Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2012)
61, *Part* 5, *pp.* 765–784

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

Nicola Barban

University of Groningen, The Netherlands

and Francesco C. Billari

Universitá Bocconi, Milan, Italy

[Received October 2010. Final revision January 2012]

Summary. We compare two techniques that are widely used in the analysis of life course trajectories: latent class analysis and sequence analysis. 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 that is obtained via the two techniques by using a data set on the life course trajectories of young adults. Then, we adopt a simulation approach to measure the ability of these two methods to classify groups of life course trajectories correctly when specific forms of 'random' variability are introduced within prespecified classes in an artificial data set. 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 latent class analysis and sequence analysis and on the other hand to illuminate the potential pitfalls in applications of these techniques.

Keywords: Categorical time series; Latent class analysis; Life course analysis; Sequence analysis

1. Introduction

In recent years, there has been a strongly growing interest in the holistic study of life course trajectories, i.e. in considering whole trajectories as units of analysis. This growth has taken place, in a somewhat unrelated way, in life course research both in the social sciences and in epidemiology. A particular focus has been the classification of individuals according to their life course trajectories, 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 by using two of the most widespread approaches to such problems: latent class analysis (LCA) and sequence analysis (SA).

The two techniques, LCA and SA, come from different statistical backgrounds. SA, in its various specifications, is based on algorithmic, or data mining, approaches using measures of dissimilarity, or distance, between individual trajectories (Abbott, 1995; Billari and Piccarreta, 2005; Elzinga, 2006). The SA approach is fully non-parametric, 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 that are 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. LCA,

Address for correspondence: Nicola Barban, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands.
E-mail: n.barban@rug.nl

 0035–9254/12/61765

in its various specifications, is based on a probabilistic modelling approach, with a finite mixture distribution as the data-generating mechanism (Hagenaars and McCutcheon, 2002; 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 procedure aims at estimating the probability of class membership for each trajectory on the basis of observed data, usually via 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.

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 and criminal careers. Using either LCA or SA techniques, individuals are assigned to homogeneous classes that are interpreted as representing typical behaviours (Roeder *et al.*, 1999; McVicar and Anyadike-Danes, 2002; Macmillan and Eliason, 2003; Nagin and Tremblay, 2005; Aassve *et al.*, 2007; Amato *et al.*, 2008). 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).

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 the life course and to analyse their determinants (Croudace *et al.*, 2003; Dunn *et al.*, 2006; Bruckers *et al.*, 2010). Other studies focus on the link between health or behavioural trajectories and later outcomes during a life course (Hamil-Luker and O'Rand, 2007; Savage and Birch, 2010). Although 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, whereas 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, subjective health and wellbeing are often measured by using discrete scales (see, for example, Sacker *et al.* (2011) and Westerlund *et al.* (2009)). Also functional disabilities and chronic diseases are often expressed by using categorical variables (Arber and Ginn, 1993; Ross and Wu, 1996). Lastly, SA methods may be used to describe the occurrence and persistence over time of particular health-related behaviours such as smoking and drinking. For these reasons, we restrict the analysis to LCA and SA techniques for categorical variables.

In the remainder of this paper, we compare the performance of 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 that is obtained via the two techniques by using a data set on the life course trajectories of young adults. Then, a simulation approach is adopted to measure the ability of these two methods to classify groups of life course trajectories correctly when specific forms of 'random' variability are introduced within prespecified classes in an artificial data set. 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 and on the other hand to illuminate the potential pitfalls in applications of these techniques.

2. Life course trajectories as categorical discrete time longitudinal data

Life course trajectories are here 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 (perhaps ordered) states. The approach that is used here can, however, be extended to the categorization of variables that are measurable on a quantitative scale (e.g. systolic blood pressure level or income) over discrete (or discretized) 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 that was proposed by Elder (1985) for the social sciences. According to Elder, life course trajectories 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 analysed by representing the original data, i.e. each individual's life course, as a sequence of states. Each individual i is associated with a variable s_{it} indicating her or his life course status at time t. As we assume that s_{it} takes a finite number of values, trajectories can be described as categorical discrete time longitudinal data. In other words, 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 represents 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 years) then starts a cohabitation lasting 5 months and then marries and remains married for 7 months can be described as

$$\text{SSSSSSSSSSSCCCCCMMMMMMM}.$$

In this case, the state space has three values (S, single; M, married; C, cohabiting) and the timescale is monthly. 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} and $i = 1, \ldots, n$. The life course trajectory of 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

$$(\text{S}, 12) - (\text{C}, 5) - (\text{M}, 7).$$

Life course sequences $\{s_{i1}, \ldots, s_{iT}\}$ can be also represented by a series of binary variables. A trajectory with a state space composed of k categories can be represented by $M = K - 1$ binary variables measured in t occasions. Our previous example would require two variables (the former indicating whether the respondent is married; the latter indicating whether she or he is cohabiting) measured on 24 occasions. 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 LCA and SA in the study of life course trajectories.

3. Latent class analysis of life course trajectories

LCA is a statistical technique which can be used to classify individuals on the basis of a set of categorical outcomes (Lazarsfeld and Henry, 1968; Goodman, 1974; Clogg, 1995). 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 called 'latent trajectory modelling' or 'longitudinal LCA' (Vermunt, 2008a; 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 $i = 1, \ldots, N$ subjects, $j = 1, \ldots, M$ life domains, $c = 1, \ldots, C$ mutually exclusive and exhaustive latent classes and $t = 1, \ldots, T$ periods. The conditional likelihood for each subject is

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

where π_{cjt} is the probability of the jth outcome being equal to 1 at time t for class c, with $0 < \pi_{cjt} < 1$.

Summing over the classes, weighted by η_c, we obtain the marginal likelihood

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

LCA assumes that the structure of correlation between observed variables is completely explained by latent factors. This condition is called 'conditional independence', i.e. $(y_{i11}, \ldots, y_{iMT} | c_i = c) \perp\!\!\!\perp (y_{i11}, \ldots, y_{iMT} | c_i = d)$ with $d \neq c$ (Hagenaars, 1988; Uebersax, 1999).

The principal drawback of using standard LCA for longitudinal data is that these models do not take into consideration the time correlation between variables. Measurements of the same variable in different time periods are considered independent. An alternative approach dealing with longitudinal measurements of categorical variables is latent transition analysis (LTA), which is also known as the class of latent Markov models or hidden Markov models (van de Pol and Langeheine, 1990; Collins and Wugalter, 1992; Vermunt, 2008a; Reboussin and Ialongo, 2010). LTA is an extension of LCA that makes it possible to model a dynamic, or changing, latent variable. The basic idea of this class of models is to describe the change over time in the categories of the latent variable, which are referred to in this context as latent statuses. LTA produces parameter estimates corresponding to the proportion of individuals in each latent status at the initial time and the probability of each possible item response conditional on latent status membership. LTA also produces a transition probability matrix, consisting of estimates of the probability of latent status membership at time $t + 1$ conditional on latent status membership at time t. This class of models typically relies on the assumption of Markovian dependence of latent statuses over time. LTA is then used to study the evolution over time in latent status membership. These models are very useful to study the transitions that happen during a life course, but they cannot directly be used to classify life course trajectories as units of analysis.

In recent years, various forms of correction have been proposed to adjust for the temporal correlation between observations, mainly by including a random effect in the model (Vermunt, 2003, 2008b; Beath and Heller, 2009; Hadgu and Qu, 1998; Scott, 2011). In later analyses, we refer to the more standard version of LCA as applied to longitudinal data.

4. Sequence analysis and optimal matching

SA is a family of algorithm-based techniques that are used to quantify the dissimilarity between categorical time series. The optimal matching (OM) algorithm is the most widely known technique in the social sciences. OM is a family of dissimilarity measures derived from what was originally proposed in the field of information theory and computer science by Levenshtein (1965). Abbott (1995) adapted OM to life course analysis. Basically, OM expresses distances between sequences in terms of the minimal amount of effort, measured in terms of subsequent

operations, that is required to change two sequences so that they become identical. A set that is composed of three basic operations to transform sequences is used: $\Omega = \{\iota, \delta, \sigma\}$, where ι 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 $\omega_z \in \Omega$, a specific cost can be assigned, $c(\omega_z)$. If z basic operations must be performed to transform one sequence into another the transformation cost can be computed as $c(\omega_1, \ldots \omega_Z) = \Sigma_{z=1}^{Z} c(\omega_z)$. The distance between two sequences can thus be defined as the minimum cost of transforming one sequence into the other, giving a symmetric matrix of pairwise distances that can be used for further statistical analysis, mainly multivariate analysis.

Although OM has been the most widespread approach to SA since Abbott's work, its use in the analysis of life course trajectories has often been criticized (Brzinsky-Fay and Kohler, 2010; Aisenbrey and Fasang, 2010). First, it is difficult to attribute a life course meaning to the sequence operations (Lesnard, 2006). Life course sequences are time referenced. Therefore, the operations imply modifications in the timescale. In particular, insertion and deletion operations warp time to match identically coded states but occurring at different moments in their respective sequences. In contrast, substituting two events maintains the original timescale of events without warping time. A simple solution to avoid 'indel' (insertion and deletion) 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 in the use of OM for life course analysis because of their arbitrariness and the weak link to theory (Wu, 2000). Critics argue that the resulting distances are meaningless from a sociological point of view (Levine, 2000). In the case in which there is no clear ranking between different states, the definition of cost is necessarily arbitrary. A common practice is therefore to set constant costs independently of the states that are substituted. This is equal to setting $c(\iota) = c(\delta)$ and $c(\sigma) = 2c(\delta)$. Using this approach, $c(\iota)$ 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 $c(\sigma) = 2\,c(\delta)$ is that, in a constant cost framework, substitution is equivalent to a deletion followed by an insertion. Alternatively, it is possible to adopt empirical costs, i.e. using substitution costs that are inversely proportional to transition frequencies between two states (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 can be defined as $c(\sigma; a, b) = c(\sigma; b, a) = 2 - p_{t,t+1}(a,b) - p_{t,t+1}(b,a)$ 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 criticism of this approach is that transitions at different times are qualitatively different. For this reason, Lesnard (2006) proposed a modification of the Hamming distance using dynamic costs. The 'dynamic Hamming distance' is based on time varying substitution costs $c_t(\sigma; a, b)$.

Third, it is not clear how to treat missing data and censoring by using SA. In fact, unequal sequence length due to censoring should not contribute to the distance between the sequences. A common practice is to restrict the analysis to sequences of the same length to avoid compar-

ing sequences of different length. Elzinga (2006) proposed 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 to assess a similarity measure. 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 and last k elements of a sequence. Elzinga (2006) reviewed in detail different distance measures based on subsequences. According to Elzinga's proposal, 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. The most used measure based on subsequences is the longest common subsequence (LCS). The theoretical basis of subsequence-based measures originates from information science and their great advantage is that the researcher does not need to specify any operation costs.

5. The consistency of latent class analysis and sequence analysis: an example using real life course data

In recent decades, the transition to adulthood has been one of the most important areas in the life course literature within the social sciences (Aassve *et al.*, 2007; Abbott, 1995). Following the seminal work of Modell *et al.* (1976), the transition to adulthood can be described by referring to a number of important events. For instance these events can be finishing school, beginning full-time employment, entering a non-marital cohabitation, becoming a parent and being married. The fact that these events can occur in different orders and at different ages yields an enormous number of possible combinations. In this section, we analyse data from a published study (Amato *et al.*, 2008) on life course trajectories of young adults to test the consistency of the classification that is obtained via LCA and SA.

5.1. The data

The data that are used in this analysis come from waves I and III of the US National Longitudinal Study of Adolescent Health. The survey is a longitudinal and nationally representative data set of American adolescents who were in grades 7–12 in 1994–1995. In the first wave, data were collected through in-home interviews with the adolescent participants and one of their parents. Typically, the parent interview was completed by the biological mother. Adolescents were interviewed again in a second wave 1 year later in 1996, again in a third wave collected in 2001–2002 when their ages ranged from about 18 to 25 years and finally in a fourth wave in 2008–2009. The data contain retrospective information on age at marriage, cohabitation, childbearing, education and full-time employment. Our analysis is restricted to women 23 years of age or older at wave III. The final sample size is 2290. The analysis focuses on young women for two reasons. First, the timing of family formation events tends to be earlier for women than for men. Second, the decision to exclude men from our analysis rests on substantial limitations of the data. Indeed, as pointed out by Amato *et al.* (2008), there is a systematic misreporting of childbirths in the fertility history modules.

5.2. Empirical analysis

Amato *et al.* (2008) proposed the use of LCA to create transitions to adulthood pathways for women between the ages of 18 and 23 years. Input data include states of cohabitation, marriage, parenthood, full-time employment and school attainment measured in six periods. The original

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%). Fig. 1 shows the estimates of a latent class model where we replicated the original analysis. The number of classes has been selected by looking at the Bayesian information criterion for various models.

Would SA lead to the same results? The first possible test is to run an SA with the same data and to compare the groups that are obtained by the two methods. Family formation trajectories can be described by the joint occurrence of the five variables that were described above. The resulting sequence is six periods long and the state space is composed of $2^5 = 32$ 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 SA, we calculated a dissimilarity matrix using different distances (see Section 4).

Starting from each of these dissimilarity matrices, a cluster analysis was conducted by using the Ward algorithm (Ward, 1963). The Ward clustering algorithm can be briefly described as follows. Consider N sequences to be clustered. Let $d(i, j)$ denote the distance between the ith and the jth sequence. The total dispersion, i.e. the amount of dispersion within the whole data set, is usually measured as $T = \Sigma_{i,j}\, d(i, j)$. Suppose now that the whole sample is partitioned into G clusters. The dispersion within the gth cluster is $W_g = \Sigma_{i,j \in g}\, d(i, j)$, and the dispersion within the G groups can be summarized as $W_G = \Sigma_{g=1}^{G} W_g$. 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 cluster 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 Ward's algorithm the two clusters to be joined are selected by minimizing the increase in the within-groups dispersion consequent on the reduction of the number of partitions:

$$\Delta(g|g_{\mathrm{L}}, g_{\mathrm{R}}) = W_g - W_{g\mathrm{L}} - W_{g\mathrm{R}} = W_{G-1} - W_G \tag{2}$$

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) classifications having a decreasing number of clusters, $\{P_{\max}, P_{\max-1}, \ldots, P_1\}$, '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 in the number of clusters.

The agreement in classification between the two techniques is measured by the strength of association between the classes obtained, calculated by using the corrected version of the Rand index (Rand, 1971). 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

$$P_s = \sum\sum P_{ij}^2 + \sum\sum P_{ij}(1 - P_{i+} - P_{+j} + P_{ij}) \tag{3}$$

$$= 1 - \sum P_{i+}^2 - \sum P_{+j}^2 + 2\sum\sum P_{ij}^2. \tag{4}$$

This measure of agreement has the advantage that it can be used even if the sizes of the two clusters (k_1 and k_2) differ. However, the Rand index makes no correction for chance agreements. Therefore, it is not possible to tell whether a specific value of P_s is 'large' or 'small', because its

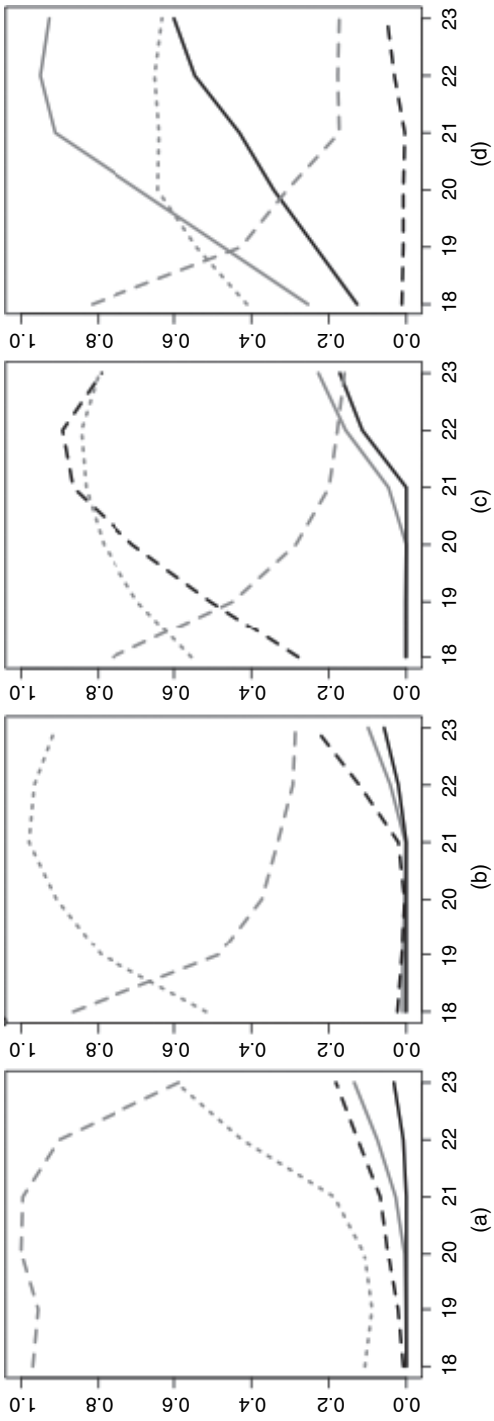

Fig. 1. Latent class representation of early family formation, women 18–23 years old, replicated from Amato *et al.* (2008) (– – –, cohabitation; ——, marriage; ——, children; – – –, in school; – – –, full-time work): (a) college–no family formation; (b) high school–no family formation; (c) cohabiting without children; (d) married mothers; (e) single mothers; (f) cohabiting mothers; (g) inactive

Table 1. Agreement in classification between LCA and SA techniques: application to real life course data on transition to adulthood

	Corrected Rand index
OM with empirical costs	0.59
LCS	0.55
OM with constant costs	0.55
Dynamic Hamming distance	0.50
Hamming distance	0.19

value when individuals are classified at random (i.e. $P_{ij} = P_{i+}P_{+j}$) is not 0, and depends on P_{i+} and P_{j+}. Therefore, 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 is

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

This statistic equals 1 for perfect agreement, $\Omega = 0$ for chance agreement and $\Omega < 0$ when agreement is less than expected by chance.

The LCA and SA have been performed using the R packages `poLCA` and `TraMineR` (Linzer and Lewis, 2011; Gabadinho *et al.*, 2011).

5.3. Results

In this application (see the results in Table 1), OM with empirically derived costs gives the closest solution to the classes that were identified by LCA in the original study by Amato *et al.* (2008). The corrected Rand index indicates that the percentage of couples classified in agreement is equal to 0.59. If we do not correct for the percentage of agreement due to chance, we would obtain a Rand index of 0.88. The LCS distance does not imply any cost settings. The cluster solution that is obtained with this method is very similar to the OM solution (0.55 corrected Rand index). 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 classifications. However, the cluster solutions that are obtained with the Hamming distance diverges substantially from the LCA solution.

This comparison does not support the use of one particular approach, but it gives a first indication of the consistency of different statistical methods for life course analysis. In particular, it is interesting to note that the two methods that give closest agreement with LCA are OM with empirical costs and LCS. In what follows we shall analyse simulated data, comparing LCA and SA by using these two distances. Although LCA and some SA techniques give consistent results, it is not possible to draw any conclusion about the reliability of the methods if the generating mechanism of life course sequences is unknown.

6. Simulation study

We propose a simulation approach to study the factors affecting the robustness of classification

by using LCA and SA. Consequently, we use a simulated data set that mimics the real data that were presented in Section 5. The simulation procedure can be summarized in four steps.

(a) Define artificial typical groups of life course trajectories.
(b) Introduce variability in the timing, quantum and sequencing of life course trajectories.
(c) Classify the individuals of the artificial data set by using LCA and SA.
(d) Compare classifications that are obtained with the two techniques with the groups as defined in the original data set.

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 that is proposed in this paper, however, follows a different approach. Instead of starting from a stochastic generating mechanism, the reliability of SA techniques is tested by increasing the level of heterogeneity among sequences belonging to the same original group. We adopt a 'black box' approach in which we focus only on the variability of life course trajectories without being interested in the generating process. In this way, we attempt to investigate the robustness of different techniques under different variations.

6.1. Defining typical groups of life course trajectories

Let us define four different groups of life course trajectories representing family formation. For simplicity, we define a state space composed of only three states S, C and M, indicating different partnership statuses. We can thus consider S as single, C as cohabiting and M as married. For each sequence, we set the length equal to 30 and S as the initial state. Then, we repeat every typical sequence 250 times, obtaining an artificial data set of 1000 observations. The data set can be considered as a monthly (quarterly) collection of data indicating the marital or union status of an individual. Let us define four 'typical' groups of sequences: group 1,

$$(\mathrm{S}, 10) - (\mathrm{C}, 10) - (\mathrm{M}, 10);$$

group 2,

$$(\mathrm{S}, 20) - (\mathrm{C}, 5) - (\mathrm{M}, 5);$$

group 3,

$$(\mathrm{S}, 10) - (\mathrm{C}, 5) - (\mathrm{M}, 10) - (\mathrm{C}, 5);$$

group 4,

$$(\mathrm{S}, 20) - (\mathrm{M}, 10).$$

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

6.2. Introducing variability in the typical sequences

To test the reliability of the two methods, we introduce random perturbations in the timing, quantum and sequencing of trajectories. Thus, we introduce a series of sequence operators that modify life trajectories. These operations introduce variability in the groups. Even if these operations do not have a specific meaning in the social sciences, we try to mimic some analysed within the life course literature or due to the observational plan.

Let us define the following operators.

(a) Postponement: with probability p (postponement rate), copy status from time t to time $t+1$, e.g.

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

(b) Slicing: with probability p (slicing rate), exchange two subsequences of the same length. In our simulations the length of subsequences is set to 5, e.g.

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

(c) Inversion: with probability p (inversion rate), exchange all the elements C with elements M, e.g.

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 *M M M M M M M M M M* C C C C C C C C C C

(d) Mutation: with probability p (mutation rate), substitute sequence status at time t with a random element of the alphabet, e.g.

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

(e) Truncation: with probability p cut the sequence at time t, with t randomly chosen. In our simulations, truncation is limited to the second half of the sequence, i.e. $t > 15$, e.g.

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 C C C C C C C C C C M M M M

The operators proposed are meant to introduce variations in the various components of life course introducing variability between sequences. Some individuals may postpone (or anticipate) a transition, whereas others invert the 'order' in which events happen. The order of events is modified in a radical way under inversion and partially under slicing. 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 at certain points in time. In contrast, truncation is a common feature in the observation of life course data. Using this disturbance strategy allows the reliability of different methods to be tested without assuming any generating mechanism of the data. Fig. 2 illustrates the effects of different sequence operators.

6.2.1. How to measure variability in timing, quantum and sequencing

6.2.1.1. Timing. The exit time from the first state is a crucial event in the transition to adulthood. As a naive indicator of timing, we define the age at the first transition. We define an indicator τ as the proportion of a sequence that is spent in the initial status:

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

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

6.2.1.2. Quantum. The number of events is a key element that characterizes a life course trajectory. We define a simple indicator ρ as the number of transitions per time period:

$$\rho = \frac{\#\{s_{(t-1)} \neq s_t\}}{T}.$$

6.2.1.3. Sequencing. The order in which events occur is crucial in several studies on life course. For example, it may be relevant to study the divergence of a life trajectory from the normative order of events. For this reason, we propose as an indicator the number of non-normative transitions, i.e. the transitions that diverge from a given sequence of events considered normative in the society. We define an indicator ς as the proportion of normative transitions over the total number of transitions:

$$\varsigma = \frac{\text{number of normative transitions}}{\text{total number of transitions}}.$$

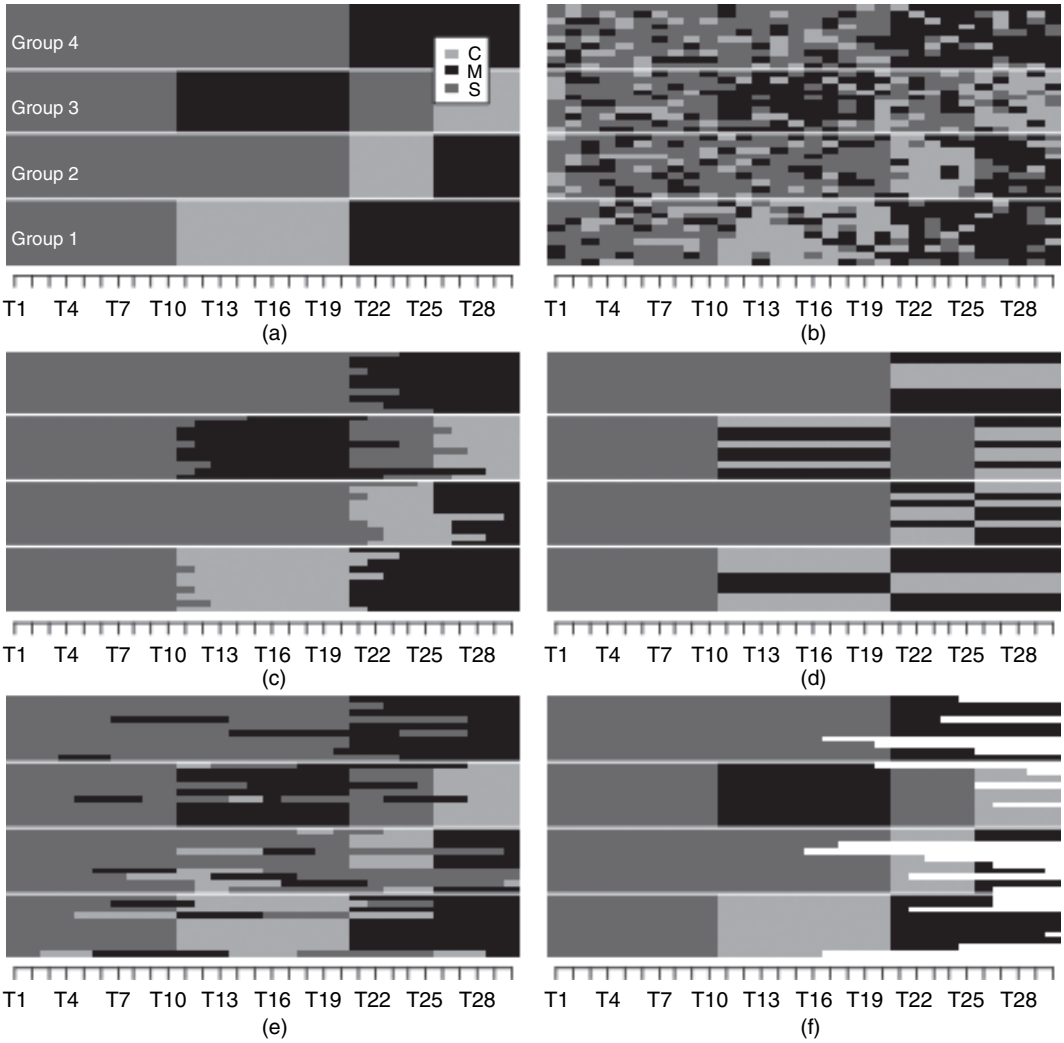

Fig. 2. Effects of various sequence operators on the initial sample: (a) original sample; (b) mutation, $p = 0.5$; (c) postponement, $p = 0.5$; (d) inversion, $p = 0.5$; (e) slicing, $p = 0.5$; (f) truncation, $p = 0.5$

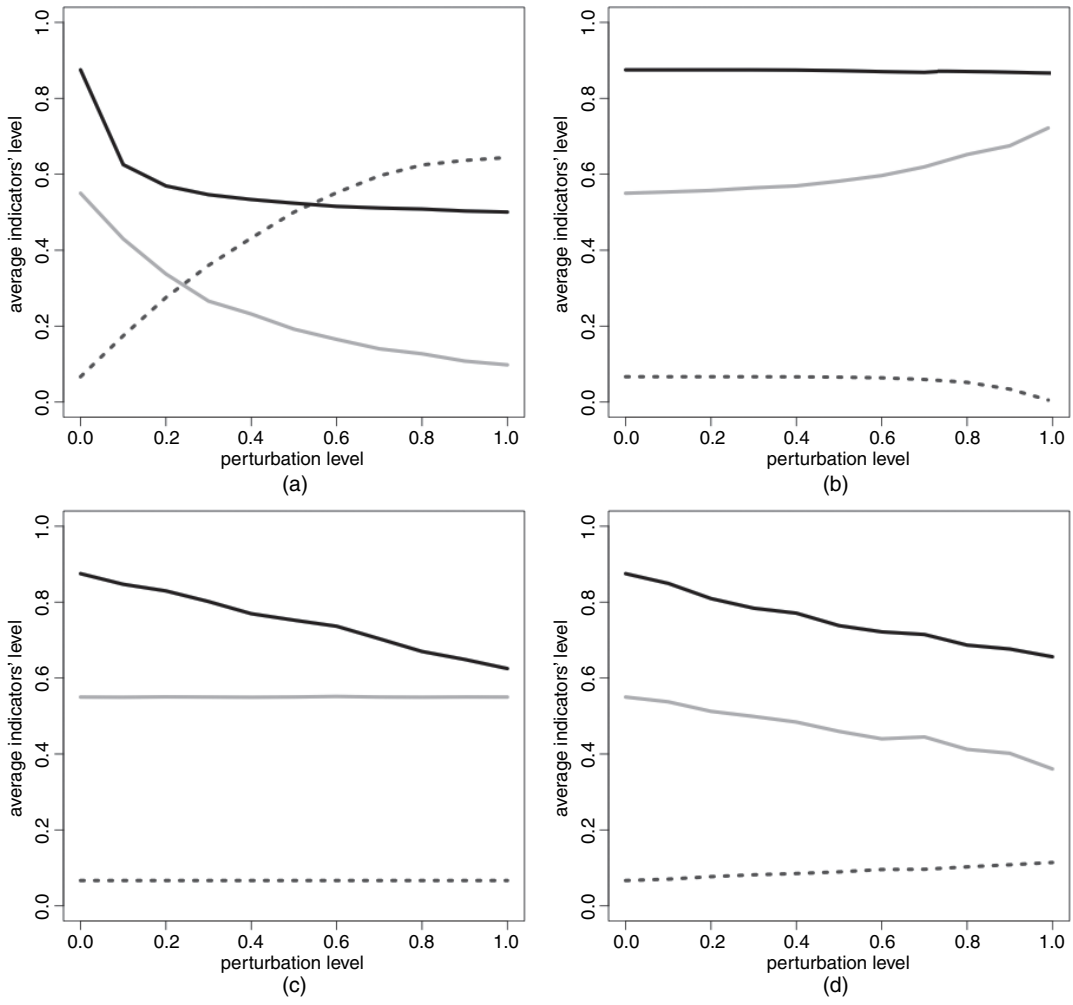

Fig. 3. Average levels of timing (——) quantum (– – –) and sequencing (——) indicators by various perturbation levels from 0 to 1 in (a) mutation, (b) postponement, (c) inversion and (d) slicing

The indicators of timing, quantum and sequencing range between 0 and 1 for each sequence. The operators that are defined above have different effects on indicators. Postponement introduces a major change in timing whereas quantum and sequencing remain unaltered. Inversion modifies only sequencing because it transforms an entire category of events into another. Slicing modifies both sequencing and quantum. Last, mutation has a massive effect on quantum, but it also affects the other two dimensions introducing completely random variations. The effects of the sequence operators are illustrated in Fig. 3. The average values of the indicators for timing, quantum and sequencing in the initial sample (without perturbations) are respectively $\bar{\tau} = 0.55$, $\bar{\rho} = 0.07$ and $\bar{\varsigma} = 0.88$.

7. Simulation results

We simulate a total number of 1000 samples of 1000 observations each applying different levels of perturbation. The original sample is modified by using separately the sequence operators

Table 2. Simulation results: classification agreement of LCA and SA (OM and LCS) under various sequence operators†

p		*Results for inversion*			*Results for postponement*			*Results for mutation*			*Results for slicing*			*Results for truncation*		
		LCA	*OM*	*LCS*	*LCA*	*OM*	*LCS*	*LCA*	*OM*	*LCS*	*LCA*	*OM*	*LCS*	*LCA*	*OM*	*LCS*
0.1	Mean	0.710	0.812	0.812	0.947	1.000	1.000	0.979	0.997	0.997	0.687	0.927	0.932	0.740	0.655	0.659
	Variance	0.008	0.001	0.001	0.013	0.000	0.000	0.006	0.000	0.000	0.009	0.001	0.001	0.018	0.001	0.002
0.2	Mean	0.570	0.673	0.673	0.942	1.000	1.000	0.965	0.975	0.973	0.604	0.859	0.871	0.675	0.579	0.580
	Variance	0.006	0.000	0.000	0.014	0.000	0.000	0.008	0.000	0.000	0.006	0.007	0.005	0.015	0.001	0.001
0.3	Mean	0.472	0.556	0.556	0.947	0.995	0.995	0.958	0.925	0.921	0.519	0.731	0.767	0.622	0.513	0.519
	Variance	0.005	0.000	0.000	0.016	0.000	0.000	0.004	0.001	0.001	0.002	0.015	0.013	0.010	0.002	0.002
0.4	Mean	0.425	0.476	0.479	0.952	0.981	0.982	0.914	0.827	0.815	0.447	0.615	0.654	0.604	0.469	0.480
	Variance	0.006	0.001	0.002	0.011	0.000	0.000	0.003	0.001	0.001	0.003	0.006	0.011	0.014	0.005	0.007
0.5	Mean	0.422	0.364	0.446	0.893	0.950	0.952	0.825	0.669	0.648	0.398	0.579	0.587	0.569	0.456	0.455
	Variance	0.004	0.003	0.000	0.019	0.001	0.001	0.001	0.002	0.002	0.002	0.004	0.004	0.007	0.011	0.011
0.6	Mean	0.413	0.476	0.479	0.872	0.896	0.894	0.663	0.466	0.429	0.363	0.547	0.559	0.550	0.475	0.493
	Variance	0.008	0.001	0.000	0.012	0.002	0.002	0.003	0.003	0.003	0.002	0.003	0.003	0.011	0.014	0.015
0.7	Mean	0.470	0.556	0.556	0.772	0.782	0.776	0.408	0.244	0.213	0.354	0.527	0.530	0.561	0.494	0.500
	Variance	0.008	0.001	0.001	0.009	0.003	0.004	0.002	0.002	0.002	0.003	0.003	0.003	0.014	0.018	0.019
0.8	Mean	0.579	0.670	0.670	0.563	0.535	0.562	0.139	0.079	0.070	0.333	0.513	0.524	0.567	0.569	0.560
	Variance	0.006	0.001	0.001	0.007	0.008	0.007	0.002	0.003	0.004	0.002	0.002	0.003	0.009	0.014	0.016
0.9	Mean	0.698	0.819	0.819	0.224	0.195	0.195	0.003	0.011	0.011	0.294	0.517	0.528	0.587	0.629	0.623
	Variance	0.009	0.001	0.001	0.002	0.002	0.002	0.000	0.000	0.000	0.004	0.004	0.004	0.010	0.008	0.009
0.99	Mean	0.908	0.979	0.979	0.003	0.003	0.003	0.000	0.000	0.000	0.268	0.502	0.508	0.605	0.652	0.650
	Variance	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.004	0.004	0.009	0.005	0.005

†Perturbation levels from 0 to 1. Average level of Ω and simulation variance.

(mutation, postponement, mutation, slicing and truncation) that were described in Section 6.2. For each sample, we estimate a latent class model with four classes and we calculate OM and LCS matrices of dissimilarity. We then apply cluster analysis by using Ward's algorithm to classify individuals into four groups. The groups that are obtained are compared with the original groups by using the corrected Rand index. Table 2 and Fig. 4 report the average rate of agreement between the original groups and the results that are obtained by LCA and SA (respectively OM and LCS). Specifically, for each sequence operator, we apply 10 different levels of perturbation $p = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99$ and we calculate the average classification agreement with the original groups.

With simulated data, we apply LCA and SA in the same way that we did with real data. Following a common approach for demographic studies, we estimated OM distances by using costs dependent on transition frequency. LCA was conducted, by defining two binary variables in each time period indicating whether the individual is single S, cohabiting C and married M. To avoid local maxima we run the model three times and we choose the model with the minimum value of the Bayesian information criterion. For practical purposes, both the number of classes and the number of clusters is fixed at 4. Varying the number of classes gave similar results in terms of classification performance.

The results (Table 2 and Fig. 4) show that classification is sensitive to the transformations that are induced by sequence operators. With increasing variability in the sample, the classification performance decreases. As expected, the classification performance of all the methods decreases rapidly with random mutation. Mutation, in fact, can be considered a benchmark since it introduces the maximum amount of variation. The agreement under postponement decreases more slowly. In particular small postponement levels 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 state. But a large postponement has also an effect on quantum, since it reduces the number of transitions in trajectories and reduces the variability between different groups of sequences. Inversion has the maximum confounding effect at level 0.5. At that point, exactly 50% of sequences have all 'C' inverted with 'M' and vice versa. With greater of levels of inversion, 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 and the classification performance decreases almost linearly. The performance of classification under truncation follows a U-shape. An increase in the truncation levels affects the number of censored individual sequences. It follows that high truncation rates are associated with sequences that are shorter on 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 level of truncation is high.

To summarize the results, we propose a measure of the overall performance. Let R be the number of simulations, and $\Omega_r^{\{\mathrm{LCA};\mathrm{OM};\mathrm{LCS}\}}$ the correct Rand index for the sample r under different sequence operators. A simple index of the overall goodness of the 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 3 summarizes the results.

Results obtained in our simulations suggest some considerations about the reliability of LCA and SA. First, there is no evidence of a technique having superior performances under all the

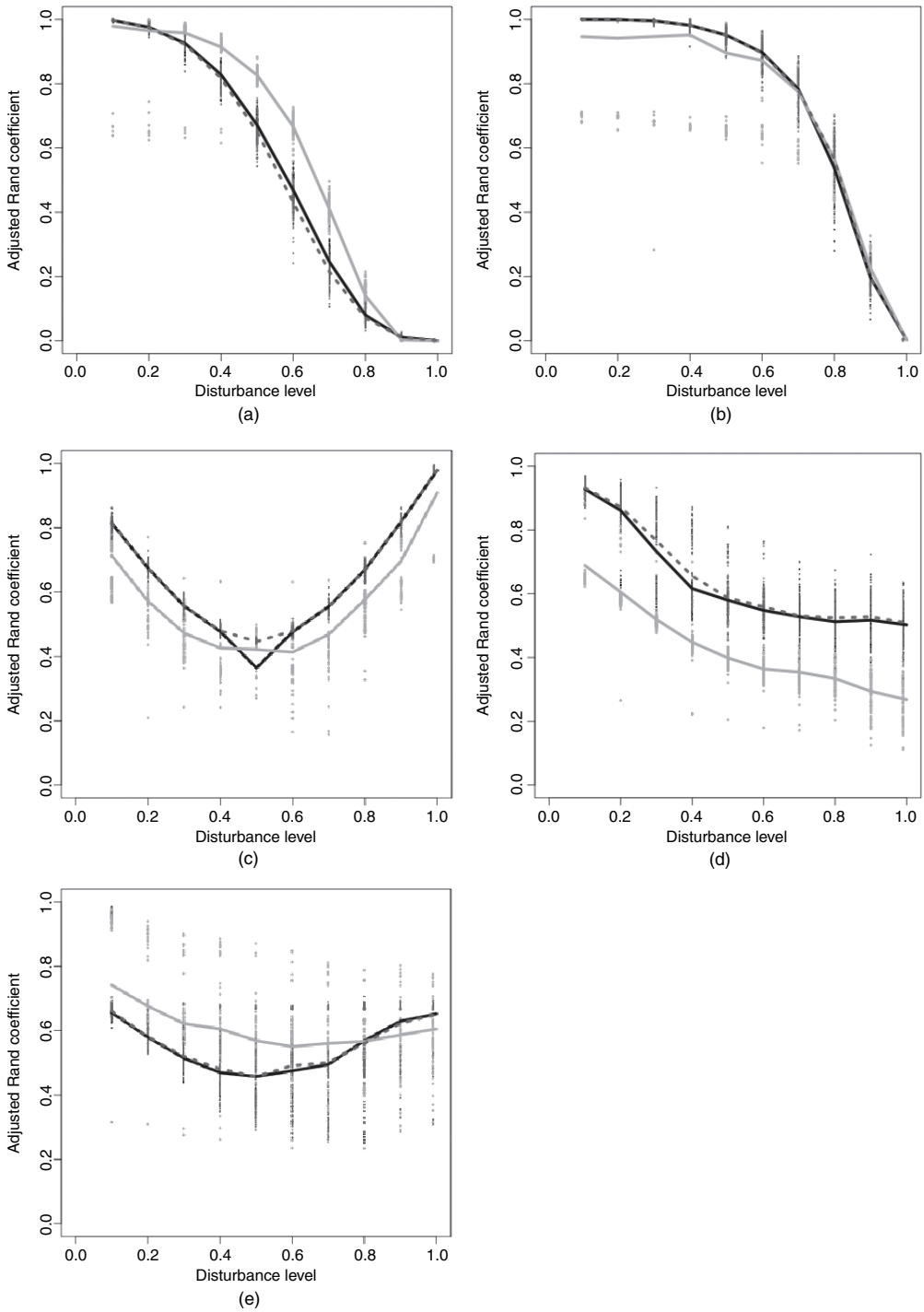

Fig. 4. Simulation results—classification agreement of LCA (——) and SA (OM (——) and LCS (– – –)) under various sequence operators (perturbation levels from 0 to 1; dots indicate the level of classification in each simulation and lines represent average values of Ω): (a) mutation; (b) postponement; (c) inversion; (d) slicing; (e) truncation

Table 3. Overall mean classification agreement ($\bar{\Omega}$) of LCA and SA (OM and LCS) under various sequence operators†

	Mutation	*Postponement*	*Inversion*	*Slicing*	*Truncation*
LCA	0.940	0.947	0.600	0.609	0.702
OM	0.899	0.988	0.647	0.785	0.612
LCS	0.892	0.988	0.661	0.802	0.616

†Perturbation levels from 0 to 0.5.

sources of variation. In fact we do not observe a technique that performs better in all the cases. Despite that, according to our simulations, SA shows greater agreement under inversion and slicing. In contrast, LCA performs better under mutation and truncation. Results from postponement indicate slightly better results for SA. Second, the classifications with LCA are less precise. Using simulated data it is possible to have an indication of the variability of the estimated agreement rates. Under all the sources of error, the results that are obtained with LCA exhibit more variability. Third, the differences between OM and LCS are minimal. Both methods, in fact, produce very similar results. Although the two distances are qualitatively different, the results that are obtained in all the sources of variability are very similar.

8. Discussion

Two techniques have been applied to the classification of life course trajectories as sequences: LCA and SA.

Our analyses on real and artificial data show that the two techniques give consistent results in classifying life course trajectories. Although we do not find a clear superiority of one method, our results show that SA (both OM and LCS) seems to perform better when life course sequences have variations within a group due to sequencing (inversion and slicing operators) and timing (the postponement operator). In contrast, LCA has better results when the variations are completely random (mutation operator) and when data are truncated. 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 data. Nevertheless, mutation can be interpreted as a measurement error, since individuals may be misclassified during repeated measurements. Censoring is also an important issue in life course research. Attrition can be affected by some observable (or unobservable) characteristics of the subjects participating in the study. A parametric model that takes into account the probability of participating in the follow-up can be incorporated in LCA (see, for example, Lin *et al.* (2002)).

Latent class models involve an estimation procedure where the number of parameters depends on the length of the sequence and the number of states. Conversely, SA is a purely algorithmic approach that does not make any assumption about the data-generating mechanism. In life course analysis, it might be necessary to study long trajectories with a large state space (take, for instance, weekly data on job careers), which can lead to inefficient estimation by using LCA. In contrast, the calculation of pairwise distances in the SA approach can be computationally burdensome in large samples.

Our study, of course, has limitations in its scope. It applies to life course sequences, i.e. categorical discrete time longitudinal data, and not to continuous-valued data nor to continuous time longitudinal data. It applies to classification and not to other methods of analysis (e.g.

regression). Also, life course classification may be influenced by other factors (e.g. the length of sequences, the dimension of the state space and the classification algorithm). Moreover, for simplicity, we based our analysis on a fixed number of classes whereas, in both LCA and SA, the number of classes is usually derived by data. Nevertheless, this study represents one of the first attempts to test the reliability of holistic methods for life course analysis.

Acknowledgements

An earlier version of this paper was presented at the conference on 'Statistical challenges in lifecourse research' (Leeds, July 13th–14th, 2010). Participants at the conference provided very useful comments and suggestions. We also thank the Associate Editor and two referees, as well as Jane Klobas, Aart Liefbroer, Stefano Mazzuco, Gianfranco Lovison and Ingrid Svensson, and seminar participants at Bocconi University and the University of Padua for comments and suggestions that helped to improve this paper. This research uses data from the National Longitudinal Survey of Adolescent Health, which is a programme project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with co-operative funding from 23 other federal agencies and foundations (Harris *et al.*, 2009). Special acknowledgment is due to Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the data files for the survey is available from `http://www.cpc.unc.edu/addhealth`. No direct support was received from grant P01-HD31921 for this analysis.

The National Longitudinal Study of Adolescent Health is a longitudinal study of a nationally representative sample of adolescents in grades 7–12 in the USA during the 1994–1995 school year. The cohort has been followed into young adulthood with four in-home interviews, the most recent in 2008, when those in the sample were aged 24–32 years. The survey combines longitudinal survey data on respondents' social, economic, psychological and physical wellbeing with contexual data on the family, neighbourhood, community, school, friendships, peer groups and romantic relationships, providing unique opportunities to study how social envionments and behaviours in adolescence are linked to health and achievement outcomes in young adulthood. The fourth wave of interviews expanded the collection of biological data to understand the social, behavioural and biological linkages in health trajectories as the cohoft ages through adulthood.

References

Aassve, A., Billari, F. C. and Piccarreta, R. (2007) Strings of adulthood: a sequence analysis of young British women's work-family trajectories. *Eur. J. Popln*, **23**, 369–388.

Abbott, A. (1995) Sequence analysis: new methods for old ideas. *A. Rev. Sociol.*, **21**, 93–113.

Aisenbrey, S. and Fasang, A. (2010) New life for old ideas: the "second wave" of sequence analysis bringing the course back into the life course. *Sociol. Meth. Res.*, **38**, 420–462.

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

Arber, S. and Ginn, J. (1993) Gender and inequalities in health in later life. *Socl Sci. Med.*, **36**, 33–46.

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

Billari, F. C. (2001) The analysis of early life courses: complex descriptions of the transition to adulthood. *J. Popln Res.*, **18**, 119–142.

Billari, F. C. and Piccarreta, R. (2005) Analyzing demographic life courses through sequence analysis. *Math. Popln Stud.*, **12**, 81–106.

Bruckers, L., Serroyen, J., Molenberghs, G., Slaets, H. and Goeyvaerts, W. (2010) Latent class analysis of persistent disturbing behaviour patients by using longitudinal profiles. *Appl. Statist.*, **59**, 495–512.

Brzinsky-Fay, C. and Kohler, U. (2010) New developments in sequence analysis. *Sociol. Meth. Res.*, **38**, 359–364.
Clogg, C. C. (1995) Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (eds G. Arminger, C. C. Clogg and M. E. Sobel). New York: Plenum.
Collins, L. and Wugalter, S. (1992) Latent class models for stage-sequential dynamic latent variables. *Multiv. Behav. Res.*, **27**, 131–157.
Croudace, T., Jarvelin, M., Wadsworth, M. and Jones, P. (2003) Developmental typology of trajectories to nighttime bladder control: epidemiologic application of longitudinal latent class analysis. *Am. J. Epidem.*, **157**, 834–842.
Dunn, K. M., Jordan, K. and Croft, P. R. (2006) Characterizing the course of low back pain: a latent class analysis. *Am. J. Epidem.*, **163**, 754–761.
Elder, G. H. (1985) *Life Course Dynamics: Trajectories and Transitions, 1968–1980*. Ithaca: Cornell University Press.
Elzinga, C. (2006) Sequence analysis: metric representations of categorical time series. *Sociol. Meth. Res.*, **38**, 463–481.
Gabadinho, A., Ritschard, G., Müller, N. S. and Studer, M. (2011) Analyzing and visualizing state sequences in R with TraMineR. *J. Statist. Softwr.*, **40**, 1–37.
Goodman, L. A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.
Hadgu, A. and Qu, Y. (1998) A biomedical application of latent class models with random effects. *Appl. Statist.*, **47**, 603–616.
Hagenaars, J. (1988) Latent structure models with direct effects between indicators: local dependence models. *Sociol. Meth. Res.*, **16**, 379–405.
Hagenaars, J. A. and McCutcheon, A. L. (2002) *Applied Latent Class Analysis*. Cambridge: Cambridge University Press.
Hamil-Luker, J. and O'Rand, A. M. (2007) Gender differences in the link between childhood socioeconomic conditions and heart attack risk in adulthood. *Demography*, **44**, 137–158.
Hamming, R. (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.*, **26**, 147–160.
Harris, K. M., Halpern, C. T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P. and Udry, J. R. (2009) The National Longitudinal Study of Adolescent Health: research design.University of North Carolina at Chapel Hill, Chapel Hill. (Available from `http://www.cpc.unc.edu/projects/addhealth/design`.)
Lazarsfeld, P. F. and Henry, N. W. (1968) *Latent Structure Analysis*. Boston: Houghton Mifflin.
Lesnard, L. (2006) Optimal matching and social sciences. *Manuscript*. Observatoire Sociologique du Changement, Paris.
Levenshtein, V. I. (1965) Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, **10**, 707–710.
Levine, J. (2000) But what have you done for us lately?: commentary on Abbott and Tsay. *Sociol. Meth. Res.*, **29**, 34–40.
Lin, H., Turnbull, B. W., McCulloch, C. E. and Slate, E. H. (2002) Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *J. Am. Statist. Ass.*, **97**, 53–65.
Linzer, D. A. and Lewis, J. B. (2011) poLCA: an R package for polytomous, variable latent class analysis. *J. Statist. Softwr.*, **42**, 1–29.
Macmillan, R. and Eliason, S. R. (2003) Characterizing the life course as role configurations and pathways. In *Handbook of the Life Course* (eds J. T. Mortimer and M. J. Shanahan), pp. 529–554. New York: Springer.
McVicar, D. and Anyadike-Danes, M. (2002) Predicting successful and unsuccessful transitions from school to work by using sequence methods. *J. R. Statist. Soc.* A, **165**, 317–334.
Modell, Jr, F. F. and Hershberg, T. (1976) Social change and transitions to adulthood in historical perspective. *J. Famly Hist.*, **1**, 7–32.
Morey, L. and Agresti, A. (1984) The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. *Educ. Psychol. Measmnt*, **44**, 33–37.
Mouw, T. (2005) Sequences of early adult transitions: a look at variability and consequences. In *On the Frontier of Adulthood: Theory, Research, and Public Policy* (eds R. Settersten, F. Furstenberg and R. Rumbaut). Chicago: University of Chicago Press.
Nagin, D. S. and Tremblay, R. (2005) Developmental trajectory groups: fact or a useful statistical fiction? *Criminology*, **43**, 873–904.
Piccarreta, R. and Billari, F. C. (2007) Clustering work and family trajectories by using a divisive algorithm. *J. R. Statist. Soc.* A, **170**, 1061–1078.
Pickles, A. and Croudace, T. (2010) Latent mixture models for multivariate and longitudinal outcomes. *Statist. Meth. Med. Res.*, **19**, 271–289.
van de Pol, F. and Langeheine, R. (1990) Mixed Markov latent class models. *Sociol. Methodol.*, **20**, 213–247.
Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Statist. Ass.*, **66**, 846–850.
Reboussin, B. A. and Ialongo, N. S. (2010) Latent transition models with latent class predictors: attention deficit hyperactivity disorder subtypes and high school marijuana use. *J. R. Statist. Soc.* A, **173**, 145–164.

Roeder, K., Lynch, K. and Nagin, D. S. (1999) Modeling uncertainty in latent class membership: a case study in criminology. *J. Am. Statist. Ass.*, **94**, 766–767.
Ross, C. E. and Wu, C.-L. (1996) Education, age, and the cumulative advantage in health. *J. Hlth Socl Behav.*, **37**, 104–120.
Sacker, A., Worts, D. and McDonough, P. (2011) Social influences on trajectories of self-rated health: evidence from Britain, Germany, Denmark and the USA. *J. Epidem. Commty Hlth*, **65**, 130–136.
Savage, J. S. and Birch, L. L. (2010) Patterns of weight control strategies predict differences in women's 4-year weight gain. *Obesity*, **18**, 513–520.
Scott, M. A. (2011) Affinity models for career sequences. *Appl. Statist.*, **60**, 417–436.
Uebersax, J. (1999) Probit latent class analysis: conditional independence and conditional dependence models. *Appl. Psychol. Measmnt*, **23**, 283–297.
Vermunt, J. (2003) Multilevel latent class models. *Sociol. Methodol.*, **33**, 213–239.
Vermunt, J. K. (2008a) Latent class models in longitudinal research. In *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (ed. S. Menard), vol. 1, pp. 373–385. Burlington: Elsevier.
Vermunt, J. (2008b) Latent class and finite mixture models for multilevel data sets. *Statist. Meth. Med. Res.*, **17**, 33–51.
Ward, J. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Statist. Ass.*, **58**, 236–244.
Westerlund, H., Kivimäki, M., Singh-Manoux, A., Melchior, M., Ferrie, J. E., Pentti, J., Jokela, M., Leineweber, C., Goldberg, M., Zins, M. and Vahtera, J. (2009) Self-rated health before and after retirement in France (GAZEL): a cohort study. *Lancet*, **374**, 1889–1896.
Widmer, E. D. and Ritschard, G. (2009) The de-standardization of the life course: are men and women equal? *Adv. Lif. Course Res.*, **14**, 28–39.
Wilson, C. (2006) Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environ. Planng* A, **38**, 187–204.
Wu, L. (2000) Some comments on "Sequence analysis and optimal matching methods in sociology: review and prospect". *Sociol. Meth. Res.*, **29**, 41–64.