Classifying life course trajectories: a comparison of latent class and sequence analysis

Nicola Barban
DONDENA "Carlo F. Dondena" Centre for Research on Social Dynamics,
Università Bocconi, Milan, Italy
nicola.barban@unibocconi.it

Francesco C. Billari
DONDENA "Carlo F. Dondena" Centre for Research on Social Dynamics,
Department of Decision Sciences and IGIER
Università Bocconi, Milan, Italy

Draft version.

Abstract

In this article we compare two techniques that are widely used in the analysis of life course trajectories, i.e. latent class analysis (LCA) and sequence analysis (SA). In particular, we focus on the use of these techniques as devices to obtain classes of individual life course trajectories. We first compare the consistency of the classification obtained via the two techniques using an actual dataset on the life course trajectories of young adults. Then, we adopt a simulation approach to measure the ability of these two methods to correctly classify groups of life course trajectories when specific forms of "random" variability are introduced within pre-specified classes in an artificial datasets. In order to do so, we introduce simulation operators that have a life course and/or observational meaning. Our results contribute on the one hand to outline the usefulness and robustness of findings based on the classification of life course trajectories through LCA and SA, on the other hand to illuminate on the potential pitfalls of actual applications of these techniques.

1 Introduction

In recent years, there has been a significantly growing interest in the holistic study of life course trajectories, i.e. in considering whole trajectories as a unit of analysis, both in a social science setting and in epidemiological and medical studies. A particular focus of such research has been the classification of individuals according to life course trajectories, so to develop typical classes, or groups, of trajectories. This paper contributes to this line of research by assessing the robustness and consistency of the findings obtained using two of the most widespread approaches to such problem, latent class analysis (LCA from now onwards) and sequence analysis (SA from now onwards).

The two techniques, LCA and SA, come from different statistical background. Sequence Analysis, in its various specifications, is based on algorithmic, or data mining, approaches aimed at making use of measures of dissimilarity, or distance, between individual trajectories (see, e.g., Abbott, 1995; Abbott and Tsay, 2000; Billari and Piccarreta, 2005; Elzinga, 2006; Brzinsky-Fay and Kohler, 2010). The SA approach is fully nonparametric, and the standard output of the first step of SA analyses is a matrix of dissimilarities. In the second step, SA-based dissimilarity matrices are then used as inputs in data reduction techniques, mainly cluster analysis or multidimensional scaling. Groups obtained via data reduction can be used, in a third step, in subsequent analyses, e.g. on the determinants or consequences of life course trajectories. Latent Class Analysis, in its various specifications, is based on a probabilistic modeling approach, with a finite mixture distribution as the data generating mechanism (see, e.g., Hagenaars and McCutcheon, 2002; Lin et al., 2002; Reboussin et al., 2002; Beath and Heller, 2009; Bruckers et al., 2010; Pickles and Croudace, 2010). The underlying hypothesis in LCA models is that individuals belong to a finite number of classes (i.e., the values of a categorical variable) that cannot be observed. The estimating procedure aims at estimating the probability of class membership for each trajectory based on observed data via, usually, a likelihood function. LCA can also be embedded in more complex structural models, where the determinants and consequences of trajectories are included in the model, or life course trajectories are seen in parallel with other processes. Estimates are commonly obtained through an EM algorithm. In terms of classification, LCA also provides the contribution of every observed variable on the definition of classes.

In the social sciences, the analysis of life course trajectories has been applied to elicit typical pathways in the transition to adulthood, professional careers, family and fertility, criminal careers. Using either LCA or SA techniques, individuals are assigned to homogeneous classes that are interpreted as representing typical behaviors (Aassve et al., 2007; McVicar and Anyadike-Danes, 2002; Blair-Loy, 1999; Macmillan and Eliason, 2003; Amato et al., 2008; Nagin and Tremblay, 2005; Roeder et al., 1999; Tremblay et al., 2004; Groff et al., 2010). The resulting distribution in groups can be used to test a specific theory or to compare cohorts, subpopulations or the same population across time and/or space (Billari, 2001; Widmer and Ritschard, 2009). Furthermore, class membership can be used as an explanatory variable for further analyses (McVicar and Anyadike-Danes, 2002; Mouw, 2005; Billari and Piccarreta, 2005; Amato et al., 2008). Sequence analysis has also been used in geographical and mobility studies focusing on transitions that occur not only in time, but also in space. The resulting trajectories represent a set of transitions that individuals experience across time in different locations in space. For example, a SA approach has been used to describe the trajectories of tourists choice behavior (Bargeman et al., 2002; Shoval and Isaacson, 2007), or to classify individuals based on their mobility and daily-activity patterns (see e.g. Wilson, 2001; Schlich and Axhausen, 2003; Stovel and Bolan, 2004; Wilson, 2008; Saneinejad and Roorda, 2009; Vanhulsel et al., 2010)

In biostatistics and epidemiology, most applications make use of LCA or related models. LCA models are used to identify typical patterns in the evolution of health status during life course and to analyze their determinants (see e.g. Hayford, 2009; Dunn et al., 2006; Harrison et al., 2009; Bruckers et al., 2010; Croudace et al., 2003). Other studies focus on the link between health or behavioral trajectories and later outcomes during life course (Hamil-Luker and O'Rand, 2007; Lajunen et al., 2009; Berge et al., 2010; Savage and Birch, 2010; Haviland et al., 2007). Despite SA techniques were first used in genetics and biostatistics to compare DNA sequences, there are no applications of SA in the study of the evolution of health trajectories during the life course. This is partially motivated by the fact that these studies generally focus on the evolution across time of continuous variables, while SA techniques are generally used to describe trajectories of discrete states. Nevertheless, a large array of medical applications can be described as a sequence of discrete states. For example, the evolution of BMI across life course can be described in categories (e.g. underweight, normal weight, overweight or obese status) using suitable thresholds. Also, SA methods may be used to describe the occurrence and persistence of particular health status such as hypertension, depression or physical limitations.

For what follows, this paper will particularly focus on the event-based interpretation of holistic approaches to the analysis of life courses. Within this interpretation, the aim of holistic methods is to study simultaneously the *timing* of events in the life course (when do events happen?, e.g. when do individuals experience their first sexual intercourse or smoke their first cigarette), their *sequencing* (in which order do events happen?, e.g. do individuals have a child prior to marriage or stop smoking before the birth of a child), and their *quantum* (how many events happen?, e.g. how many births do they have) (Billari, 2005).

In the remainder of this paper, we compare the performance LCA and SA and test their consistency. In particular, we focus on the use of LCA and SA as devices to obtain classes of individual life course trajectories. After a brief introduction and review of the relevant literature, we compare the consistency of the classification obtained via the two techniques using an actual dataset on the life course trajectories of young adults. Then, a simulation approach is adopted to measure the ability of these two methods to correctly classify groups of life course trajectories when specific forms of “random” variability are introduced within pre-specified classes in an artificial datasets. In order to do so, we introduce simulation operators that have a life course and/or observational meaning. The results obtained contribute on the one hand to outline the usefulness and robustness of findings based on the classification of life course trajectories through LCA and SA, on the other hand to illuminate on the potential pitfalls of actual applications of these techniques.

2 Life course trajectories as categorical time series

Life course trajectories can be described as the observation, over the course of an individual’s time (i.e. age), of a number of events (i.e. life events) triggering a change in a corresponding number of categorical states. The approach used in the analyses can however, without loss of generality, be extended to states that are measurable on a quantitative scale (e.g. systolic blood pressure level, income) over discrete time units. It can also be used to represent the life course of units other than individuals (e.g., households, organizations, institutions, . . .).

The concept of trajectory derives from the interdisciplinary systematization of the life course paradigm proposed by Elder (1985), in which life course trajectories usually refer to the joint occurrence of events in multiple life domains. For example, one may want to have a representation of the evolution of union status, childbearing and work history.

Trajectories can be analyzed by representing the original data, i.e. each individual's life course, as a sequence of states. Each individual i can be associated to a variable s_{it} indicating her/his life course status at time t. As one can assume that s_{it} takes a finite number of values, trajectories can be described as categorical time series. In other terms, trajectories can be represented as strings or sequences of characters, with each character denoting one particular state. The state-space, (i.e the alphabet from which sequences are constructed) has a finite number of elements and represent all the possible states that an individual can take in each time period. For instance, a woman who is single for 12 months since the start of our observation (e.g., age 18), then starts a cohabitation lasting 5 months and then marries and remains married for 7 months can be described as follows:

$$SSSSSSSSSSSSCCCCCMMMMMMM$$

In this case, the state-space has 3 values (S=single; M=married; C=cohabiting).

More formally, let us define a discrete-time stochastic process $S_t : t \in T$ with state-space $\Sigma = \{\sigma_1, \ldots \sigma_K\}$ with realizations s_{it} with $i = 1 \ldots n$. The life course trajectory of the individual i is described by the sequence $s_i = \{s_{i1} \ldots s_{iT}\}$.

For practical reasons, a more compact representation of sequences, which we shall use later on, involves counting the repetitions of a state, which in the former example becomes as follows:

(S,12)-(C,5)-(M,7)

Life course sequences $\{s_{i1} \ldots s_{iT}\}$ can be alternatively represented by a series of vectors $\{\mathbf{y_{it}}, \ldots \mathbf{y_{iT}}\}$ where the K categories of s_{it} are represented by $M = K-1$ binary variables. This representation is particularly useful in the latent class framework, where the series of binary observations are included in the model through a logistic link.

We now briefly review the use of Latent Class Analysis and Sequence Analysis in the study of life course trajectories.

3 Latent Class Analysis of life course trajectories

Latent Class Analysis (LCA) is a statistical technique used (also) to classify individuals based on a set of categorical outcomes (Lazarsfeld and Henry, 1968; Goodman, 1974; McCutcheon, 1987; Clogg, 1995; Hagenaars and McCutcheon, 2002). The underlying assumption of LCA is that individuals belong to classes that are unobserved (latent), but for which observed data provide adequate information on class membership through a likelihood function. When data are collected longitudinally, the use of LCA is usually defined "latent trajectory modeling" or "longitudinal latent class analysis" (Vermunt, 2008b; Beath and Heller, 2009; Collins and Wugalter, 1992).

In the LCA framework, it is convenient to represent the life course trajectory as a series of binary vectors indicating the simultaneous occurrence of states in different life domains. Let us assume that there are i subject, $j = 1, \ldots, M$ life domains, $c = 1, \ldots, C$ classes and $t = 1, \ldots, T$ periods. The conditional likelihood for each subject is:

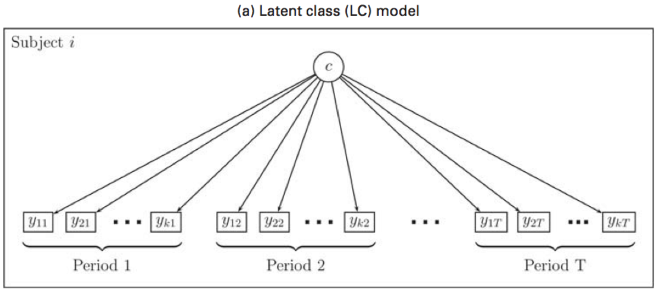
$$P(y_{i11}.,\dots,y_{iMT}|c_i=c)=\prod_{t=1}^{T}\prod_{j=1}^{M}\pi_{cjt}^{yijt}(1-\pi_{cjt})^{1-y_{ijt}},$$

where π_{cjt} is the probability of jth outcome =1 at time t for class c, constrained to be between zero and one by transformation through, for example, the logistic scale. Summing over the classes, weighted by η_c, one obtains the marginal likelihood:

$$P(y_{i11},\dots,y_{iMT})=\sum_{c=1}^{C}\eta_c P(y_{i11},\dots,y_{iMT}|c_i=c)$$

LCA assumes that the structure of correlation between observed variables is completely explained by latent factors. This condition is called “conditional indipendence”, that is $P(y_{i11}.,\dots,y_{iMT}|c_i=c) \perp\!\!\!\perp P(y_{i11}.,\dots,y_{iMT}|c_i=d)$ with $d \neq c$ (Espeland and Handelman, 1989; Hagenaars, 1988; Uebersax, 1999). The longitudinal structure of the model can be represented by figure 1.

Figure 1: Latent class structure for longitudinal data, (Beath and Heller, 2009)



The principal drawback of using standard LCA for longitudinal data is that these models do not take in consideration the time correlation between variables. The same variable measured in different time periods is, in fact, considered independent. In the recent years, various forms of correction have been proposed to adjust for temporal correlation between observations, mainly including a random effect in the model (Vermunt, 2008a; Beath and Heller, 2009; Hadgu and Qu, 1998; Vermunt, 2003). In later analyses, we refer to the more standard version of LCA applied to longitudinal data.

4 Sequence analysis and Optimal Matching

Sequence analysis is a family of algorithm based techniques used to quantify distances between categorical time series. Optimal Matching algorithm (OM) is the most known technique that has been applied to social science. The development of OM started in the seventies and the technique has been described in details by Kruskal (1983). Basically, OM expresses distances between sequences in terms of the minimal amount of effort, measured in terms of edit operations, that is required to change two sequences such that they become identical. A set that is composed of three basic operations to transform sequences is used: $\Omega = \{i, \delta, \sigma\}$, where i denotes *insertion* (one state is inserted into the sequence), δ denotes *deletion* (one state is deleted from the sequence) and σ denotes *substitution* (one state is replaced by another state). To each of these elementary operations $\omega_k \in \Omega$, a specific cost can be assigned, $c(\omega_k)$. If K basic operations must be performed to transform one sequence into another the transformation cost can be computed as $c(\omega_1, \ldots \omega_K) = \sum_{k=1}^{K} c(\omega_k)$.

A specific cost can be assigned to each operation, and the total cost of applying a series of edit operations can be computed as the sum of the costs of single operations. The distance between two sequences can thus be defined as the minimum cost of transforming one sequence into the other one. Hence, the resulting output is a symmetric matrix of pairwise distances that can be used for further statistical analysis, mainly multivariate analysis. Optimal Matching is a family of dissimilarity measures derived from the measure originally proposed in the field of information theory and computer science by Vladimir Levenshtein (Levenshtein, 1965). Abbott (1995) adapted OM to social science assigning to three elementary operations different costs, based on the social differences between states (Lesnard, 2006). The choice of the operations' costs determines the matching procedure and influences the results obtained. This is a major concern about the use of this technique in social sciences (Wu, 2000). A common solution for assessing the substitution costs is to use the inverse of the transition probability, in order to assign higher costs to the less common transitions (Piccarreta and Billari, 2007).

4.1 Sequence-based alternatives to Optimal Matching Algorithm

The use of OMA in the analysis of life course trajectories has often been criticized. (for a recent review see Brzinsky-Fay and Kohler, 2010; Aisenbrey and Fasang, 2010).

First, it is difficult to attribute a sociological meaning to the sequence operations (Lesnard, 2006). In biology the three edit operations used in OM are of little theoretical relevance since there is no resemblance with bio-chemical processes. However, differently from biological sequences, social sequences are time referenced. Therefore, the edit operations in social sequences imply modifications in the time scale. In particular, insertion and deletion operations warp time in order to match identically coded states but occurring at different moments in their respective sequences. On the other hand, substituting two events conserve the original time scale of events without warping time. A simple solution to avoid *indel* operations is to use the Hamming distance (Hamming, 1950). The Hamming distance measures the minimum number of substitutions required to change one string into the other.

Second, the choice of costs is a major concern on the use of OM for social sciences because their arbitrariness and the weak link to theory. Critics argue that the resulting distances are meaningless from a sociological point of view (Levine, 2000). In the case in which there is no a clear ranking between the different states, the definition of cost is necessarily arbitrary. A common practice is to set constant costs independent to the states that are substituted. This is equal to set $c(i) = c(\delta)$ and $c(\sigma) = 2c(\delta)$. Using this approach, $c(i)$ is a scaling factor, and the dissimilarity between two sequences is proportional to the (minimum) number of operations that are needed to transform one into another, with double weight given to substitution. The reason for setting $c(\sigma) = 2c(\delta)$ is that, in a constant cost framework, substitution is equivalent to a deletion followed by an insertion. Alternatively, it is possible to adopt a data-driven approach, i.e. using substitution costs that are inversely proportional to transition frequencies (Piccarreta and Billari, 2007). Consider two states, a and b. Let $N_t(a)$ and $N_t(b)$ be the number of individuals experiencing respectively a and b at time t, and $N_{t,t+1}(a,b)$ be the number of individuals experiencing a at time t and b at time $t+1$. The transition frequency from a to b is

$$p_{t,t+1}(a,b) = \frac{\sum_{t=1}^{T-1} N_{t,t+1}(a,b)}{\sum_{t=1}^{T-1} N_t(a)} \tag{1}$$

The cost of substituting a for b is $c(\sigma; a,b) = c(\sigma; b,a) = 2 - p_{t,t+1}(a,b) - p_{t,t+1}(b,a)$ if $a \neq b$. This cost specification takes into account the occurrence of the events weighting more those transitions that are less frequent. A possible critic is that transitions at different age are qualitatively different. For this reason, Lesnard (2006) proposes a modification of the Hamming distance using dynamic costs. The “Dynamic Hamming Distance” (DHD) is based on time-varying substitution costs $c_t(\sigma; a,b)$.

Third, it is not clear how to treat missing data and censoring among sequences. In fact, unequal sequence length due to censoring should not contribute to distance between sequences. A common practice is to restrict the analysis to sequences of the same length in order to avoid distortions due to comparing sequences of different length. Elzinga (2006) proposes different measures for categorical time series that are valid for sequences of different length and do not require cost specification. The basic idea is to compare the number of common subsequences of two sequences in order to asses a similarity measures. A subsequence is a sequence that can be derived from another sequence by deleting some elements without changing the order of the remaining elements. For example, ABD is a subsequence of $ABCDE$. Remarkable subsequences are the prefix and the suffix of a sequence, that are, respectively, the first (last) k elements of a sequence. Elzinga (2006) reviews in details different distance measures based on subsequences. The basic idea is that two sequences are very similar if they have in common long subsequences. In this way, the length of common subsequences can be used as an indicator of the similarity of two strings. Suitable measures based subsequences are: the longest common subsequence (LCS); the longest common prefix (LCP) and the longest common suffix (RLCP). The theoretical basis of these measures come from information science and their great advantage is that the researcher does not need to specify any operation costs.

Other solutions that have been proposed rely on OM with some modifications. For example, Hollister (2009) and Gauthier et al. (2009) analyze different cost specification, while Halpin (2010) proposes a modified version of the algorithm where OMs elementary operations are weighted inversely with episode length.

5 The consistency of LCA and SA: an example using real life course data

One of the main challenges of studying the life course is the complexity of life course data (Giele and Elder, 1998). It is a common practice in life course analysis to identify sensible periods of the life course using a set of different markers coming from different life domains. For instance, transition to adulthood can be described with five life course transitions: finishing school, beginning full-time employment, entering a non-marital cohabitation, becoming a parent, and getting married. The fact that these transitions can occur in different orders and at different ages yields to an enormous number of possible combinations. To study the diverse experiences of transition to adulthood, it is necessary to reduce the number of pathways to a manageable number. Amato et al. (2008) propose to use latent class analysis to create family formation pathways for women between the age of 18 and 23. Input variables include cohabitation, marriage, parenthood, full-time employment, and school attainment. Data ($n = 2,290$) come from Waves I and III of the National Longitudinal Study of Adolescent Health (Add Health). The analysis revealed seven latent pathways: college- no family formation (29%), high school–no family formation (19%), cohabitation without children (15%), married mothers (14%), single mothers (10%), cohabiting mothers (8%), and inactive (6%). Figure 2 shows the estimates of a latent class model.

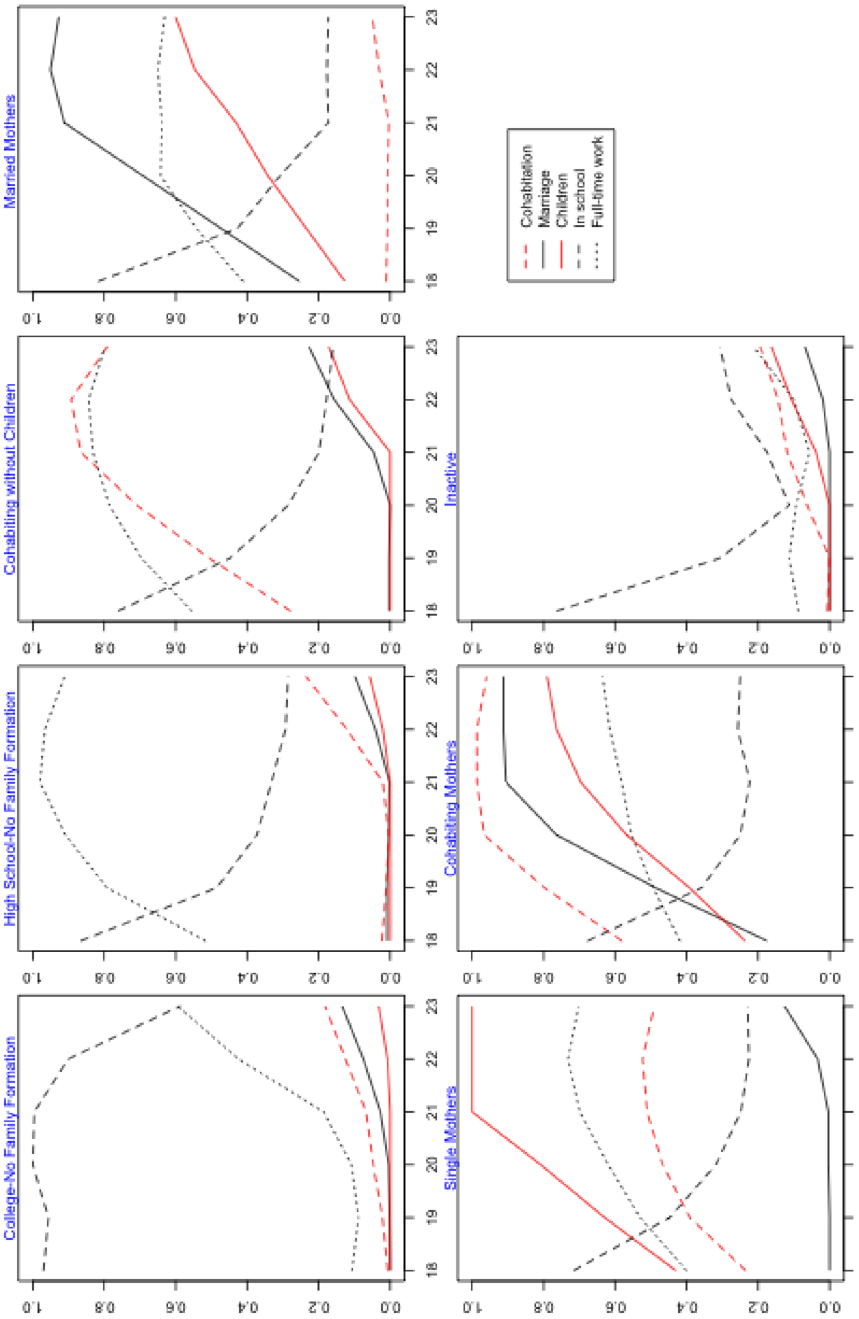
Would a sequence analysis lead to the same results? The first possible test is to run a sequence analysis with the same data and compare the groups obtained by the two methods. Family formation trajectories can be described by the joint occurrence of the five variables described above. The resulting sequence is 6 period long and the state-space is composed $2^5 = 32$ elements resulting from the combination of the possible states. It follows that the number of possible sequences is 32^6. To compare the LCA solution with sequence analysis, we calculated the dissimilarity matrix using different distances: OM with transition costs; Longest Common Subsequence (LCS); OM with constant costs; Dynamic Hamming Distance (DHD); Longest Common Prefix (LCP); Longest Common Suffix (RLCP); Hamming distance. Starting from each of these dissimilarity matrices, a cluster analysis is conducted using the Ward algorithm. Then we derive a measure of agreement in classification between the LCA solution and the cluster solutions derived by the SA approaches. The agreement in classification is measured with the Rand index (Rand, 1971) that measures the proportion of couples of observations classified in the same group by two cluster solutions. The corrected version of Rand Index (Morey and Agresti, 1984) accounts for the agreement due to chance. Results are presented in table 1. A detailed description of the clustering method and the classification index is presented in section 6.3.

Table 1: Agreement in classification between LCA and SA techniques

	Rand index	Corrected Rand index
OM with empirical costs	0.88	0.59
Longest common subsequence (LCS)	0.87	0.55
OM with constant costs	0.86	0.52
Dynamic Hamming distance (DHD)	0.86	0.50
Longest Common Prefix (LCP)	0.77	0.26
Longest Common Suffix (RLCP)	0.71	0.19
Hamming distance	0.71	0.19

In this example, Optimal Matching with empirical-derived costs gives the closest solution to the classes identified by LCA. The rand index is 0.88 meaning that among all the possible pairs of observations, almost the 90% are classified in the same group using the two methods. The corrected version of the Rand index accounts for the proportion of agreement due to chance and reduces the percentage of couples classified in agreement to

Figure 2: Latent class representation of early family formation. Women 18-23 years old. Add-health, (Amato et al., 2008)



59%. The LCS distance does not imply any cost settings. The cluster solution obtained with this method is very similar to the OM solution (0.87 Rand index, 0.55 the corrected version). Using constant costs does not substantially decrease the agreement with respect to the OM version with empirical costs. Also the use of dynamic costs based on the age of the respondent does not change the percentage of agreement between the two classification. On the other hand, the cluster solutions obtained with the remaining distances (LCP; RLCP and Hamming distance) diverges substantially from the LCA solution presented in the paper by Amato et al. (2008).

This example does not motivate the use of a particular distance respect to the others, but gives a first indication on the consistence of different statistical methods for life course analysis. In particular it is interesting to notice that, in this case, the two methods that lead to a closer solution to LCA are OM with transition costs and LCS. In the simulations presented in this paper, we compare LCA with these two methods for sequence analysis. Although the different approaches for life course classification seem to be consistent (in particular between LCA and OM), it is not possible to draw any conclusion on the reliability of the methods if the generating mechanism of life course sequences is unknown.

6 A simulation study

We propose a simulation approach to study the factors affecting the goodness of LCA and SA techniques. The simulation procedure can be summarized in 4 steps:

1. Define typical groups of life course trajectories
2. Introduce variability in timing, quantum and sequencing
3. Classify individuals of the artificial dataset using Latent Class and Optimal Matching techniques
4. Compare classification obtained with the two techniques with the real groups

A simulation approach to test the reliability of SA techniques has been previously proposed by Wilson (2006) to test the performances of the *ClustalG* multiple alignment package. The simulation study proposed in this paper, however, follows a different approach. Instead of starting from a stochastic generating mechanism, the reliability of SA techniques is tested increasing the level of heterogeneity among groups of sequences.

6.1 Defining typical groups of life course trajectories

Let us define 4 different groups of life course trajectories using a simple state-space composed by the states S,C,M. For each sequence, we set the length equal to 30 and S as initial state. Then, we repeat every typical sequence 250 times obtaining an artificial dataset of 1000 observations. The dataset can be considered as a monthly (quarterly) collection of data indicating the marital/union status of an individual. One, for example, can consider S as single, C as cohabiting and M as married. Let us define 4 "typical" groups of sequences:

1. (S,10)-(C,10)-(M,10)

2. (S,20)-(C,5)-(M,5)
3. (S,10)-(C,5)-(M,10)-(C,5)
4. (S,20)-(M,10)

Where (X, t) indicates t periods in state X. Individuals from group 1 are single for 10 periods than they cohabit for 10 periods and then they stay in marriage until the end of the sequence. Groups differ for timing, quantum and order. For example, group 1 differs from group 2 because individuals exit state S earlier, from group 3 because of the order of states M and C and from group 4 because they experiences state C.

6.2 Introducing variability in the typical sequences

To test the reliability of the two methods, we introduce random perturbations in timing, quantum and sequencing of trajectories. The idea is to confound the latent groups modifying sequences with different sources of noise. Thus, we introduce a series of sequence operators that modify life trajectories. These operations introduce variability in the groups. Even if these operations have not a specific meaning in the social sciences, we tried to mimic some behaviors observed by individuals during the life course.

Let us define the following operators:

- Postponement

  With probability p (postponement rate), copy status from time t to time $t + 1$

  S S S S S S S S S S **C** C C C C C C C C C **M**M M M M M M M M M

  S S S S S S S S S S **S** C C C C C C C C C **C** M M M M M M M M M

- Slicing

  With probability p (slicing rate), exchange two subsequence of the same length

  S S S S S S S S S S **C C C C C** C C C C **C** **MMMM**M M M M M M

  S S S S S S S S S S **C** **MMMM**C C C C **C C C C C** M M M M M M

- Inversion

  With probability p (inversion rate), exchange all the elements C with elements M

  S S S S S S S S S S **C C C C C C C C C C** M M M M M M M M M M

  S S S S S S S S S S **MMMMMMMMMM**C C C C C C C C C C

- Mutation

  With probability p (mutation rate), substitute sequences status at time t with a random element of the alphabet.

  S S S S **S** S S S S S **C** C C C C C **C** C C C **M**M M M **M**M M M M M

  S S S S **M**S S S S S **S** C C C C C **C** C C C **C** M M M **C** M M M M M

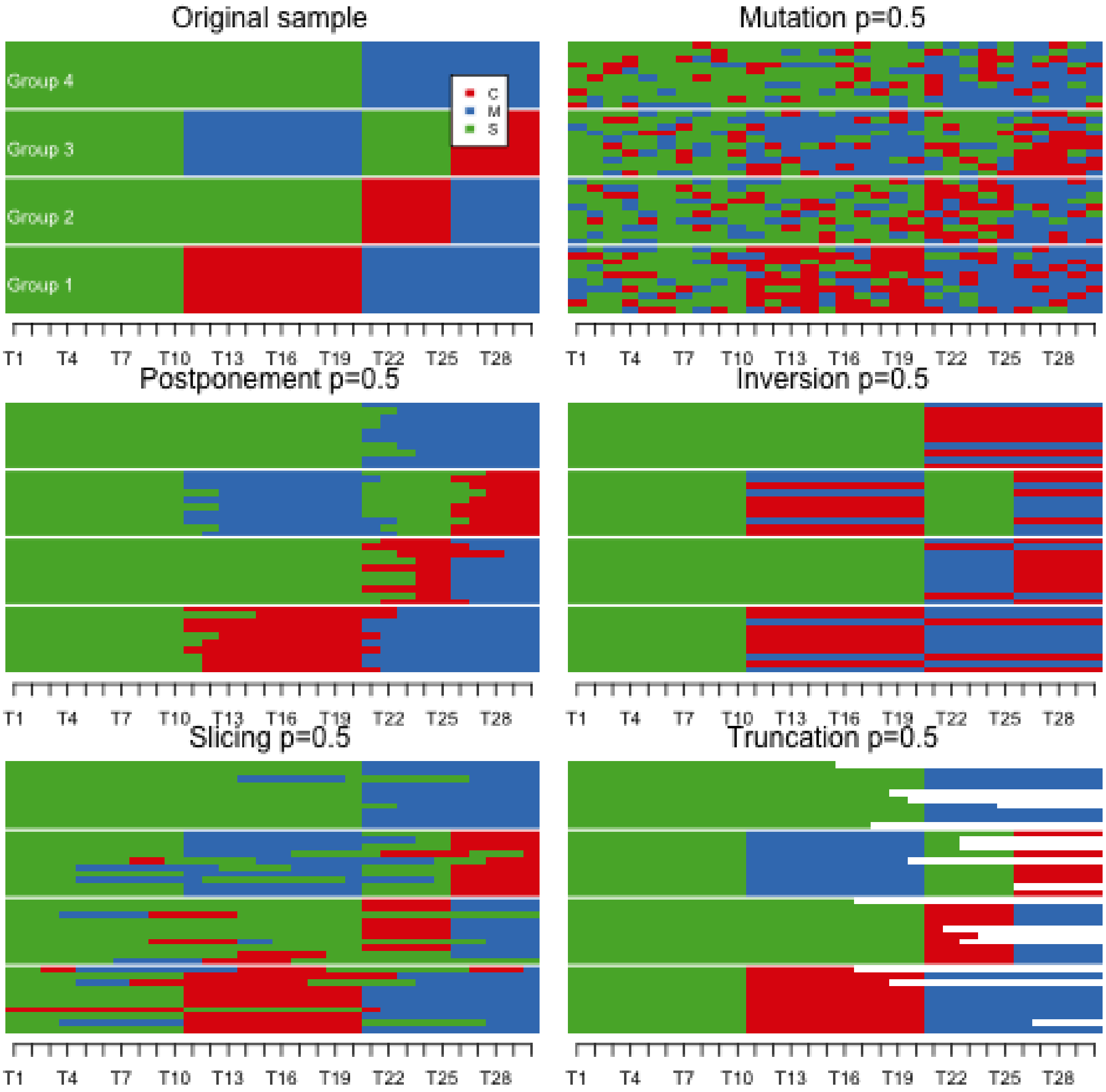
- Truncation

  With probability p cut sequence at time t, with t randomly chosen.

  S S S S S S S S S S C C C C C C C C C C M M M M **M**M M M M M

  S S S S MS S S S S S C C C C C C C C C C M M M

Figure 3: Effects of different sequence operators



The operators proposed are meant to introduce variations in the different components of life course introducing variability among sequences. The general idea is to modify the sequences mimicking the behavior of real life course trajectories. For example, some individuals may postpone (or anticipate) a transition, while others invert the "order" in which events happen. Mutation does not have a direct life course interpretation, but it can be described as a source of measurement error, since it may occur that individuals are randomly misclassified across time. Using this disturbance strategy allows to test the reliability of different methods without assuming any generating mechanism of the data.

6.2.1 How to measure variability in Timing, Quantum and Sequencing

- **Timing** The *tempo* dimension of a transition is the timing in which a change of state occurs. The exit time from the first time is a crucial transition in many demographic studies (i.e. leaving parental home, entering the first union, having the first child). As a naive indicator of timing, we define the age at first transition. The standardized indicator τ expresses the proportion of a life sequence spent in the initial status. Precocious individuals have a low value of τ, on the contrary τ increases with postponement.

$$t_{min} = min\{s_{(t-1)} \neq s_t\} \qquad t = 1, \ldots, T$$

$$\tau = t_{min}/T$$

- **Quantum**. The number of events is a key element that characterizes a life course trajectory. The concept of *Quantum* indicates the likelihood of an individual to experience transitions. A simple indicator can be expressed by the overall number of transitions. The standardized value ρ indicates the number of transitions per time period.
$$\rho = \frac{\#\{s_{(t-1)} \neq s_t\}}{T}$$
- **Sequencing** The order in which events occur is crucial in the study of life course. For example, it may be relevant to study the divergence of a life trajectory from the normative course of transition. For this reason, we propose as an indicator, the number of non-normative transition. That is, the transitions that diverge from a given sequence of events considered normative in the society. The standardized value ς indicates the proportion of normative transitions over the total number of transition.
$$\varsigma = \frac{\text{Number of normative transitions}}{\text{Total number of transitions}}$$

The three indicators range between 0 and 1.

These operators modify different dimension of life courses. Postponement introduces a major change in timing while the other two dimensions remain unaltered. Inversion modifies only the order of events because it transforms an entire category of events into another. Slicing modifies both the order and the quantum of events. Last, mutation has a massive effect on quantum, but it affects also the other two dimensions introducing completely random variations. The effect of the sequence operators are illustrated in figure 4.

6.3 Classification

Once defined a new dataset, modified by the previous "sequence operators", it is possible to apply the alternative classification procedures. While LCA requires less specifications by the researcher, in sequence analysis one need to specify the costs (only in case of OM) and the clustering procedure. Following the most common approach in SA for demographic studies, we estimate OM distances using costs proportional to transition rates and we use standard Ward algorithm for clustering. Ward clustering algorithm (Ward, 1963) can be briefly described as it follows. Consider N individuals to be clustered according to their sequences. Let $d(i, j)$ denote the distance between the ith and the jth

Figure 4: Effects in timing, quantum and sequencing. Mutation, Postponement

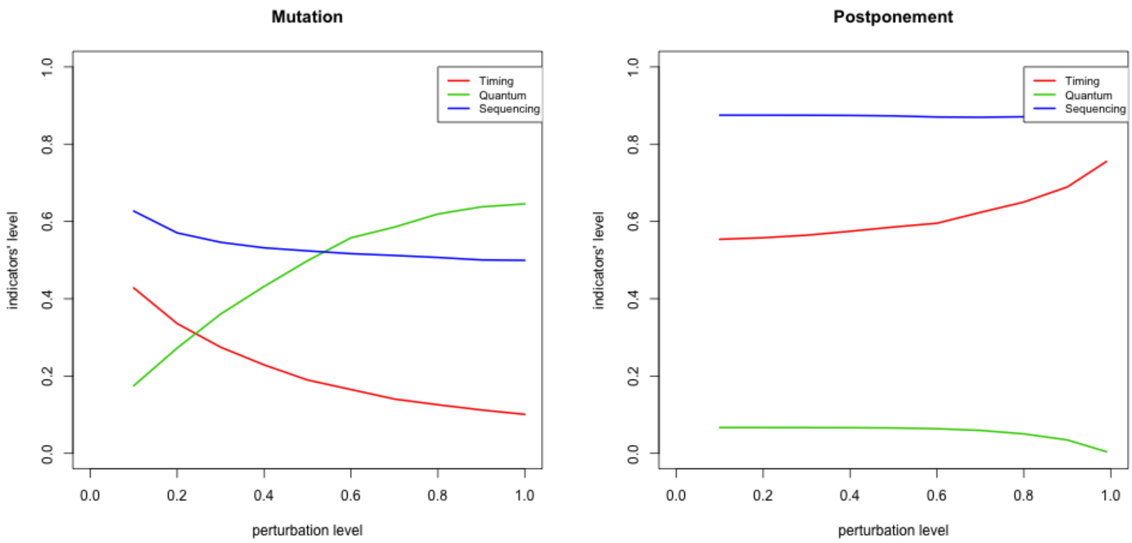
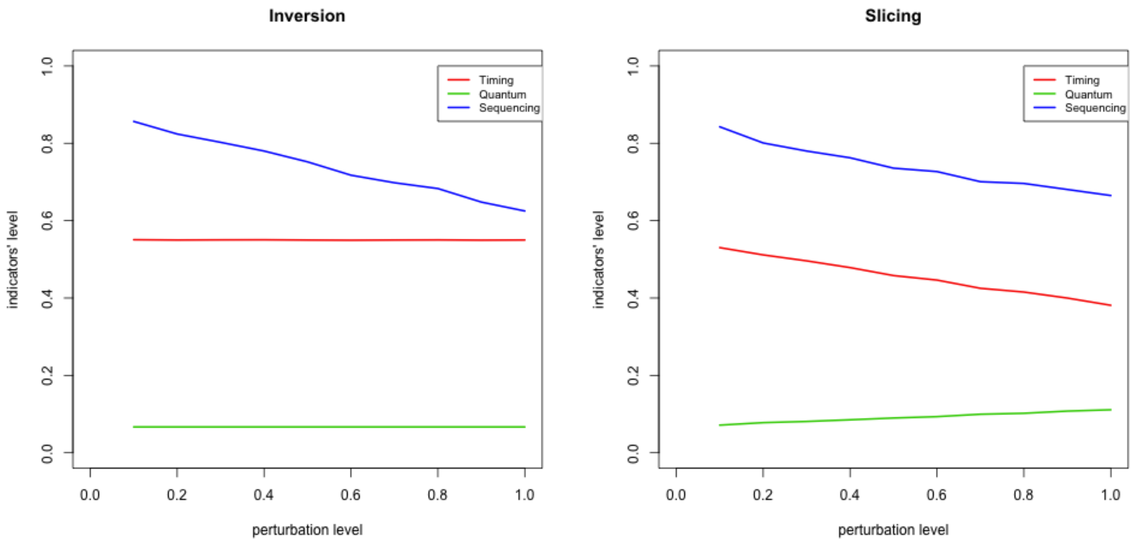


Figure 5: Effects in timing, quantum and sequencing. Inversion, Slicing



individual sequences. The total dispersion, i.e. the amount of dispersion within the whole data set, is usually measured as $T = \sum_{i,j} d(i,j)$. Suppose now that the whole sample is partitioned into G clusters. The dispersion within the gth cluster is $W_g = \sum_{i,j \in g} d(i,j)$, and the dispersion within the G groups can be summarized as $W_G = \sum_{g=1}^{G} W_g$. The adequacy of a clustering solution is often evaluated by referring to $R_G^2 = 1 - W_G/T$, which is the proportion of the total dispersion accounted for by the G clusters. By construction, if $G-1$ clusters are obtained by joining two clusters, say g_L and g_R, out of a number of G, into a single one g, it follows that $W_G < W_{G-1}$, and $R_G^2 > R_{G-1}^2$. Hierarchical agglomerative clustering algorithms proceed by sequentially joining pairs of clusters: they differ in the criterion that is followed to select which clusters must be joined. In Wards algorithm the two clusters to be joined are selected by minimizing the increase in the within-groups dispersion consequent on the reduction of the partitions degree:

$$\Delta(g|g_L, g_R) = W_g - W_{gL} - W_{gR} = W_{G-1} - W_G \tag{2}$$

or, equivalently, by minimizing $\Delta R_{G-1}^2 = R_G^2 - R_{G-1}^2$. The result of this hierarchical procedure is a sequence of (nested) clusters solutions having a decreasing number of clusters, $\{P_{max}, P_{max-1}, ..., P_1\}$, max being the maximum number of clusters that we can define, coinciding with N, the total number of cases. Given a partition P_G, the P_{G-1} partition is determined by (conditionally) maximizing R_{G-1}^2, i.e. by minimizing the decrease in the R^2 due to the reduction of the number of clusters.

Latent Class has been conducted setting binary variables in each time period indicating if the individual is single (S), cohabiting (C) and married (M). To avoid local maxima we run the model 3 times and we choose the model with the minimum BIC. For practical purposes, both the number of classes and the number of clusters is set fixed. The analyses conducted varying the number of classes give similar results in terms of classification performances.

6.4 Classification performances

The goodness of classification is measured examining the association rate between the classes obtained by the two methods and the original groups. we measure how the association between the real and the actual groups changes according to different levels of disturbance. Association rate is measured with a modified version of Rand index (Rand, 1971). Rand index measures the proportion of couples of observations that are classified in the same group by two (or more) judges. Suppose that in the population of interest, there are k_1 clusters in the first solution and k_2 clusters in the second. Let P_{ij} be the probability that a randomly selected individual is classified in cluster i in the first solution and cluster j in the second solution. Rand's statistic is defined to be the probability that a randomly selected pair is classified in agreement. This probability equals

$$\begin{aligned} P_s &= \sum\sum P_{ij}^2 + \sum\sum P_{ij}(1 - P_{i+} - P_{+j} + P_{ij}) \qquad (3) \\ &= 1 - \sum P_{i+}^2 - \sum P_{+j}^2 + 2\sum\sum P_{ij}^2 \qquad (4) \end{aligned}$$

This measure of agreement has the advantage that can be used even if the size of the two clusters (k_1 and k_2) differ. On the other hand, Rand index makes no correction for chance agreement. Therefore, it is not possible to tell whether a specific value of P_s is "large" or "small", because its value when individuals are classified at random (i.e. $P_{ij} = P_{i+}P_{+j}$) is not zero, and depends on P_{i+} and P_{j+}. This can constitute a disadvantage when the replicability of different classifications are being compared. In this paper, we use the corrected version of the Rand Index (Morey and Agresti, 1984) that properly takes into account the proportion of agreement due to chance. The corrected version of Rand's statistic equals

$$\Omega = \frac{2\sum\sum P_{ij}^2 - 2(\sum P_{i+}^2)(\sum P_{+j}^2)}{\sum P_{i+}^2 + \sum P_{+j}^2 - 2(\sum P_{i+}^2)(\sum P_{+j}^2)}. \tag{5}$$

This statistic equals one for perfect agreement, $\Omega = 0$ for chance agreement, and $\Omega < 0$ when agreement is less than expected by chance.

7 Simulation results

We simulated 1000 samples for each sequence operator applying different level of disturbance. For each sample, we estimate a latent class model with 4 classes and we calculated OM and LCS matrix of dissimilarity. Then we apply a cluster analysis using Ward algorithm to classify individuals in 4 groups. The groups obtained are compared with the original groups using the corrected Rand index. Figure 6 and table 3 report the average rate of agreement between the original groups and the results obtained by latent class analysis and sequence analysis (OM and LCS). Results show that classification is sensitive to the transformations inducted by sequence operators. With the increasing of variability in the sample, the classification goodness decreases. As expected, the performances of all the methods decrease rapidly with random mutation. Mutation, in fact, can be considered a benchmark since it introduces the maximum amount of variation. The agreement rate under postponement decreases more slowly. In particular small postponement rates do not seem to affect the probability of good classification. However, precision decreases with higher disturbance levels. Postponement principally affects timing, since it extends the amount of time spent in the initial status. But a massive postponement has also an effect in quantum, since it reduces the amount of transitions in trajectories and reduces the variability between different groups of sequences. Inversion has the maximum confounding effect at rate 0.5. At that point, exactly half of sequences get all "C" inverted with "M" and vice-versa. With greater inversion rates, the order of sequences changes and, in turn, variability within groups is reduced. Therefore, classification becomes straightforward. Slicing has an effect both on sequencing and quantum of life course and the classification decreases almost linearly. The performances of classification under truncation follow a U-shape. An increase in truncation rate affects the number of censored individual sequences. It follows that high truncation rates are associated with sequences that are shorter in average. For this reason (since truncation is randomly assigned to the second half of the sequence), we observe an increase in classification agreement when the truncation rate is high.

The results obtained by our simulations suggest some considerations about the reliability of these classification methods. First, there are no evidence of a methodology that have superior performances under all the sources of variation. In fact we do not observe a methodology that perform better in all the cases. Despite that, according to our simulations, LCA has better performances under mutation and truncation. On the other hand, SA shows greater agreement in inversion and slicing. Results from postponement indicate a substantial equivalence of the techniques with slightly better results for sequence analyses. Second, the classifications with latent class analysis seem to be less precise. Using simulated data it is possible to have an indication on the variability of the estimated agreement rates. Under all the sources of error, the results obtained with LCA exhibit more variability. Third, the differences between OM and LCS are minimal. Both the methods, in fact, produce very similar results. Although the two distances are qualitatively different, the results obtained in all the sources of variability are very similar.

To summarize the results we propose a measure of the overall performance. Let R be the number of simulations, and $\Omega_r^{\{LCA;OM;LCS\}}$ the corrected Rand index for the sample r under different sequence operators. A simple index of the overall goodness of classification is the expected Rand index $\bar{\Omega}$.

$$\bar{\Omega} = \frac{1}{R}\sum_{r=1}^{R} \Omega_i \tag{6}$$

$\bar{\Omega}$ can be interpreted as the expected agreement between the true groups and the estimated classification. Table 2 summarizes the results. Sequence analysis techniques seem to have better performances under postponement, inversion and slicing. Latent class analysis gives better results under mutation and in case of data truncation.

Table 2: Classification rate

	Mutation	Postponement	Inversion	Slicing	Truncation
LCA	0.586	0.713	0.566	0.427	0.608
OM	0.520	0.735	0.638	0.632	0.549
LCS	0.509	0.737	0.647	0.646	0.552

Figure 6: Classification results

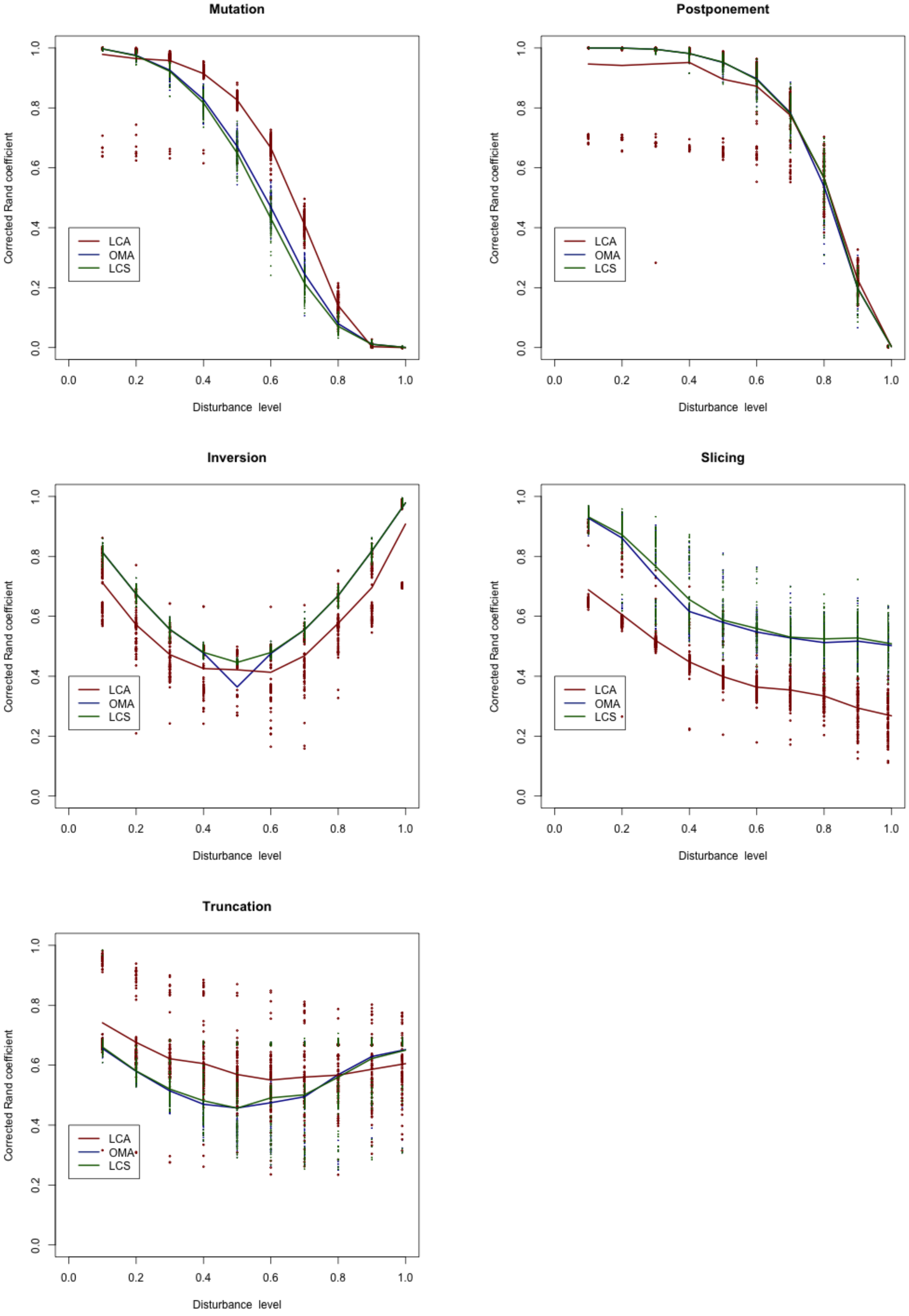


Table 3: Simulation results

		Inversion			*Postponement*			*Mutation*			*Slicing*			*Truncation*		
		LCA	OM	LCS	LCA	OM	LCS	LCA	OM	LCS	LCA	OM	LCS	LCA	OM	LCS
0.1	Mean	0.7100	0.8120	0.8120	0.9470	1.0000	1.0000	0.9790	0.9970	0.9970	0.6870	0.9270	0.9320	0.7400	0.6550	0.6590
	Var	0.0077	0.0005	0.0005	0.0131	0.0000	0.0000	0.0064	0.0000	0.0000	0.0087	0.0005	0.0005	0.0176	0.0003	0.0023
0.2	Mean	0.5700	0.6730	0.6730	0.9420	1.0000	1.0000	0.9650	0.9750	0.9730	0.6040	0.8590	0.8710	0.6750	0.5790	0.5800
	Var	0.0063	0.0004	0.0004	0.0143	0.0000	0.0000	0.0078	0.0001	0.0001	0.0062	0.0068	0.0049	0.0146	0.0009	0.0008
0.3	Mean	0.4720	0.5560	0.5560	0.9470	0.9950	0.9950	0.9580	0.9250	0.9210	0.5190	0.7310	0.7670	0.6220	0.5130	0.5190
	Var	0.0047	0.0003	0.0003	0.0163	0.0000	0.0000	0.0041	0.0005	0.0005	0.0018	0.0146	0.0134	0.0102	0.0021	0.0024
0.4	Mean	0.4250	0.4760	0.4790	0.9520	0.9810	0.9820	0.9140	0.8270	0.8150	0.4470	0.6150	0.6540	0.6040	0.4690	0.4800
	Var	0.0061	0.0007	0.0002	0.0111	0.0002	0.0002	0.0026	0.0011	0.0011	0.0027	0.0064	0.0107	0.0137	0.0053	0.0067
0.5	Mean	0.4220	0.3640	0.4460	0.8930	0.9500	0.9520	0.8250	0.6690	0.6480	0.3980	0.5790	0.5870	0.5690	0.4560	0.4550
	Var	0.0043	0.0028	0.0000	0.0187	0.0006	0.0005	0.0013	0.0021	0.0016	0.0023	0.0037	0.0038	0.0071	0.0113	0.0115
0.6	Mean	0.4130	0.4760	0.4790	0.8720	0.8960	0.8940	0.6630	0.4660	0.4290	0.3630	0.5470	0.5590	0.5500	0.4750	0.4930
	Var	0.0075	0.0006	0.0002	0.0123	0.0015	0.0016	0.0025	0.0025	0.0025	0.0018	0.0027	0.0033	0.0111	0.0142	0.0151
0.7	Mean	0.4700	0.5560	0.5560	0.7720	0.7820	0.7760	0.4080	0.2440	0.2130	0.3540	0.5270	0.5300	0.5610	0.4940	0.5000
	Var	0.0075	0.0005	0.0005	0.0085	0.0025	0.0042	0.0020	0.0016	0.0017	0.0026	0.0025	0.0026	0.0140	0.0176	0.0194
0.8	Mean	0.5790	0.6700	0.6700	0.5630	0.5350	0.5620	0.1390	0.0790	0.0700	0.3330	0.5130	0.5240	0.5670	0.5690	0.5600
	Var	0.0057	0.0006	0.0006	0.0065	0.0079	0.0066	0.0015	0.0003	0.0004	0.0024	0.0024	0.0033	0.0089	0.0142	0.0164
0.9	Mean	0.6980	0.8190	0.8190	0.2240	0.1950	0.1950	0.0030	0.0110	0.0110	0.2940	0.5170	0.5280	0.5870	0.6290	0.6230
	Var	0.0085	0.0005	0.0005	0.0020	0.0018	0.0019	0.0000	0.0000	0.0000	0.0039	0.0041	0.0040	0.0104	0.0076	0.0089
0.99	Mean	0.9080	0.9790	0.9790	0.0030	0.0030	0.0030	0.0000	0.0000	0.0000	0.2680	0.5020	0.5080	0.6050	0.6520	0.6500
	Var	0.0142	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0041	0.0039	0.0037	0.0086	0.0048	0.0051

8 Discussion

In the last decade, holistic methods for life course analysis have become more and more common. Instead of focusing only on life course transitions, the object of the study is the entire life trajectory. Life course trajectories can be described as categorical time series where time is associated to life states. Using longitudinal or retrospective data, it is, in fact, possible to describe individuals' life courses as age-referenced sequences of events. Rather than modeling directly the probability of the occurrence of a particular event, holistic methods attempt to individuate important patterns in the data using a data mining approach. In the literature of life course analysis we can distinguish two principal approaches: latent class analysis and sequence analysis.

It is not clear, however, how reliable are these methods in detecting effectively patterns in the data. A bigger critic that have been moved to these techniques is exactly their reliability and the difficulties in testing it. For this reason, we propose a simulation approach to investigate the reliability of classification techniques in life course analysis. Furthermore, we propose a method to simulate life sequencing without making any assumptions on the generating mechanism of the data. Starting from homogeneous groups of life trajectories, we introduce different sources of variability that, mimicking individuals' behavior, transform life courses in different dimensions. This approach allows to test if there are substantial differences in detecting groups of life trajectories.

Our simulation results show that the two methods are consistent. Although we do not found the absolute superiority of a method respect to the other, our results show that OM and LCS seem to have better performances when life course sequences are modified in the ordering of transitions (inversion and slicing). On the other hand, LCA has better results when the variations are completely random (mutation). Although random mutation may be common in some scientific fields, i.e. biology or information theory, a random disturbance appears to be quite unlikely in life course analysis. Individuals may experience unexpected events in life course, but usually these events are associated with a duration and rarely have no effect on the following part of the life trajectory. Nevertheless, mutation can be interpreted as a measurement error, since individuals may be misclassified during repeated measurements.

Overall, the results obtained in this paper justify the use of sequence analysis (in particular OM and LCS) for the study of life course. Our sequence operators do not cover all the possible variation that can occur in life course. That otherwise would be impossible. Also, life course classification may be influenced by other factors (i.e. the length of sequences, the dimension of the state-space and the classification algorithm). Despite that, this study presents some limitations, it represents one of the first attempt to test the reliability of holistic methods for life course analysis.

References

Aassve, A., F. C. Billari, and R. Piccarreta (2007, Jan). Strings of adulthood: A sequence analysis of young british women's work-family trajectories. *European Journal of Population 23*(3-4), 369–388.

Abbott, A. (1995). Sequence analysis: new methods for old ideas. *Annual Review of Sociology 21*(1), 93–113.

Abbott, A. and A. Tsay (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research 29*(1), 3.

Aisenbrey, S. and A. Fasang (2010, Jan). New life for old ideas: The "second wave" of sequence analysis bringing the"course"back into the life course. *Sociological Methods & Research 38*(3), 420–462.

Amato, P., N. Landale, and T. Havasevich-Brooks (2008, Jan). Precursors of young women's family formation pathways. *Journal of Marriage and Family 70*, 1271–1286.

Bargeman, B., C. Joh, and H. Timmermans (2002). Vacation behavior using a sequence alignment method. *Annals of Tourism Research 29*(2), 320–337.

Beath, K. J. and G. Z. Heller (2009, Jan). Latent trajectory modelling of multivariate binary data. *Stat Model 9*(3), 199–213.

Berge, J. M., M. Wall, K. W. Bauer, and D. Neumark-Sztainer (2010, Apr). Parenting characteristics in the home environment and adolescent overweight: a latent class analysis. *Obesity (Silver Spring) 18*(4), 818–25.

Billari, F. C. (2001, Jan). The analysis of early life courses: complex descriptions of the transition to adulthood. *Journal of Population Research 18*(2), 119–142.

Billari, F. C. (2005). Life course analysis: two (complementary) cultures? some reflections with examples from the analysis of the transition to adulthood. *Advances in Life Course Research 10*, 261–281.

Billari, F. C. and R. Piccarreta (2005, Jan). Analyzing demographic life courses through sequence analysis. *Mathematical Population Studies 12*, 81–106.

Blair-Loy, M. (1999, Mar). Career patterns of executive women in finance: An optimal matching analysis. *The American Journal of Sociology 104*(5), 1346–1397.

Bruckers, L., J. Serroyen, G. Molenberghs, H. Slaets, and W. Goeyvaerts (2010, Jan). Latent class analysis of persistent disturbing behaviour patients by using longitudinal profiles. *Journal of the Royal Statistical Society. Series C 59*, 495–512.

Brzinsky-Fay, C. and U. Kohler (2010, Jan). New developments in sequence analysis. *Sociological Methods Research 38*(3), 359–364.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, and M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences*. New York: Plenum.

Collins, L. and S. Wugalter (1992, Jan). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research 27*(1), 131–157.

Croudace, T., M. Jarvelin, M. Wadsworth, and P. Jones (2003, Jan). Developmental typology of trajectories to nighttime bladder control: Epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology 157*(9), 834–842.

Dunn, K. M., K. Jordan, and P. R. Croft (2006, Apr). Characterizing the course of low back pain: a latent class analysis. *American Journal of Epidemiology 163*(8), 754–61.

Elder, G. H. (1985, Jan). *Life course dynamics: trajectories and transitions, 1968-1980.* Ithaca, NY: Cornell Univ Press.

Elzinga, C. (2006). Sequence analysis: Metric representations of categorical time series. *Sociological Methods Research 38*(3), 463–481.

Espeland, M. and S. Handelman (1989, Jan). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics 45*(2), 587–599.

Gauthier, J., E. Widmer, P. Bucher, and C. Notredame (2009). How much does it cost?: Optimization of costs in sequence analysis of social science data. *Sociological Methods Research 38*(1), 197–231.

Giele, J. Z. and G. H. Elder (1998, Jan). *Methods of life course research: qualitative and quantitative approaches.* Los Angeles, CA: SAGE Publications.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*, 215–231.

Groff, E. R., D. Weisburd, and S.-M. Yang (2010, Mar). Is it important to examine crime trends at a local "micro" level?: A longitudinal analysis of street to street variability in crime trajectories. *Journal of Quantitative Criminology 26*(1), 7–32.

Hadgu, A. and Y. Qu (1998, Jan). A biomedical application of latent class models with random effects. *Journal of the Royal Statistical Society. Series C 47*(4), 603–616.

Hagenaars, J. (1988). Latent structure models with direct effects between indicators: local dependence models. *Sociological methods & research 16*(3), 379–405.

Hagenaars, J. A. and A. L. McCutcheon (2002, Jan). *Applied latent class analysis.* Cambridge, UK: Cambridge University Press.

Halpin, B. (2010, Feb). Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods Research 38*(3), 365–388.

Hamil-Luker, J. and A. M. O'Rand (2007, Feb). Gender differences in the link between childhood socioeconomic conditions and heart attack risk in adulthood. *Demography 44*(1), 137–158.

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal 26*(2), 147–160.

Harrison, W. J., B. M. Bewick, M. S. Gilthorpe, A. J. Hill, and R. M. West (2009, Oct). Longitudinal latent class analysis of alcohol consumption. *Journal of Epidemiology & Community Health 63*(Suppl 2), 35–35.

Haviland, A., D. Nagin, and P. Rosenbaum (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods 12*(3), 247.

Hayford, S. R. (2009, Nov). The evolution of fertility expectations over the life course. *Demography 46*(4), 765–783.

Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods Research 38*(2), 235–264.

Kruskal, J. (1983, Apr). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review 25*(2), 201–237.

Lajunen, H.-R., A. Keski-Rahkonen, L. Pulkkinen, R. J. Rose, A. Rissanen, and J. Kaprio (2009, Oct). Leisure activity patterns and their associations with overweight: a prospective study among adolescents. *Journal of Adolescent 32*(5), 1089–103.

Lazarsfeld, P. F. and N. W. Henry (1968). *Latent Structure Analysis.* Boston, MA: Houghton Mifflin.

Lesnard, L. (2006). Optimal matching and social sciences. *Manuscript, Observatoire Sociologique du Changement (Sciences Po and CNRS), Paris.(http://laurent. lesnard. free. fr/).*

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Dokl. 10*, 707–710.

Levine, J. (2000). But what have you done for us lately?: Commentary on abbott and tsay. *Sociological Methods Research 29*(1), 34–40.

Lin, H., B. Turnbull, C. McCulloch, and E. Slate (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association 97*(457), 53–66.

Macmillan, R. and S. R. Eliason (2003). Characterizing the life course as role configurations and pathways. In J. T. Mortimer and M. J. Shanahan (Eds.), *Handbook of the Life Course*, pp. 529–554. New York, NY: Springer.

McCutcheon, A. C. (1987). *Latent Class Analysis.* Beverly Hills, CA: Sage Publications.

McVicar, D. and M. Anyadike-Danes (2002, Jan). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society. Series A 165*(2), 317–334.

Morey, L. and A. Agresti (1984, Mar). The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement 44*(1), 33.

Mouw, T. (2005, Jan). Sequences of early adult transitions: A look at variability and consequences. In R. Settersten, F. Furstenberg, and R. Rumbaut (Eds.), *On the Frontier of Adulthood: Theory, Research, and Public Policy.* Chicago, IL: University of Chicago Press.

Nagin, D. S. and R. Tremblay (2005). Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology 43*(4), 873–904.

Piccarreta, R. and F. C. Billari (2007, Jan). Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society. Series A 170*(4), 1061 – 1078.

Pickles, A. and T. Croudace (2010, Jun). Latent mixture models for multivariate and longitudinal outcomes. *Statistical Methods in Medical Research 19*(3), 271–89.

Rand, W. M. (1971, May). Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association 66*(336), 846–850.

Reboussin, B., M. Miller, K. Lohman, and T. Have (2002, Jan). Latent class models for longitudinal studies of the elderly with data missing at random. *Journal of the Royal Statistical Society. Series C 51*(1), 69–90.

Roeder, K., K. Lynch, and D. S. Nagin (1999). Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association 94*(447), 766–767.

Saneinejad, S. and M. Roorda (2009, Jan). Application of sequence alignment methods in clustering and analysis of routine weekly activity schedules. *Transportation Letters 1*(3), 197–211.

Savage, J. S. and L. L. Birch (2010, Mar). Patterns of weight control strategies predict differences in women's 4-year weight gain. *Obesity (Silver Spring) 18*(3), 513–20.

Schlich, R. and K. Axhausen (2003). Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation 30*(1), 13–36.

Shoval, N. and M. Isaacson (2007). Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers 97*(2), 282–297.

Stovel, K. and M. Bolan (2004, Jan). Residential trajectories: Using optimal alignment to reveal the structure of residential mobility. *Sociological methods & research 32*, 559–598.

Tremblay, R. E., D. S. Nagin, J. R. Séguin, M. Zoccolillo, P. D. Zelazo, M. Boivin, D. Pérusse, and C. Japel (2004, Jul). Physical aggression during early childhood: trajectories and predictors. *Pediatrics 114*(1), e43–50.

Uebersax, J. (1999, Jan). Probit latent class analysis: Conditional independence and conditional dependence models. *Applied Psychological Measurement 23*(4), 283–297.

Vanhulsel, M., C. Beckx, D. Janssens, K. Vanhoof, and G. Wets (2010). Measuring dissimilarity of geographically dispersed space–time paths. *Transportation forthcoming.*

Vermunt, J. (2003). Multilevel latent class models. *Sociological Methodology*, 213–239.

Vermunt, J. (2008a). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research 17*(1), 33.

Vermunt, J. K. (2008b). *Latent Class Models in Longitudinal Research*, Chapter Handbook of Longitudinal Research: Design, Measurement, and Analysis. Burlington, MA: Elsevier.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association* (58), 236–244.

Widmer, E. D. and G. Ritschard (2009, Mar). The de-standardization of the life course: Are men and women equal?. *Advances in Life Course Research 14*(1-2), 28–39.

Wilson, C. (2001, Jan). Activity patterns of canadian women: Application of clustalg sequence alignment software. *Transportation Research Record 1777*, 55–67.

Wilson, C. (2006). Reliability of sequence-alignment analysis of social processes: Monte carlo tests of clustalg software. *Environment and Planning A 38*(1), 187–204.

Wilson, C. (2008, Jan). Activity patterns in space and time: calculating representative hagerstrand trajectories. *Transportation 35*, 485–499.

Wu, L. (2000). Some comments on" sequence analysis and optimal matching methods in sociology: Review and prospect". *Sociological Methods & Research 29*(1), 41–64.