining whether the clustering algorithm returns the same solution across several sub-samples. The clusterwise Jaccard Bootstrap Mean (CJBM) (Hennig, 2007), is a measure that uses the bootstrap to re-sample the data and to compute the Jaccard similarities of the original clusters to the most similar clusters in the re-sampled data. As proposed by Hennig (2008), when CJBM is below 0.6, the cluster solution should not be trusted. If CJBM is above 0.85, the classification technique generates highly stable clusters. A CJBM between 0.6 and 0.85 suggests some structure, but exact cluster membership is uncertain.

Cluster quality and stability measures have to be combined with theoretical interpretation for sound typology decisions. Therefore, the last step in creating an SA typology is to provide a substantively meaningful interpretation of the clusters. Visualisation tools such as sequence index plots (Scherer, 2001) and sequence medoid plots (Gabadinho, Ritschard, Müller & Studer, 2011) facilitate interpretation of cluster solutions. In sequence index plots each sequence is represented by a line composed of differently colored segments, with colors representing states and the length of the segments being proportional to the time spent in a state. A sequence index plot summarises large amounts of information in a single graph: order, prevalence and timing of states and overall variability within and between sequences. The medoid sequence is an observed sequence whose average distance to all the other sequences in a cluster is minimal.

The SA-cluster solution is affected by many factors: the sequence encoding, the choice of a metric, and the choice of a clustering technique. Here, we present no sensitivity analyses of our results since our purpose is mainly to elaborate on the global methodologies of applying and comparing SA and LCA.

Latent Class Analysis

Latent Class Analysis (LCA)[ii] is a statistical technique for the analysis of multivariate categorical data (see e.g. Hagenaars & McCutcheon, 2002). To concisely explain the LC-model, we need some concepts and notation. First, we write $y = y_1 \dots y_n$ to denote an observed sequence of length n. Second, we denote the latent class model Θ as a set of R conditional probability distributions $T = \{\theta_1, \dots \theta_R\}$ over the observable states, each of these characterising precisely one of the R latent classes. Furthermore, the model needs a specification of the probability that a sequence is generated from any of these latent classes, the vector $P = (\pi_1, \dots \pi_R)$ wherein π_j denotes the probability that a sequence is generated from θ_j. Thus a complete LC-model can be specified as $\Theta = (R, T, P)$. The LC-model states that the probability of observing a particular sequence, given the model, equals

$$Prob(y|T) = \sum_{r=1}^{R} \pi_r \prod_{i=1}^{n} Prob(y_i|\theta_r).$$

So, the model states that, given a fixed latent class r, the consecutive observed states are statistically independent and this assumption is known as ‘local independence’. This mixture model (e.g. McLachlan & Peel, 2000) is closely related to supervised Naive Bayes classifiers (Hand & Yu, 2001; Vermunt & Magidson, 2003). Despite the highly implausible assumption of local independence (Rennie, Shih & Karger, 2003), such models often perform quite well for classification tasks because dependencies often are equal across classes or cancel out (Zhang, 2005). Of course, some sequences may be extremely (un-)likely to be generated from some of the latent classes. If the number of classes is well chosen, each observed sequence is relatively (much) more likely to have been generated from one particular latent class than from any of the other classes. Therefore, class membership of each specific sequence is often decided by assigning the sequence to the class with the highest probability of generating that sequence.

Local independence implies that, for each latent class, observing the sequence $aaabbb$ is precisely as likely as observing the sequences $bbbaaa$ or $ababab$ or any other of the 20 discernable permutations of these six observations. So, local independence is a counter-intuitive assumption in the context of modelling life course sequences. Indeed, as can be observed from Figure 2, life courses mainly differ in the timing and selection of states, not in the orderings of states. The assumption of equal probability of observing any ordering of a given collection of states arises from the assumption that, given a conditional distribution θ_i, the observable states are generated by just sampling the alphabet of states according to θ_i. The observation that order-differences are rare (Figure 2) in fact constitutes a statistical test of this assumption: the assumption should be rejected because of observing so little variation of state-orderings. Therefore, applying the LC-model to describe life courses requires an interpretation of that model that, on the one hand, includes the assumption of local independence but that, on the other hand, makes variation in the ordering of states implausible. Such an interpretation is amply discussed in Appendix 2: it is assumed that different classes arise by different “template sequences” that are edited, state by state, such that

1. successive edits are statistically independent and
2. edits resulting in an actual change of a template-state are implausible.

So, in this interpretation, sampling observable states from class-conditional distributions is replaced by sampling of edits and applying them to class-conditional templates. If one additionally assumes that there is a unique, most likely path of edits, OM-distances between pairs of sequences are monotone with the probability that these pairs result from (editing) the same class-specific template. If this assumption is valid, the observed scarcity of order differences within classes is explained. In Appendix 2, we detail and formalise this interpretation and the unifying SA-LC assumption of one dominant edit path.

Just like in cluster analysis, when using LCA, one has to decide on the optimal number of classes. The number of latent classes to a large extent determines the fit of the model: the more latent classes, the “easier” it becomes to accommodate the diversity of the observed sequences. As the number of classes increases, the likelihood of the model generating the sequences increases, but at the risk of fitting to noise and at the expense of estimating more model parameters. Although the LCA model itself does not automatically determine the number of latent classes, a variety of goodness of fit statistics are available (Lin & Dayton, 1997).

Using statistics such as BIC and relative entropy (Vermunt & Magidson, 2013), one can gain information about model fit against the number of latent classes. One usually looks for a minimum in the BIC-curve. Relative entropy is between zero and one, with values near one indicating high certainty in classification and values near zero indicating low certainty. Here, we will use information on both BIC and relative entropy.

Like in SA, visualisation tools can be used to facilitate the interpretation of the latent class solution. By estimating, separately for each latent class, the sequence state with the highest frequency at each time point, we construct a model state sequence for each latent class. The resulting sequence model state plot in LCA is comparable to the medoid plot in SA in the sense that it aims at summarizing the key features of a cluster. However, whereas in SA the medoid is an actually observed sequence, the model state sequence in LCA might be a non-existing sequence. An interpretation of the latent class can also be obtained through visual inspection of the sequence index plot. Thus, the key decision in LCA to be made is on the number of latent classes and their interpretation can be aided by sequence index plots and sequence model state plots.

Typology Comparison

Both SA and LCA can assist researchers to detect structures in sequence data by segmenting the life course sequences into clusters or classes. However, the most basic distinction between SA and LCA is the way in which the classes or types are defined. An SA typology is based on a distance measure, a clustering procedure and a set of statistics that determine the quality of the cluster solution, while an LCA typology is obtained via the maximisation of a likelihood function that derives from a probabilistic model. Two questions arise in this context: (i) how similar are both typologies and (ii) can we decide on which solution has to be preferred?

A number of tools are available to judge how similar the two solutions are. The simplest tool is a cross tabulation of both typologies. Let S be a set of N data-items and let U denote an SA-typology over S: $U = \{U_1, U_2, \dots, U_R\}$ with $U_i \cap U_j = \emptyset$; and $\cup_i U_i = S$. Similarly, let V denote an LCA-typology over S: $V = \{V_{1,} V_2, \dots, V_C\}$ with $V_i \cap V_j = \emptyset$; and $\cup_i V_i = S$. An $R \times C$ cross-tabulation summarises the overlap between the two typologies by listing the numbers $n_{ij} = |U_i \cap V_j|$. A quantification of the overlap or agreement between two classifications can be achieved through the Rand index. The Rand index (Rand, 1971) is built upon counting pairs of items on which two typologies agree or disagree. As used above, the N item pairs in S can be classified into one of four types N_{11}: the number of pairs of which the members are in the same cluster in both U and V; N_{00}: the number of pairs of which the members are in different clusters in both U and V; N_{01} : the number of pairs of which the members are in the same cluster in U but in different clusters in V; and N_{10}: the number of pairs of which the members are in different clusters in U but in the same cluster in V. These numbers can be calculated using the n_{ij}. N_{11} and N_{00} can be used as indicators of agreement between U and V. The Rand index (R) is defined as: $R = (N_{00} + N_{11})/(N_{00} + N_{01} + N_{10} + N_{11})$, which ranges from 0 (no pair classified in the same way under both typologies) to 1 (identical typologies). In our illustration, we will adopt the adjusted Rand index (Hubert & Arabie, 1985) used by Barban & Billari (2012) to compare typologies.

The adjusted Rand index provides us with information about how different the two typologies are, but does not provide any clue about whether one or the other typology is superior. We suggest that the concept of construct validity as developed in psychometrics (Cronbach & Meehl, 1955; Ross, Wright & Anderson, 2013) can be fruitfully applied to decide which one of a set of alternative typologies is preferable. A typology can be viewed as a theoretical construct that is always part of a larger nomological network, i.e. a set of relationships between the concept of interest (the typology) and other concepts. If one has expectations about the statistical relationship between a typology and other concepts, one can measure the strength of these statistical relationships for each of the typologies. The stronger these relationships, the more likely it is that the measure of a construct is valid. This approach is often used in psychology to assess the validity of psychological constructs (e.g. Leary, Kelly, Cottrell & Schreindorfer, 2013). The nomological network approach is illustrated in Figure 1. A life course typology is part of a nomological network that specifies how this typology is related to other variables, either determinants (x_1….x_i) or consequences (y_1…y_i). Given our knowledge about the expected relationships in this network, one can assess how strongly alternative typologies are related to the rest of the nomological network. In general, one would prefer the typology that shows the strongest associations with other variables in the network. Given that we can use substantive information about the expected relationship between a life course typology and related concepts (e.g. levels of education, religiosity and well-being), this approach offers an elegant, theory-driven solution to the dilemma of deciding between alternative life course typologies.

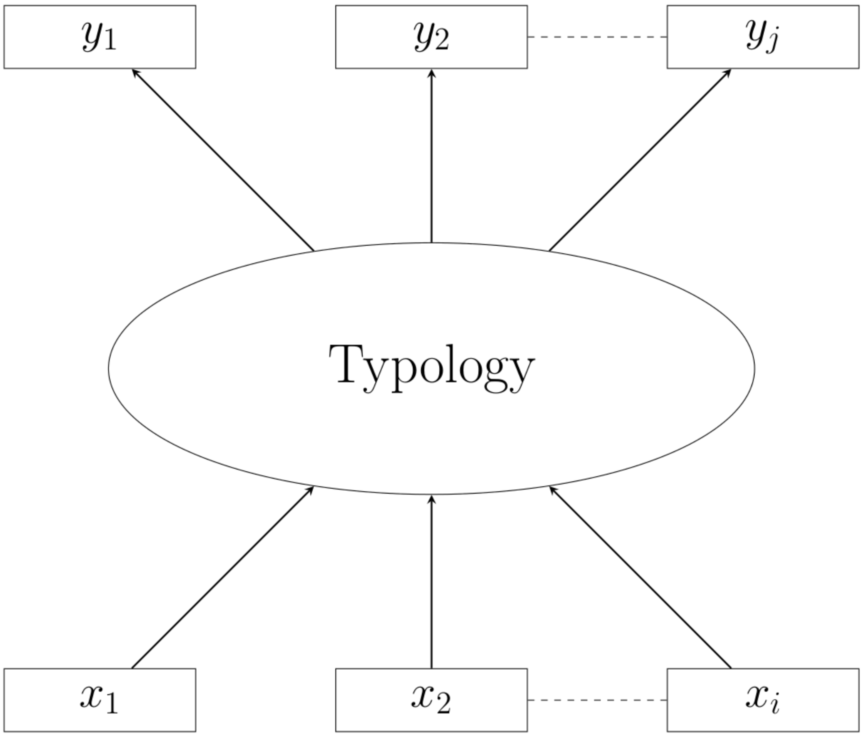


Figure 1. A representation of a nomological network surrounding a life course typology

The statistical application of this idea of construct validity depends on the kind of relationships within the nomological network. If one examines relationships between the typology and consequences, one can use linear or logistic regression, depending on the measurement level of the dependent variables. If one examines the relationship between typologies and their determinants, the easiest test of construct validity would be to estimate separate multinomial logit models with each of the competing typologies as the dependent variable, and a joint set of predictors. What complicates matters here, is that one cannot simply compare the fit-statistics of such models, as the dependent variable (class membership) differs between typologies, and thus one cannot use standard indicators of model fit. However, an alternative procedure is possible by swapping the dependent and independent variables, and predicting the available background variables from the SA- and LCA-generated typologies. Given that the dependent variables are the same now, one can use BIC and other fit-indices to judge which typology is more strongly related to the background variable of interest. One can repeat this procedure for multiple background variables, and decide upon the best typology by comparing the sets of BIC values for the alternative typologies. Alternatively, one can use a MANOVA-approach to compare the quality of the different typologies. Here, we do not use such a multivariate MANOVA-approach because the underlying distributional assumptions are hard to test thoroughly. However, such multivariate tests and the instruments for analysing the power of such tests are readily available (e.g. Faul, Erdfelder, Lang & Buchman, 2007) to the interested reader.

Methods and Data

Data

We used a subset of the Family and Fertility Survey (Festy & Prioux, 2002). This subset (Elzinga & Liefbroer, 2007) includes 10,301 female respondents from 15 countries born between 1960 and 1964. Full monthly event history information was available regarding respondents' fertility and partnerships between ages 18 and 30, their country of birth, years of education after age 15, religion and parental divorce, as shown in Tables 1 and 2.[iii]

Table 1. Definition of social background variables used in the typology comparison

Abbreviation	Meaning
Edu1	No education after age 15
Edu2	0-3 years of education after age 15
Edu3	3-5 years of education after age 15
Edu4	5+ years of education after age 15
Pardiv0	Parents not divorced
Pardiv1	Parents divorced
Pardiv3	Parents' divorce not known
Reli0	Not religious
Reli1	Catholic
Reli2	Protestant
Reli3	Other religion
Reli4	Religion unknown

Table 2. Number of respondents per country and percentage of the respondents per category of the social background variables

	Region	edu1	edu2	edu3	edu4	pardiv0	pardiv1	pardiv3	reli0	reli1	reli2	reli3	reli4	Nr. Resp.
1	Estonia	9.51	48.59	41.20	0.70	66.20	22.18	11.62	47.54	0.00	41.20	11.27	0.00	284
2	Czech Republic	5.10	23.81	50.34	20.75	85.37	14.29	0.34	--	--	--	--	--	294
3	France	8.02	44.32	22.94	24.72	87.53	10.91	1.56	--	--	--	--	--	449
4	New Zealand	--	--	--	--	--	--	--	--	--	--	--	--	460
5	Hungary	21.58	30.56	25.85	22.01	85.04	14.53	0.43	39.10	47.65	8.12	3.21	1.92	468
6	Latvia	0.85	22.67	39.41	37.08	77.12	19.07	3.81	31.78	20.55	17.37	26.69	3.60	472
7	Lithuania	1.17	15.37	33.46	50.00	80.54	18.09	1.36	8.17	80.93	0.78	8.56	1.56	514
8	Slovenia	18.02	20.14	32.69	29.15	92.40	7.42	0.18	21.02	68.90	0.18	8.66	1.24	566
9	Netherlands	4.08	35.85	22.84	37.22	85.02	9.98	4.99	39.94	34.80	19.21	5.90	0.15	661
10	Spain	36.95	22.36	10.84	29.85	97.32	2.68	0.00	17.40	78.05	0.54	3.08	0.94	747
11	Austria	20.88	30.05	33.51	15.56	90.03	9.44	0.53	30.72	59.18	3.59	6.38	0.13	752
12	Canada	--	--	--	--	82.72	15.31	1.96	3.80	46.73	33.77	15.71	0.00	764
13	Italy	30.47	15.48	18.43	35.63	97.79	2.21	0.00	8.72	90.05	0.49	0.61	0.12	814
14	Portugal	--	--	--	--	94.38	5.18	0.44	--	--	--	--	--	908
15	U.S.A.	50.65	0.28	8.47	40.60	75.84	24.07	0.09	8.47	29.05	49.67	12.71	0.09	2148

The table is ordered by the total number of respondents per region

"–": Data not available

We organised the sequence data as a state sequence (STS) data set (Ritschard, Gabadinho, Studer & Müller, 2009). STS is a chronologically ordered list of the states based on the survey information. We distinguish six family formation states: living single (S), unmarried cohabitation (U), marriage (M), living single with a child/children (SC), cohabitation with a child/children (UC), and marriage with a child/children (MC). Therefore, the sequence data consist of 144 monthly family-life statuses; an example from one respondent is shown below.

$$\overbrace{S \ldots S}^{87}\ \overbrace{M \ldots M}^{56}\ \overbrace{MC \ldots MC}^{11}$$

This person has first spent 87 months in the Single state, followed by 56 months in the Married state and 11 months in the Married with Children state.

Methods

All methods introduced in the previous sections were applied to data set. All analyses were performed in the R software environment for statistical computing and graphics on a 3.2 GHz CPU, 32GB RAM and 64-bit PC, using the R packages TraMineR (OM), stats (hierarchical clustering), WeightedCluster (cluster decision), fpc (bootstrapping), poLCA (LCA), flexclust (Rand index), and nnet (multinomial logistic regression).

Results

SA Typology

Following the steps outlined earlier, we first compute a distance matrix using the TraMineR software (Gabadinho et al., 2011). There are many different metrics to construct distances between sequences. The choice for either of these metrics may affect the nature of the resulting typology. So, one should be aware of the differences between these metrics (see Robette & Bry (2012), Elzinga & Studer (2015) and Studer & Ritschard (2016) for a detailed discussion of these issues). Here, we experimented with a variety of distance measures: OM with various cost settings and a sequence-based vector representation (Elzinga & Studer, 2015) with various parameter settings. OM with indel-cost of 4 and substitution-cost of 2 (the default setting of TraMineR) generates the best solution. Here, we present the results for the selected OM-metric only. The cost setting used implies that all substitutions were equally costly and that mere deletion or insertion never occurred. The solutions for other distance measures can be obtained from the first author upon request.

In our example, we use, for reasons already explained, hierarchical clustering (Ward's method). In Table 3, values of ASW, PBC, and HC are presented for solutions with two to eight clusters. The values in Table 3 show that a solution with six clusters is favored, as this solution combines the highest values of ASW (0.63) and PBC (0.35) with the lowest value of HC (0.10). The next step is to test the stability of the cluster solution by using the CJBM statistic. This statistic indicates to what extent sequences are likely to be assigned to the same cluster over a large number of random draws from the sample. The higher this likelihood, the more stable the cluster solution is. The CJBM statistic for the six clusters of the SA-6 solution varies between 0.45 and 0.76, falling short of the 0.85-level, which is considered to indicate high cluster stability. This suggests that quite a few clusters are not very stable. Table 4 shows that for the SA-7 solution, the CJBM statistics vary between 0.42 and 0.76.

Table 3. Values of cluster quality statistics

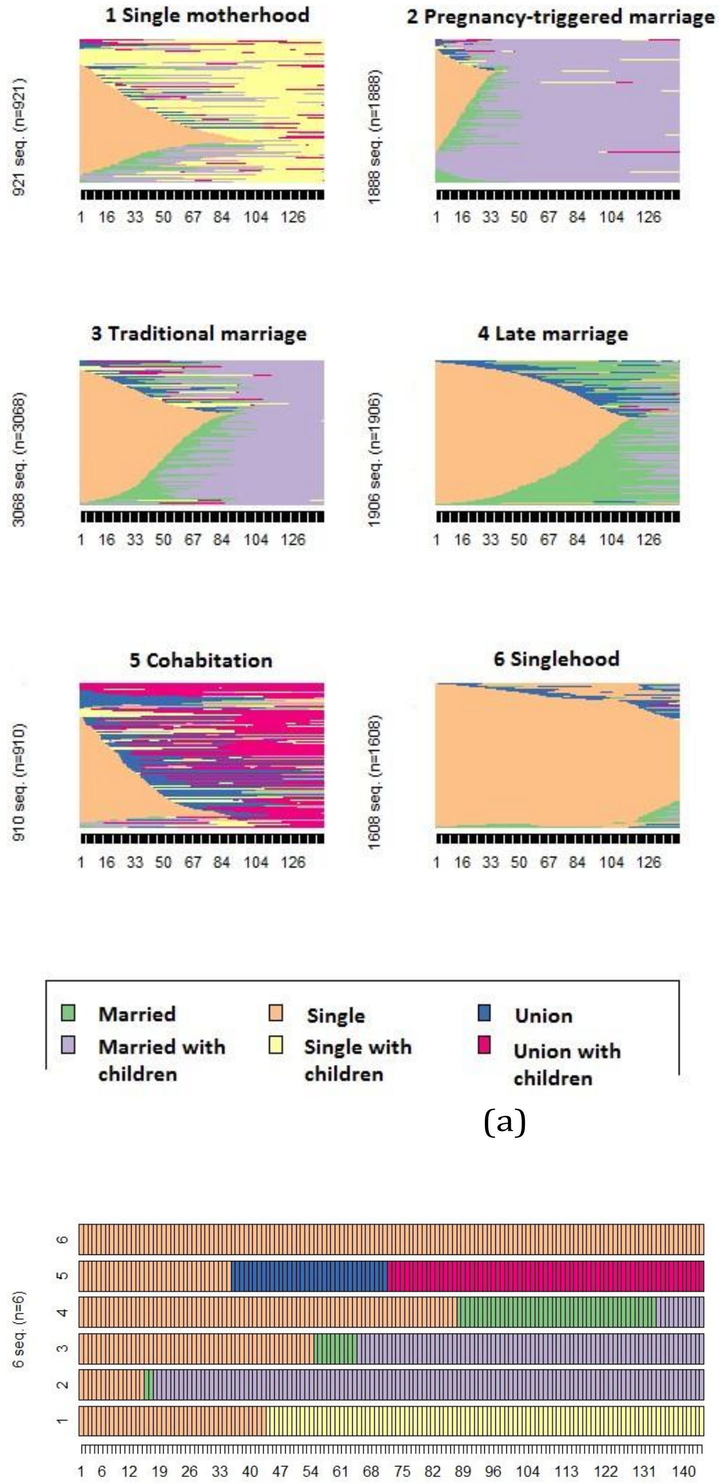
Number of clusters	PBC	ASW	HC
2	0.48	0.34	0.21
3	0.49	0.29	0.22
4	0.62	0.34	0.13
5	0.57	0.31	0.14
6	0.63	0.35	0.10
7	0.59	0.33	0.11
8	0.58	0.33	0.10

Note: PBC (maximal value preferred), ASW (maximal value preferred) and HC (minimal value preferred)

Table 4. Values of CJBM statistics of all six clusters of the OM optimal solution

CJBM	Cluster 1	Cluster2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
SA-6	0.57	0.66	0.54	0.49	0.45	0.76	NA
SA-7	0.62	0.69	0.54	0.42	0.48	0.46	0.76

The final step is to interpret the cluster solution. Figure 2a presents a sequence index plot to show how individuals within each cluster move between states over time. Figure 2b shows a sequence medoid plot that presents the sequence with the smallest average distance to the other sequences in the pertaining cluster. In Figure 2a, cluster 1 mainly consists of sequences starting as ’single’ (S) followed by a transition to ’single with children’ (SC). This suggests that a meaningful label to this cluster could be ’single motherhood’, which is confirmed by the sequence medoid plot shown in Figure 2b. Figure 2b shows that the medoid sequence of cluster 1 spent a spell of 43 months as Single (S), followed by a spell of 101 months being single with children (SC). Based on the interpretation of these plots, labels are assigned to all six clusters. Next to the ’single motherhood’ cluster (8.9% of the cohort), we found a cluster that we label ’pregnancy-triggered marriage’ (18.3%) as these sequences often have less than nine months between marriage and parenthood, a cluster labeled ’traditional marriage’ as marriage is usually followed rather soon by motherhood (29.7%), a cluster labeled ’late marriage’ (18.5%), a cluster labeled ’cohabitation’ as many of these sequences are characterised by spells of cohabitation either with or without children (8.8%), and finally a cluster labeled ’singlehood’ (15.6%).



(a)

(b)

Figure 2. Sequence Index plot (a) and Sequence Medoid plot (b) of the OM-6 solution

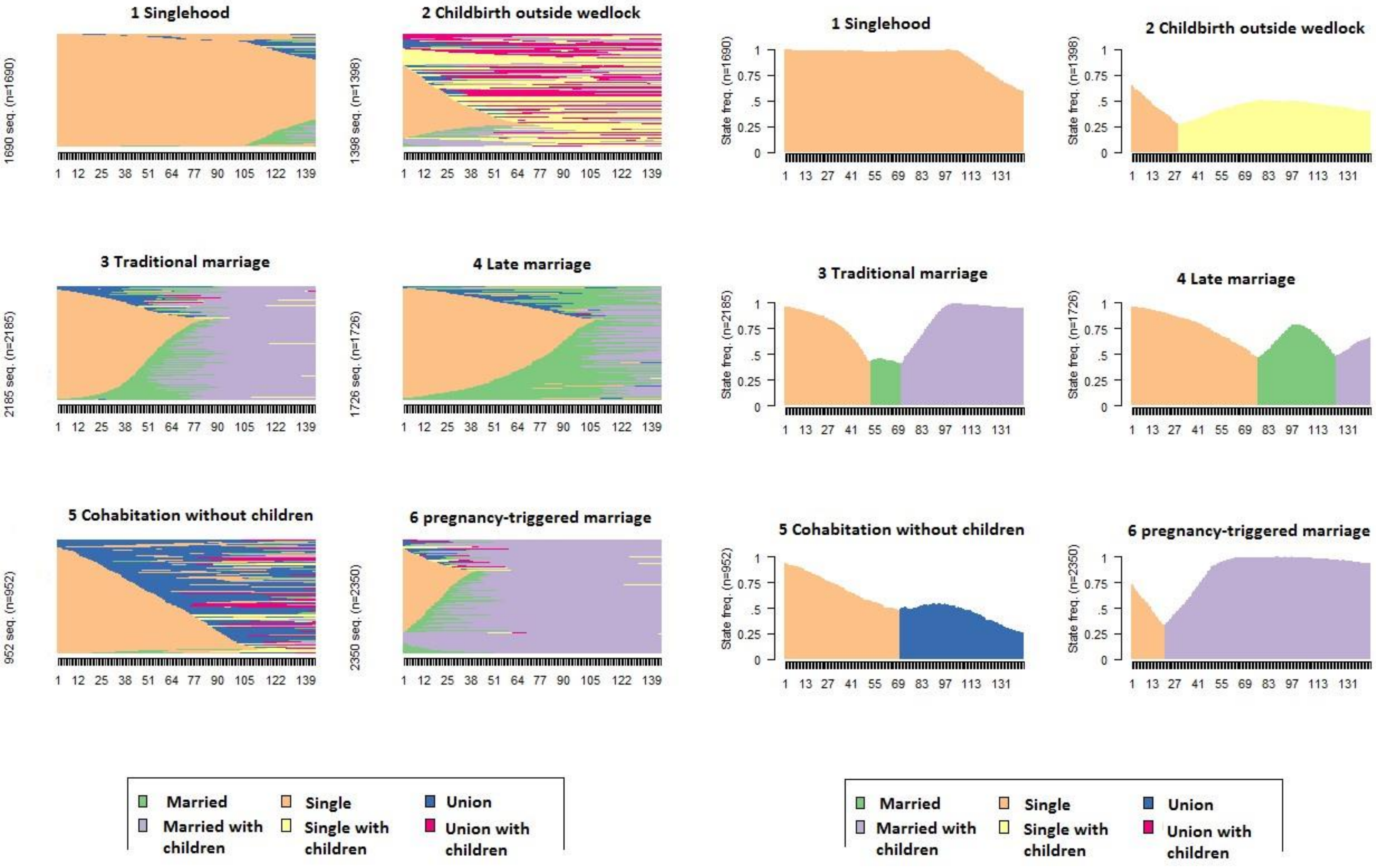
LCA Typology

The first step in a LCA is to decide on the number of classes to generate. In a stepping-stone paper, Frahley & Rafterey (1998, Figs. 3 and 4) demonstrated how BIC can be used for cluster-model comparison. Within one type of covariance-structure, these authors plotted the BIC for different numbers of clusters, expecting the BIC to first decrease and then increase again around the optimal number of clusters. Here, following the principle set out by these authors, we compare LC-models that only differ in their number of clusters. So, theoretically, one could expect BIC first to drop drastically with the increase in the number of classes, followed by a slow decrease and finally by an increase again, the latter due to the large number of parameters estimated. High relative entropy (close to 1) indicates good model fit and therefore a desirable number of latent classes. In Table 5, the observed BIC and relative entropy values of the 2- to 8-class solutions are presented. The BIC and relative entropy values of the LCA typology in our example do not exactly show the expected pattern. One observes a drastic decrease in BIC values up to about a five-class solution and a much slower decline up till an eight-class solution. Relative entropies for all solutions are close to 1, suggesting high certainty in classification. Given that all entropies are close to 1, it is hard to base model-selection on relative entropy. The decrease in BIC obviously slows down after five classes, and we decided to examine the six- and seven-class solutions in more detail.

Table 5. Values of latent class analysis model fit statistics: BIC (minimal value preferred), and relative entropy (closest to one value preferred)

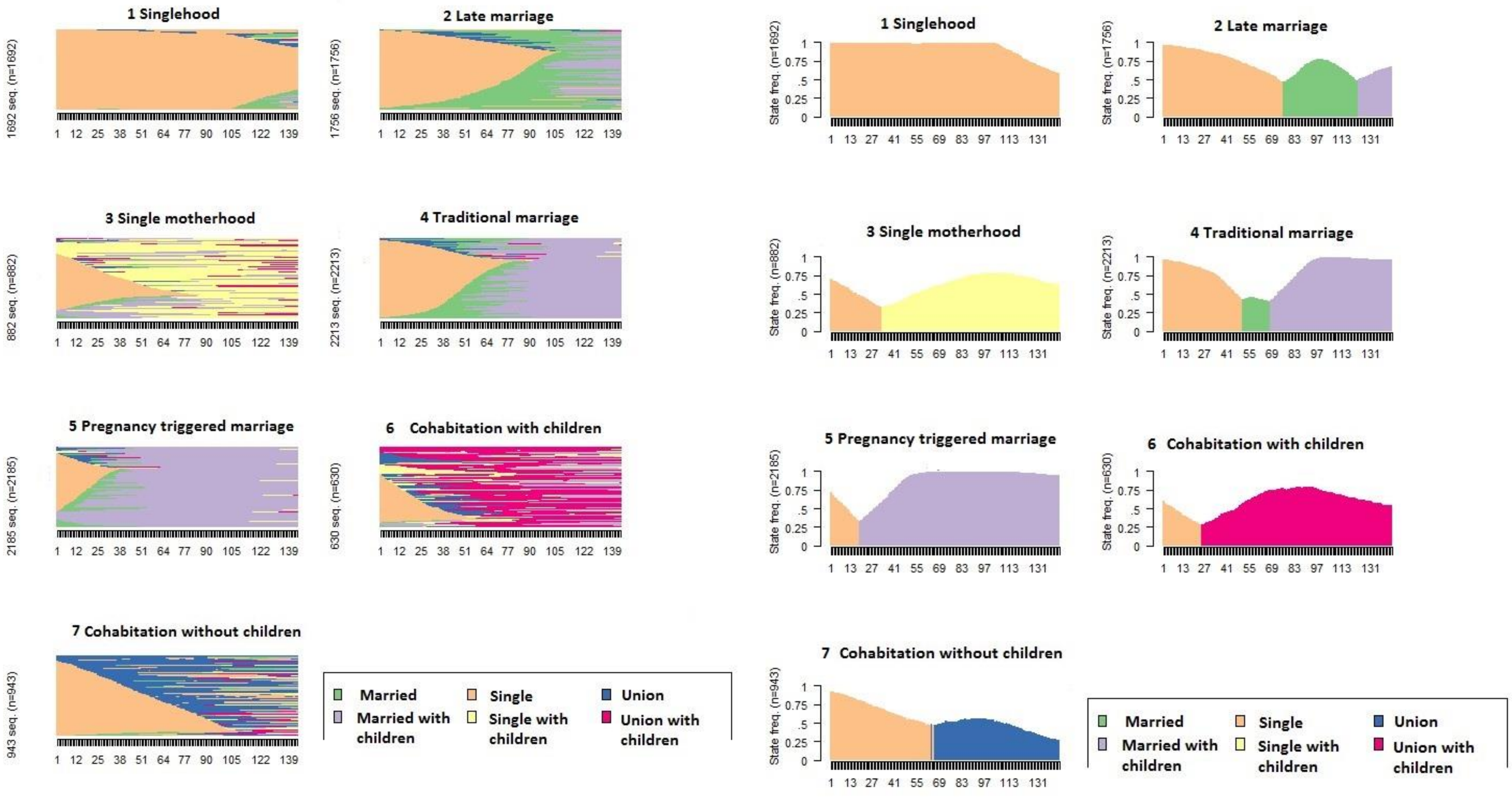
Number of clusters	BIC*10^6	relative entropy
2	3.0	0.9993
3	2.6	0.9992
4	2.3	0.9982
5	2.1	0.9980
6	2.0	0.9979
7	1.9	0.9976
8	1.8	0.9975

The interpretation of the LCA typology can be facilitated by sequence index plots and sequence model state plots (Figures 3 and 4). The six-class solution (for short: LCA-6, presented in Figures 3a and 3b) partitions female respondents into classes that we interpret as ’singlehood’ (16.4%), ’childbirth outside marriage’ (13.6%), ’traditional marriage’ (21.2%), ’late marriage’ (16.8%), ’cohabitation without children’ (9.2%), and ’pregnancy-triggered marriage’ (22.8%). LCA-7 (Figures 4a and 4b) generates five classes that are quite comparable to LCA-6, namely ’singlehood’ (16.4%), ’late marriage’ (17.0%), ’traditional marriage’ (21.5%), ’pregnancy-triggered marriage’ (21.2%), and ’cohabitation without children’ (9.2%). The main difference with LCA-6 is that instead of one class dominated by sequences with children outside marriage, there are now two classes. One of them can be interpreted as ’cohabitation with children’ (6.1%), and the other as ’single motherhood’ (8.6%). As it is hard to judge which class number is optimal, we decide to compare both solutions to the SA-6 solution in the next section.



(a) (b)

Figure 3. Sequence Index plot (a) and Sequence Model State plot (b) of the LCA-6 solution



(a) (b)

Figure 4. Sequence Index plot (a) and Sequence Model State plot (b) of the LCA-7 solution

Typology Comparison

As a first step in the typology comparison, we simply compare the frequency distributions of each cluster or class in SA and LCA solutions. SA-6 and both LCA typologies have four major clusters in common. These are ’pregnancy-triggered marriage’, ’late marriage’, ’traditional marriage’ and ’singlehood’. As shown in Figure 5, the traditional marriage cluster is somewhat larger in the SA-6 solution (30%) than in either of the LCA solutions (around 21%). In reverse, the pregnancy-triggered marriage cluster is somewhat smaller in SA-6 (18%) than in both LCA solutions (around 22%). The late marriage cluster (17%) and the singlehood cluster (16%) roughly match each other in all three typologies. The smaller clusters differ considerably, both between the SA and LCA solutions, and between the two LCA solutions. In SA-6, the two smaller clusters are single motherhood (9%) and cohabitation (9%). In the LCA-6 solution, the two smaller clusters are interpreted as childbirth outside marriage (14%) and cohabitation without children (9%). In the LCA-7 solution, the three smaller clusters are labeled as single motherhood (9%), cohabitation with children (6%) and cohabitation without children (9%). Note that the cohabitation cluster in the SA solution includes both cohabiters with and without children. Besides, the LCA-6 solution combined single motherhood and cohabitation with children into one class. Only in the LCA-7 solution do these three groups form separate clusters. Thus, the key difference between the three classifications is the way that cohabiters with children are classified. They are classified separately in the LCA-7 solution, grouped together with single mothers in the LCA-6 solution and grouped together with cohabiters with children in the SA-6 solution.

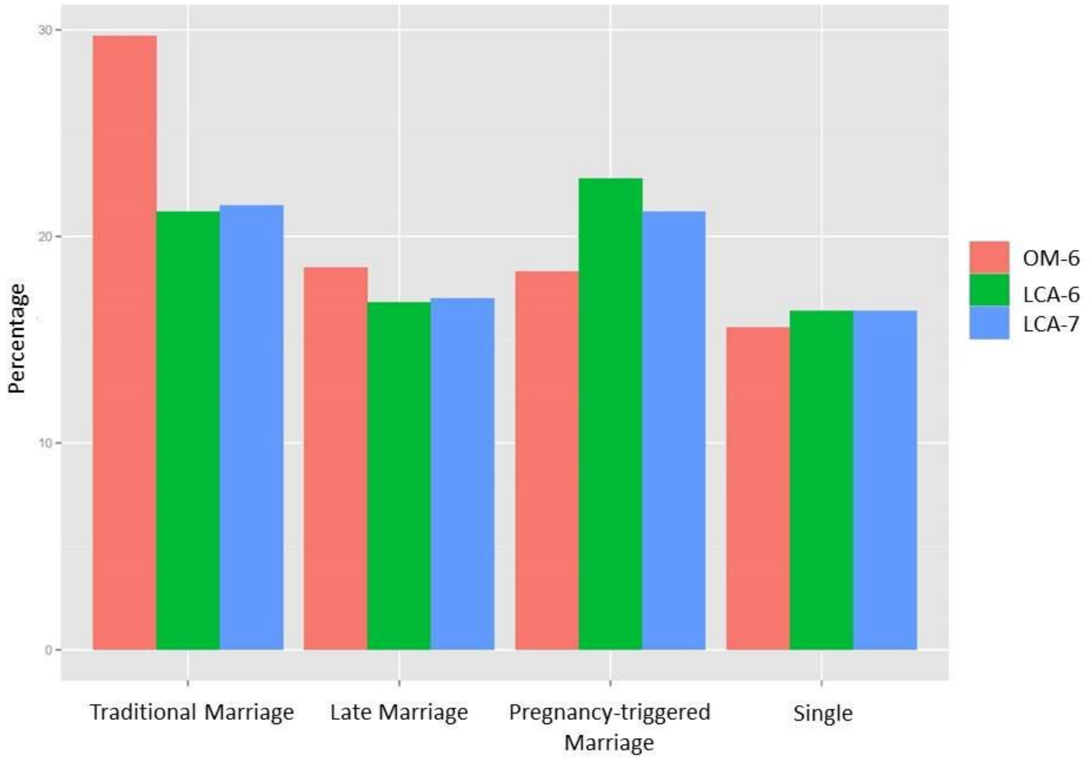


Figure 5. Percentage of respondents in the four large clusters or classes of OM-6, LCA-6 and LCA-7

Given that the substantive interpretation and labeling of SA-6 and LCA-7 were quite similar, we use these cluster solutions to illustrate the usefulness of cross-tabulation (Table 6). If the typologies would be exactly the same, there would be a permutation of its rows and columns such that only the diagonal elements would have positive values, whereas the rest of the table would be empty; when the numbers of classes in the two typologies differ, one may expect that members of one class in the one typology are distributed over a small number of classes in the other typology

(Agresti, 2001). Five classes in SA-6 and LCA-7 received the same label. In Table 6, we observe that most of the sequences that were assigned to these classes in the SA solution were assigned to the same class in LCA-7. In all, 73% of sequences are assigned to clusters with roughly the same substantive interpretation in both analyses. The major difference between SA-6 and LCA-7 is that the former only has one cluster with sequences dominated by cohabitation, whereas the latter contains two clusters, one dominated by sequences of cohabitation without children and one dominated by sequences of cohabitation with children. Table 6 shows that the sequences that are assigned to the cohabitation cluster in SA-6 are almost equally split between the two cohabitation clusters in LCA-7. Four additional differences are found between SA-6 and LCA-7. Of those sequences that are classified as traditional marriage in SA-6, 3.6% are classified as late marriage in LCA-7. Another 3.0% of the traditional marriage sequences in SA-6 are classified as pregnancy-triggered marriages in LCA-7. Thus, it seems that the traditional marriage cluster in SA-6 encompasses a broader range of marriages than the traditional marriage cluster in LCA-7. Similar results hold for the late marriage cluster in SA-6. About 3.3% of the sequences in this cluster are classified as cohabitation without children in LCA-7, and another 2.1% as single in LCA-7. A comparison of the sequence index plots in Figures 2a and 4a shows that some cohabiters who married at a relatively late age, are classified as late marriage in SA-6 but as cohabiters without children in LCA-7. Similarly, some respondents who had been single for most of the time and only married just before turning 30 are classified as late marriage in SA-6 and as single in LCA-7.

Table 6: Cross tabulation of LCA and SA typology solutions, values shown as percentages

LCA/SA	Cohab	Lmarriage	Pmarriage	Smothers	Single	Tmarriage
Cohab with c	4.14	0.00	0.04	0.45	0.00	1.50
Cohab without c	3.80	3.27	0.00	0.93	1.10	0.06
Lmarriage	0.12	13.03	0.00	0.11	0.17	3.62
Pmarriage	0.01	0.00	17.88	0.30	0.00	3.02
Smothers	0.75	0.08	0.32	6.46	0.00	0.96
Single	0.00	2.08	0.00	0.01	14.34	0.00
Tmarriage	0.03	0.05	0.09	0.69	0.00	20.63

Cohab = Cohabitation, Lmarriage = Late marriage, Pmarriage = pregnancy-triggered marriage, Smother = single mother, Tmarriage = Traditional marriage, Cohab with c = Cohabitation with children, and Cohab without c = Cohabitation without children

The adjusted Rand index for the cross-classification of SA-6 and LCA-6 is 0.59, 0.67 for the cross-classification of SA-6 and LCA-7, and 0.88 for the cross-classification of LCA-6 and LCA-7. Evidently, LCA-6 and LCA-7 have a large overlap, but also SA-6 and LCA-7 are highly comparable.

Which typology to prefer? Based on our discussion of construct validity, the strength of the statistical relationship between class membership and relevant background variables can be used to judge the quality of the typology. The stronger the relationship between class membership and other variables that are expected to be related to that typology, the better the typology performs. Therefore, we use class membership as the independent variable in multinomial logistic regression models predicting a series of relevant background variables. In this example, we use four external variables: level of education, parental divorce, level of religiosity and country. Some variables are not available for all countries, and respondents from these countries are excluded in the relevant analyses. To balance the analysis, we compare not only SA-6 with LCA-6 and LCA-7, but also add SA-7, even though the cluster quality statistics in Table 3 clearly favor SA-6 above SA-7. BIC's of all four typologies for all four dependent variables are presented in Table 7.

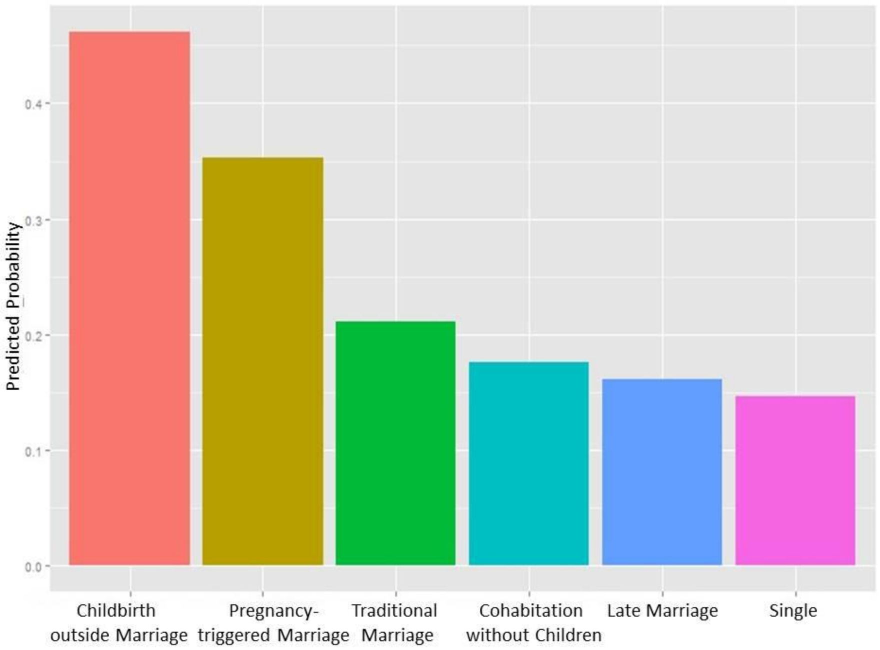
The BIC's of SA-6 are always lower than those of SA-7, implying that the added complexity of SA-7 does not improve the predictive power of the typology sufficiently to warrant the additional complexity.

Things are less clear-cut for the LCA typologies. Based on the BIC's for predicting parental divorce and religion, LCA-6 seems superior to LCA-7, while for predicting education and country, LCA-7 seems superior to LCA-6. To understand this, it is important to remember that in LCA-6, single motherhood and cohabitation with children are jointly classified in one class 'childbirth outside marriage', while in LCA-7, these are separate classes. For predicting parental divorce and religion, the distinction between single motherhood and cohabitation with children does not improve model fit, suggesting that those classified as single mothers and those classified as having a childbirth within cohabitation do not differ much in terms of their odds of experiencing a parental divorce or of being religious. However, for predicting education and country, distinguishing these two groups improves model fit. This suggests that those in the 'single motherhood' class and the 'cohabitation without children' class differ significantly from each other in terms of their distributions across countries and across levels of education. To provide a better interpretation of the meaning of this latter difference, we calculate predicted probabilities (Figure 6) of having no further education after age 15. In LCA-6, those classified as 'childbirth outside marriage' have a 47% chance of having no education after age 15. In LCA-7, this group is split and those classified as single mothers have a higher chance of no additional education (48%) than those classified as having a child within cohabitation (43%). Whether to prefer the LCA-6 or LCA-7 models, depends on how substantively meaningful these differences in educational distributions are. We view them as substantively meaningful, as they show that single motherhood is more strongly linked to social disadvantage than having a child within cohabitation, and thus we prefer LCA-7 to LCA-6.

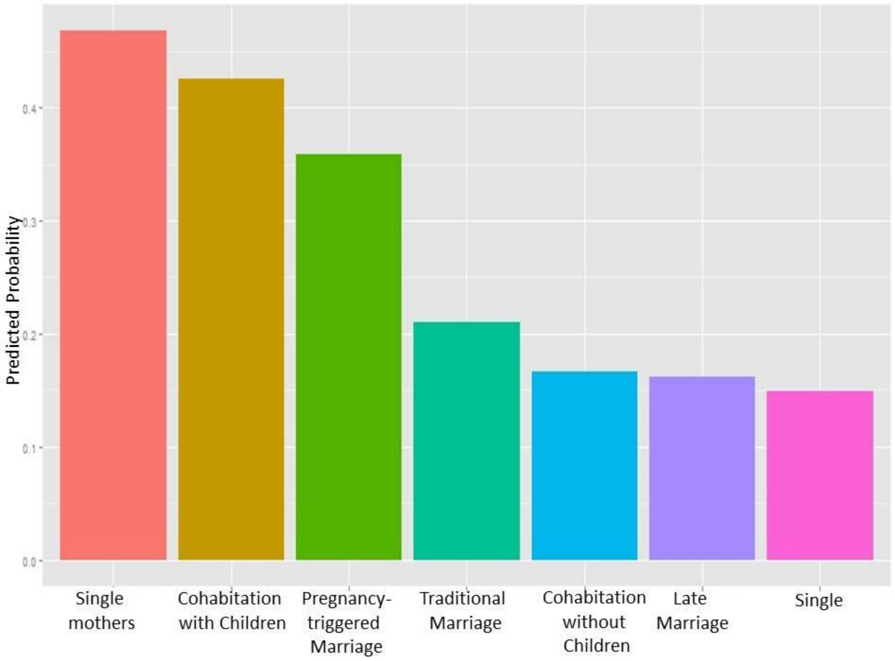
Finally, to choose between the SA-6 and LCA-7 classifications, BIC-values of the models with these two typologies were compared. Table 7 shows that the BICs for all background variables are lower for the LCA-7 typology than for the SA-6 typology. Therefore, we conclude that in in this *particular* illustration the LCA-typology is superior to the SA-typology. Overall, choosing LCA-7 seems the best decision.

Table 7. BICs of SA (OM 6 and OM 7) and LCA (LCA 6 and LCA 7), based on multinomial logistic regression models

	Education	Parental divorce	Religion	Country
OM 6	21612.34	8970.52	20308.92	51758.54
OM 7	21626.74	8986.59	20333.26	51813.44
LCA 6	21448.98	8909.06	20225.49	51583.76
LCA 7	21441.70	8927.17	20230.11	51442.98



(a)



(b)

Figure 6: Predicted probabilities of “no education after age 15” using a multinomial logistic regression model for (a) the LCA-6 solution and (b) the LCA-7 solution

Discussion and conclusion

The key question discussed in this article is whether quite different approaches to develop life course typologies, sequence analysis (SA) and latent class analysis (LCA), lead to the same typologies, and whether it is possible to decide on which of these typologies is to be preferred.

We emphasise three main contributions of this article. First, we suggest a number of statistical tools that aid decision-making about the optimal number of clusters or classes. We propose that one should make use of a combination of statistical information and substantive interpretation. The choice of the number of clusters in SA can be facilitated by using cluster quality statistics – whereas the use of BIC plots supports the choice process in LCA.

The second contribution of this paper is its suggestion to consider cluster stability as an important aspect of cluster quality.

Our third, and major, contribution consists of our proposal to validate the obtained typologies by examining their association with other variables that are known or expected to be related to the pertaining life course trajectories. This approach is based on the idea of construct validity that is central to measurement theory. Analogously, we argue that one can evaluate the quality of a typology by examining how strongly it relates to other variables within its nomological network. This validation approach is not confined to comparing SA and LCA, but can also be used to compare typologies obtained by using other distance metrics within SA or by using other clustering methods applied to the same metric.

Our presentation of SA and LCA was illustrated by a substantive comparison. As we emphasised throughout our presentation, many different decisions have to be taken within each approach, and each one of them has to be based on a combination of substantive and statistical evidence. In Appendix 1, the main practical challenges facing both methods are discussed.

Our example illustrated the steps to be taken when performing SA and LCA. One of the interesting results is that the resulting typologies were quite comparable: the adjusted Rand-index is close to 0.7. Thus, the question arises whether or not this result is a coincidence. To shed light on this question, we elaborated on an old idea of Joseph Kruskal, and present our findings in Appendix 2. Our reasoning suggests that SA and LCA will most often lead to the roughly the same typologies.

In a recent article, Mikolai and Lyons-Amos (2017) compared SA to Latent Class Growth Models (LCGM), a type of LCA that takes the temporal ordering of events into account. Theoretically, LCGM's have the advantage over the simple LCA model that the former incorporates the ordering of events in the life course whereas the latter does not. In practical terms, estimating LCGMs with a larger number of observable states and almost ten times as many respondents is practically infeasible However, Mikolai and Lyons-Amos obtained interesting results comparing LCGM to SA and also showed that in practice, results are roughly the same.

Summarising, we tried to expand upon the pioneering research by Barban & Billari (2012) by proposing guidelines on performing and comparing SA- and LCA-based typologies and by introducing a number of useful statistical tools to aid in choosing between competing typologies.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 324178 (Project: Contexts of Opportunity. PI: Aart C. Liefbroer). The authors would like to thank the anonymous reviewers for their stimulating comments to earlier versions of this article.

References

Aassve, A., Billari, F. C., & Piccarreta, R. (2007). Strings of adulthood: A sequence analysis of young British women's work-family trajectories. *European Journal of Population/Revue Européenne de Démographie, 23*(3-4), 369-388. https://doi.org/10.1007/s10680-007-9134-6

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 16*(4), 129–147. https://doi.org/10.1080/01615440.1983.10594107

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History, 16*(3) , 471-494. https://doi.org/10.2307/204500

Agresti, A. (2002) *Categorical Data Analysis* (2nd Edition). Wiley, NJ. https://doi.org/10.1002/0471249688

Bahl, L. R., & Jelinek, F. (1975). Decoding for channels with insertions, deletions and substitutions with applications to speech recognition. *IEEE Tra'nsactions on Information Theory, 21*(4), 404-411. https://doi.org/10.1109/TIT.1975.1055419

Barban, N., & Billari, F. C. (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. https://doi.org/10.1111/j.1467-9876.2012.01047.x

Bayne, C. K., Beauchamp, J. J., Begovich, C. L. & Kane, V. E. (1980). Monte Carlo comparison of selected clustering procedures. *Pattern Recognition, 12*(2), 51-62. https://doi.org/10.1016/0031-3203(80)90002-3

Billari, F. C., & Liefbroer, A. C. (2010). Towards a new pattern of transition to adulthood?. *Advances in Life Course Research*, 15(2), 59-75. https://doi.org/10.1016/j.alcr.2010.10.003

Billari, F. C. & Piccarreta, R. (2005). Analyzing demographic life courses through sequence analysis. *Mathematical Population Studies, 12*(2), 81-106. http://dx.doi.org/10.1080/08898480590932287

Brzinsky-Fay, C. (2007). Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe. *European Sociological Review, 23*(4), 409-422. https://doi.org/10.1080/08898480590932287409-422

Brzinsky-Fay, C., & Kohler, U. (2010). New Developments in Sequence Analysis. *Sociological Methods & Research, 38*(3), 359-364. https://doi.org/10.1177/0049124110363371

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with Stata. *The Stata Journal, 6*(4), 435-460.

Buchmann, M. C., & Kriesi, I. (2011). Transition to adulthood in Europe. *Annual Review of Sociology, 37*, 481–503. https://doi.org/10.1146/annurev-soc-081309-150212

Cornwell, B. (2015). *Social Sequence Analysis*. Cambridge (UK): Cambridge University Press. https://doi.org/10.1017/CBO9781316212530

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281. https://doi.org/10.1037/h0040957

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1-38.

Elzinga, C. H. (2005). Combinatorial representation of token sequences. *Journal of Classification, 22*(1), 87-118. https://doi.org/10.1007/s00357-005-0007-6

Elzinga, C. H. (2014). Sequence A152072. The On-Line Encyclopedia of Integer Sequences (2014), published electronically at http://oeis.org.

Elzinga, C. H., & Liefbroer, A. C. (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*(3-4), 225–250. https://doi.org/10.1007/s10680-007-9133-7

Elzinga, C. H., & Studer, M. (2015). Spell sequences, state proximities and distance metrics. *Sociological Methods & Research, 44*(1), 3-47. https://doi.org/10.1177/0049124114540707

Elzinga, C. H., & Studer, M. (forthcoming). Normalization of distance and similarity in sequence analysis. *Sociological Methods & Research*.

Elzinga, C. H., & Wang, H. (2013). Versatile string kernels. *Theoretical Computer Science, 495*, 50-65. https://doi.org/10.1016/j.tcs.2013.06.006

Faul, F., Erdfelder, E., Lang, A-G. & Buchman, A. (2007) G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191. https://doi.org/10.3758/BF03193146

Festy, P., & Prioux, F. (2002). An evaluation of the Fertility and Family Surveys project. United Nations Publications.

Frahley, C. & Raftery, A. E. (1998) How many clusters? Which clustering method? Answers via Model Based Cluster Analysis. *The Computer Journal, 41*(8), 578-588. https://doi.org/10.1093/comjnl/41.8.578

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1–37. https://doi.org/10.18637/jss.v040.i04

Greenberg, R. I. (2003) Bounds on the Number of Longest Common Subsequences. arXiv:cs/031030v2[cs.DM] e-print

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511499531

Hand, D.J. & Yu, K. (2001) Idiot's Bayes – Not so stupid after all. *International Statistical Review*, 69(3), 385-398.

Helske, S., Steele, F., Kokko, K., Räikkönen, E., & Eerola, M. (2014). Partnership formation and dissolution over the life course: applying sequence analysis and event history analysis in the study of recurrent events. *Longitudinal and Life Course Studies, 6*(1), 1-25. https://doi.org/10.14301/llcs.v6i1.290

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis, 52*(1), 258–271. https://doi.org/10.1016/j.csda.2006.11.025

Hennig, C. (2008). Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis, 99*(6), 1154–1176. https://doi.org/10.1016/j.jmva.2007.07.002

Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin, 83*(6), 1072. https://doi.org/10.1037/0033-2909.83.6.1072

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218. https://doi.org/10.1007/BF01908075

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kleinepier, T., de Valk, H. A., & Gaalen, R. van (2015). Life paths of migrants: A sequence analysis of Polish migrants' family life trajectories. *European Journal of Population, 31*(2), 155–179. https://doi.org/10.1007/s10680-015-9345-1

Kruskal, J. B. (1983). An overview of sequence comparison: time warps, string edits and macromolecules. *SIAM Review, 25*(2), 201-237. ttps://doi.org/10.1137/1025045

Kuncheva, L., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In *Systems, man and cybernetics, 2004 IEEE international conference* (Vol. 2), pp. 1214–1219. https://doi.org/10.1109/ICSMC.2004.1399790

Leary, M. R., Kelly, K. M., Cottrell, C. A., & Schreindorfer, L. S. (2013). Construct validity of the Need to Belong scale: Mapping the nomological network. *Journal of Personality Assessment, 95*(6), 610-624. https://doi.org/10.1080/00223891.2013.819511

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*(8), 707-710.

Liepins, G. E. (1980). Rigorous, systematic approach to automatic data editing and its statistical basis (No. ORNL/TM-7126). Oak Ridge National Lab., TN (USA). https://doi.org/10.2172/5518874

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*(3), 249–264. https://doi.org/10.3102/10769986022003249

Linzer, D. A., & Lewis, J. B. (2011). polca: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1–29. https://doi.org/10.18637/jss.v042.i10

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook* (Vol. 2). Springer. https://doi.org/10.1007/b107408

Massoni, S., Olteanu, M., & Rousset, P. (2009). Career-path analysis using optimal matching and self-organizing maps. In Advances in Self Organizing Maps. proceedings of the 7th International Workshop, WSOM 2009, St. Augustine, FL, USA, June 8-10, 2009 (Vol. 5629, p. 154-162). Springer. https://doi.org/10.1007/978-3-642-02397-2_18

McLachlan, G. & Peel, D. (2000) *Finite Mixture Models.* Wiley Series in Probability and Statistics. New York, Wiley. ISSN 1757-9597191. https://doi.org/10.1002/0471721182

Mikolai, J. & Lyons-Amos, M. (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. https://doi.org/10.14301/llcs.v8i2.415

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179. https://doi.org/10.1007/BF02294245

Mills, M. (2011). *Introducing survival and event history analysis*. London: Sage. https://doi.org/10.4135/9781446268360

Moen, P., Kelly, E. & Huang. R. (2008). Fit inside the work-family black box: An ecology of the life course, cycles of control reframing. *Journal of Occupational and Organizational Psychology, 81*(3), 411-433. https://doi.org/10.1348/096317908X315495

Pailhé, A., Robette, N., & Solaz, A. (2013). Work and family over the life course. A typology of French long-lasting couples using optimal matching. *Longitudinal and Life Course Studies, 4*(3), 196-217.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

Rennie, J. D., Shih, L., Teevan, J. & Karger, D.R. (2003) Tackling the poor assumptions of naïve Bayes text classifiers. In Fawcett, T. & Mishra (Eds.) *Proceedings of the Twentieth International Conference on Machine Learning 2003* (ICML-2003), pp 616-623.

Ristad, E. S., & Yianilos, P. N. (1998). Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(5), 522-532. https://doi.org/10.1109/34.682181

Ritschard, G., Gabadinho, A., Studer, M., & Müller, N. S. (2009). Converting between various sequence representations. In *Advances in Data Management* (pp. 155–175). Springer. https://doi.org/10.1007/978-3-642-02190-9_8

Robette, N. (2010). The diversity of pathways to adulthood in France: Evidence from a holistic approach. *Advances in Life Course Research, 15*(2–3), 89-96. https://doi.org/10.1016/j.alcr.2010.04.002

Robette, N., & Bry, X. (2012). Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodology Sociologique, 116*(1), 5-24. https://doi.org/10.1177/0759106312454635

Rossi, P. H., Wright, J. D., & Anderson, A. B. (2013). *Handbook of survey research*. Academic Press.

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European Sociological Review, 17*(2), 119–144. https://doi.org/10.1093/esr/17.2.119

Scholtus, S. (2014). Error localization using general edit operations (Discussion Paper No. 2014-14). The Hague: Statistics Netherlands. (Available at http://www.cbs.nl)

Studer, M. (2012). Analyse de données séquentielles et application à l'étude des inégalités sociales en début de carrière académique (Doctoral dissertation, PhD Thesis, Genève: Université de Genève).

Studer, M. (2013). Weighted cluster library manual. LIVES Working Papers , 24 , 1-32.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179*(2), 481-511. https://doi.org/10.1111/rssa.12125

Vermunt, J. K. & Magidson, J. (2003) Latent class models for classification. *Computational Statistics & Data Analysis, 41*(3-4), 531-537. https://doi.org/10.1016/S0167-9473(02)00179-2

Vermunt, J. K., & Magidson, J. (2005). Latent GOLD 4.0 User's Guide. Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2013). Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. Belmont, MA: Statistical Innovations Inc.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Watson, M. W., & Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics, 23*(3), 385-400. https://doi.org/10.1016/0304-4076(83)90066-0

Wu, C. J. (1983). On the convergence properties of the EM-algorithm. *Annals of Statistics, 11*, 95-103. https://doi.org/10.1214/aos/1176346060

Zhang, H. (2005) Exploring conditions for the optimality of naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence, 19*(2), 183-198. https://doi.org/10.1142/S0218001405003983

Endnotes

[i] In the present context, picking the "right" clustering technique is an unsolved problem since there is no generally accepted idea about the "structure" of life-course sequence data. Without such an idea, the choice of a clustering method is more or less arbitrary. For example, we know that some clustering techniques are well suited for sets of variables that have non-elliptical multivariate normal distributions with equally sized subpopulations (most hierarchical methods) or work well with particular distance measures (PAM). However, in the context of SA, we do not directly observe these underlying variables but only a diffuse summary measure: a distance between sequences. Bayne, Beauchamp, Begovich and Kane (1980), using bivariate distributions, tested thirteen different techniques for their classification accuracy. Their last, concluding sentence is "*However, as the complexity of the distributions increases, the differences between all of these methods decrease*". Unfortunately, this sentence still well summarises the state of affairs in unsupervised partitioning of distance matrices. Therefore, it is not surprising that often, in the present context, (agglomerative) hierarchical clustering is chosen: most people enter a phase of family formation during their early adulthood and, most often, this involves partnering and reproduction. It is not unreasonable to consider variation in this general pattern has a hierarchical structure and thus a hierarchical clustering seems warranted. Moreover, hierarchical techniques have the advantage of easily visualisable results in the form of a dendrogram. Although PAM is a good alternative, we decided to use Ward's agglomerative hierarchical method since it is by far the most frequently chosen. Of course, this method has disadvantages, which have been amply documented in the vast literature on clustering methods. Unfortunately, good alternatives like PAM also have disadvantages. Therefore, whichever method is picked, substantive validation of a cluster solution is of vital importance. Comparing different clustering methods in the context of SA is, even if possible at all, beyond the scope of this paper.

[ii] Latent class analysis is the simplest form of all latent structure models. There are other models that consider time-dependence of the sequences, such as Markov models. However, LCA is sequence oriented with the fewest assumptions (local independence). Therefore, it is the only latent structure model that can be sensibly compared to SA.

[iii] Country weights were not available for all countries. Given the illustrative purpose of our example, we decided to use unweighted data.