Dondena Working Papers

Carlo F. Dondena Centre for Research on Social Dynamics and Public Policy

Population Dynamics and Health

Comparing models for sequence data: prediction and dissimilarities

Raffaella Piccarreta
Marco Bonetti
Stefano Lombardi

Working Paper No. 113
January 2018

Università Bocconi • The Dondena Centre
Via Guglielmo Röntgen 1, 20136 Milan, Italy
http://www.dondena.unibocconi.it

ISSN-2035-2034

Comparing models for sequence data: prediction and dissimilarities

Raffaella Piccarreta
Dondena Centre for Research on Social Dynamics and Public Policy,
Bocconi Institute for Data Science and Analytics (BIDSA) and
Department of Decision Sciences, Bocconi University, Milan, Italy

Marco Bonetti*
Dondena Centre for Research on Social Dynamics and Public Policy and
Department of Policy Analysis and Public Management, Bocconi University, Milan, Italy

Stefano Lombardi
Department of Economics, Uppsala University, Uppsala, Sweden

Abstract

We consider the case when it is of interest to study the different states experienced over time by a set of subjects, focusing on the resulting trajectories as a whole rather than on the occurrence of specific events. Such situation occurs commonly in a variety of settings, for example in social and biomedical studies. Model-based approaches, such as multistate models or Hidden Markov models, are being used increasingly to analyze trajectories and to study their relationships with a set of explanatory variables. The different assumptions underlying different models typically make the comparison of their performances difficult. In this work we introduce a novel way to accomplish this task, based on microsimulation-based predictions. We discuss some criteria to evaluate one model and/or to compare competing models with respect to their ability to generate trajectories similar to the observed ones.

Keywords: Dissimilarity; Hidden Markov model; Interpoint distance distribution; Micro-simulation; Multi-state model; Optimal Matching; Sequence analysis.

Acknowledgements

We thank L. Cai for having made the SAS code to estimate the MSLT available, and for suggestions about its extension to the analysis of the FFS data. We are grateful to P. Tebaldi for having extended the Stata commands to allow for the direct input of the dissimilarity matrix to the Stata commands `mstat` and `mtest`.

Lastly, we thank the Central Commission for Statistics of Statistics Netherlands for having granted us permission to use the data on the Dutch Fertility and Family Surveys.

* Corresponding author: Marco Bonetti, Dept. of Policy Analysis and Public Management, Bocconi University, Via Roentgen 1, 20136 (MI), Milan, Italy. Tel +39 02 58365670, e-mail marco.bonetti@unibocconi.it

Comparing models for sequence data: prediction and dissimilarities

1. Introduction

We consider the case when for n subjects the activities (or states) experienced over a period of time are tracked, so that a trajectory, i.e. a finite *sequence* or ordered collection of states, is available for each subject. Typically, observation of such trajectories is right censored.

There are many applications when data of this type can be of interest. For example, in sociology, one may be interested in studying the transition to adulthood of individuals with respect to union and family formation, or to employment. In health studies, the conditions of individuals are typically observed over time; in each period, one records whether or not the patient experiences some focal event such as remission, occurrence of a disease, various degrees of severity of a disease, complications, or death. In

Objects of interest typically are the event of experiencing a state, the time of occurrence or transition into a state, or the length of the permanence in a state (see, e.g., Lawless, 2003).

The states that an individual may experience can take different forms. For instance, they can be *recurring* (when one state can be visited more than once), *transient* (when a subsequent transition to another state may occur), or *absorbing* (when no transitions to other states are possible after having experienced that state). Here we consider state trajectories that include all of these kinds.

Adopting a "holistic" approach, we will focus on the evolution of the trajectory as a whole, rather than just on the timing or occurrence of specific events. We are interested in studying the relationships that may exist between the trajectories and a set of covariates, a problem that is of course relevant in many contexts. There is increasing attention and interest in the literature on the use of model-based approached at this aim. For example, multi-state models are being used more and more to describe the occurrence of events of different kinds over time, and can prove useful to identify relevant covariates and/or to assess the effects of covariates on the evolution of the trajectories. For a review of multi-state models and their implementation one may refer to Putter et al. (2007) or Beyersmann et al. (2012), who consider models for the hazard of transitioning to specific states within the context of multi-state models with no recurrent events. The class of multi-state models is broad, and includes a number of popular models. For

example, the case of competing risks, focused on the transition from one initial state to several mutually exclusive absorbing states, is closely related to a multi-state model. Similarly, traditional survival analysis can be regarded as a special case of a multi-state model, with attention limited to only one absorbing state beyond the initial one.

Another popular approach is based on latent Markov models (see e.g. Vermunt et al., 2008; Bartolucci et al., 2013), which assume that underlying the observed trajectories there exists a latent (hidden) process, described by a Markov chain with a finite number of states. In such a framework, the latent process is indeed a multi-state model, and the observed states are treated as "realizations" of the latent ones, and they are assumed to be independent conditionally on the sequence of the hidden states.

Irrespective of the approach followed to study the trajectories' evolution, the assessment of the results and, possibly, model selection are usually based upon criteria (for example, AIC of BIC) that depend on the hypothesized data generating process. The different assumptions underlying different models typically make the comparison of their results difficult if not impossible.

A related, crucial issue is to evaluate the models' performance with respect to the *original* object of interest, which is the set of the observed trajectories.

Here, we propose to use simulated trajectories to study and compare the in-sample and the out-of-sample predictive power of competing models, that is their ability to generate trajectories that are "similar" to the observed ones. Our aim is to introduce criteria to suitably compare collections of dissimilarities computed across observed and model-generated sequences.

We explore a few distance-based methods to assess the relative merits of competing models, when applied to the same data. Specifically, we refer to data collected as part of the Fertility and Family Surveys (FFS), conducted in the 1990's in selected member States of the United Nations Economic Commission for Europe (UNECE, www.unece.org/pau/ffs/ffs.htm; Latten and De Graaf, 1997). In Section 2 we describe the data, with specific reference to the holistic approach to the analysis of life courses known in the literature as *Sequence Analysis* (SA). These data were analyzed in Lombardi (2012) and in Bonetti et al. (2013) using two alternative models for the probability of transitioning from one state to another while accounting for a set of covariates, namely the Multi-State Life Table (MSLT) approach (Cai et al. 2006, 2010) and the State Change model (SCM) (Bonetti et al., 2013). These two event history models are briefly described in Section 3. For the sake of completeness, in that section we also provide a basic illustration of latent Markov models, which may also be used to study trajectories.

In Section 4 we introduce some proposals to compare the performance of two or more competing models, and in Section 5 we illustrate their use to assess the relative goodness of fit through the predictive accuracy of models applied to the FFS data. For the sake of synthesis, we focus on the two models SCM and MSLT. We close with some comments in Section 6.

2. The Fertility and Family Survey data

We consider data arising from the Fertility and Family Surveys (FFS) study, conducted in the 1990s in selected member States of the United Nations Economic Commission for Europe (Latten and De Graaf, 1997). The same data were analyzed in Bonetti et al. (2013) and in Lombardi (2012), who focused on 1897 women from the Netherlands born between 1953 and 1962. In particular, the interest was on women's childbearing and family formation patterns and on their relationships with a set of baseline covariates. For each woman, the ordered collection of the monthly states experienced between 18 and 30 years of age was considered, summarized by a sequence $\boldsymbol{s} = (s_1, \dots, s_T)$, where T = 144 months for all women, $s_t \in \{1, \dots, M\}$ is the state visited at time t, and M is the number of possible different states.

Specifically, we consider the following states: living without a partner and having no children (**N**), married without children (**M**), in unmarried cohabitation without children (**U**), single with at least one child (**NC**), married with at least one child (**MC**), and cohabiting and having at least one child (**UC**). A compact representation of a woman's trajectory can be obtained by listing the visited states $\boldsymbol{v} = (v_1, \dots, v_h)$ (*states sequence*) and the durations $\boldsymbol{\tau} = (\tau_1, \dots, \tau_h)$ of the uninterrupted permanences in each state (*durations sequence*), with h indicating the observed total number of states visited. For example, for a woman who lived without a partner for 22 months, then cohabited for 27 months, then lived as single again for 31 months, and finally married and remained in that state for 64 more months, one has h = 4, $\boldsymbol{v}$ = (**N,U,N,M**), and $\boldsymbol{\tau}$ = (22,27,31,64). Although some states can (as in this example) be visited more than once, the "children" state is *absorbing*: after the first child is born, the woman cannot return to any of the "no children" states (deaths of children are not considered). Note that the last duration might be right-censored.

The focus is on the association between the sequences and a set of (baseline) categorical socio-demographic characteristics, namely birth cohort, level of education, religious status, and having versus not having separated or divorced parents. We also distinguish between the two birth cohorts 1953-1957 and 1958-1962. *Education* is based on the years of education received after the age of 15, and it groups the women into three classes: women who interrupted their studies,

those who proceeded with at most an additional 3 years of education, and those who received more than 3 years of additional education. *Religion* indicates whether a woman declared herself as being religious or not. Finally, *Divorce* indicates whether a woman's parents are separated or divorced (at the time of the survey). Note that the survey collected information on women at the time of the interview, i.e. after the age of 30. As a consequence, the use of *Divorce* to explain or predict sequences may be questionable. Given that most parental divorces take place during the children's adolescence, however, the assessment of the effect of parents' being divorced is not likely to be overly biased. The use of *Religion* may also be somewhat questionable because of the possibility of (rare) changes of religious status during one's life. On the other hand, *Education* is a clearly defined baseline variable, as it refers to events that occurred before the age of 18. For a detailed description of the FFS study and of its main findings we refer to Latten and De Graaf (1997).

3. The Study of states over time

A problem that typically arises when trying to explain and/or predict trajectories is that the frequency of each specific trajectory is typically very low. However, some trajectories will be similar. For instance, two trajectories may differ only by a slight misalignment of experienced event(s): the two trajectories would then have identical states sequences – for example, in the FFS study, $\boldsymbol{v_1}$**=** $\boldsymbol{v_2}$ = (**N,U,N,M**) – but slightly different durations, say $\boldsymbol{\tau_1}$ =(22,27,31,64) and $\boldsymbol{\tau_2}$ =(20,27,31,66). Or, they may differ by short spells in different states, so that they would be almost identical except for the presence of an additional state, as for $\boldsymbol{v_1}$=(**N,U,N,M**), $\boldsymbol{v_2}$=(**N,U,N,U,M**), $\boldsymbol{\tau_1}$ =(22,27,31,64), and $\boldsymbol{\tau_2}$ =(22,27,2,29,64). These ideas are at the basis of the techniques and algorithms used in Sequence Analysis, which is now an established approach to the description of life courses. SA focuses on criteria – generally based upon alignment algorithms – to measure the dissimilarity between pairs of sequences. Such dissimilarities can be used in a number of ways, and a typical application consists of the identification of clusters of individuals who experience similar if not identical trajectories.

Whilst SA aims primarily at identifying the most salient features of the observed trajectories, the study of the effects of possibly relevant covariates is clearly also quite important.

However, the analysis of the relationship between pairwise dissimilarities and covariates is very difficult, if not impossible, also due to the structural and constrained inter-relationships that exist among dissimilarities. Some proposals were introduced in the literature to draw conclusions

about the impact of covariates on the "structural features" of the trajectories, as described by clusters. Piccarreta and Billari (2007) extend the ANOVA concepts and the R^2 criterion to SA, to assess the extent to which a certain cluster solution accounts for the total sample-heterogeneity (see Section 4 for more details). Studer et al. (2011) use permutation tests to extend the ANOVA F-test along such direction. Although this approach allows concluding whether individuals with different levels of covariates experience "significantly" different trajectories, it does not provide any specific indication about the relationship between covariates and sequences. Some authors (see e.g., McVicar and Anyadike-Danes, 2001) propose to use multinomial logistic regression to relate the probability (or "risk") of experiencing trajectories in the different clusters to the explanatory variables, thus allowing one to gain more interpretable results. This approach provides reliable results only when the clusters are highly homogeneous. Clearly, within-clusters homogeneity can often be achieved only by increasing the number of clusters, that is the number of levels taken by the dependent variable in the multinomial model, with a consequent trade-off between the model's simplicity and its reliability.

The difficulties implied by the study of the trajectories-covariates relationships led some scholars to model the evolution of the transitions from one state to another in different periods, as well as the relationships between such transitions and the available explanatory variables. With no claim of being exhaustive, in the following we describe some models that may be used in this direction.

Here we are focusing on data that consist of state transitions in discrete time. Even if transitions occur continuously, when available data are interval censored a common approach is to treat time as discrete, provided that the time intervals are sufficiently narrow and that transitions are not too frequent, so that not too much information is lost. These two conditions are consistent with the assumptions that: (a) only one transition can happen within each time interval; and (b) at the beginning of each interval individuals are at risk of experiencing the allowed transitions, which may occur in correspondence of an unknown point of the time interval (Cai et al., 2010).

The standard approach to modelling the probability $P_{q\to r} = \Pr(s_{t+1} = r | s_t = q)$ of transitioning from state q to state r between times t and $(t+1)$ is the generalized multinomial logistic regression (see, e.g., Agresti, 2002):

$$\log\left\{\frac{P_{q\to r}(\mathbf{x}_t)}{P_{q\to M}(\mathbf{x}_t)}\right\} = \alpha_{qr} + \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\beta}_{qr} \quad q = 1, \dots, M; r = 1, \dots, (M-1) \tag{1}$$

where M is the reference state, $\mathbf{x}_t$ is the vector of explanatory variables at time t, and $\boldsymbol{\theta}_{qr} = (\alpha_{qr}, \boldsymbol{\beta}_{qr}^{\mathrm{T}})^{\mathrm{T}}$ is a vector parameter specific for each departure state, q, and for each of the $(M-1)$ arrival states. More parsimonious specifications can be obtained by constraining some parameters to be equal across some state transitions.

Note that in (1) the probabilities of transitioning across states depend on the past history only through the current state (and through the possibly time-varying covariate $\mathbf{x}_t$). In other words, the state visited after time *t* only depends on the state experienced at t. In some cases, it is more realistic to let the transition probabilities also depend on the *permanence* in the current state since the most recent entry in it. The resulting less restrictive models are referred to with different names, but they are all characterized by the *semi-Markov* property. Two specific such models were applied to the FFS data in Lombardi (2012) and in Bonetti et al. (2013).

The first model is the Multi-State Life Table (MSLT) model proposed by Cai et al. (2006, 2010). The probability of transitioning from state q to state r is modelled as:

$$\log\left\{\frac{P_{q\to r}(\mathbf{x}_t)}{P_{q\to M}(\mathbf{x}_t)}\right\} = \alpha_{qr} + \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\beta}_r + \gamma_r d_t \quad q = 1,\dots,M; r = 1,\dots,(M-1) \qquad (2)$$

where d_t is the time spent in the current state up to time t (since the most recent entrance into it). The duration effect γ_r is assumed to be specific for the different *arrival* states, and constant across the current states. If the current state is absorbing, then the probability of transitioning to another state is set equal to zero. Notice that the duration effect can enter the model through, say, polynomial terms, and that it is also possible to allow the duration effect to vary with one or more covariates by adding interaction terms.

Whilst MSLT adjusts for the time spent in the current state, it does not model durations directly. An alternative in this direction is the State Change Model (SCM) described in Bonetti et al. (2013). SCM separately models the time to the next *generic* transition, and the probability of transitioning to specific states conditionally on a transition occurring.

Regression models are built for the two components of the model: time-to-event regression models for the duration, and conditional multinomial regression models to relate the probabilities of transitioning to the different arrival states to a set of covariates and to the observed duration up to the transition.

For example, the (discrete) time to the next transition may be assumed to follow a geometric distribution with a parameter p that depends on the covariates through a logit link:

$$p(\mathbf{z}_j) = \exp\left(\mathbf{z}_j^{\mathrm{T}}\boldsymbol{\varphi}\right)\left[1 + \exp\left(\mathbf{z}_j^{\mathrm{T}}\boldsymbol{\varphi}\right)\right]^{-1}, \tag{3}$$

where $\mathbf{z}_j$ summarizes the information available when the j-th state is entered, and $\boldsymbol{\varphi}$ is the vector of parameters. The covariates in $\mathbf{z}_j$ can be time-varying, and $\mathbf{z}_j$ may also include the last state visited before the j-th transition.

As for the probability of transitioning to a specific state r, a variation of the generalized conditional multinomial regression models is used. Specifically, the probability of transitioning from state q to state r at the j-th transition, $P_{q\to r,j} = \Pr(v_{j+1} = r | v_j = q)$, is modelled as:

$$P_{q\to r,j}(\mathbf{x}_j) = \frac{\exp\left(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_{qr}\right)}{1 + \sum_{m=1,m\neq q}^{M-1}\exp\left(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_{qm}\right)} \quad q = 1,\dots,M; r = 1,\dots,(M-1);\ r \neq q \tag{4}$$

Note that $P_{q\to r,j}$ is now specific to the j-th transition and not to the time period t (as in Equation 2). Here $\mathbf{x}_j$ summarizes the information available at the j-th transition – which may or may not overlap with the set of covariates $\mathbf{z}_j$ in (3), and that will be allowed to include the permanence in the state visited prior to the transition (referred to as *Tval* in Section 5). The probability of transitioning into the same state, $P_{q\to q,j}$, is set to zero for every q and j. A possible drawback of SCM is a large number of parameters to estimate, and this can be mitigated by constraining some of them to be equal to zero. Relatedly, Bonetti et al. (2013) suggest a preliminary nonparametric screening procedure to select the most promising explanatory variables.

The main difference between MSLT and SCM is that while the former allows for an effective description of covariate effects on the transition probabilities, the latter allows for a direct interpretation of covariate effects on the time-to-event distribution of the time until the next transition.

Recently, attention has been also devoted to the use of latent Markov models for the analysis of sequence data. Here we briefly describe two such models, namely the Hidden Markov and the Mixture Hidden Markov models (HMMs and MHMMs, respectively), which admit as special cases a number of other specific models.

In a HMM, it is assumed that a latent (hidden) process exists, namely a first-order Markov chain with a finite number K of states, $\boldsymbol{\sigma} = (\sigma_1,\dots,\sigma_T)$, $\sigma_t \in (1,\dots,K)$. The evolution of the Markov chain is described by the vector of initial probabilities $\boldsymbol{\pi}_0 = (\pi_{01},\dots,\pi_{0K})^{\mathrm{T}}$, whose k-th element $\pi_{0k} = \Pr(\sigma_1 = k)$ is the probability to start the sequence with the k-th state, and by the matrix

of the transition probabilities across latent states. The (k, r)-th term of such matrix $\boldsymbol{\Pi}$ is the probability of transitioning from the k-th to the r-th state, $\pi_{k \to r} = \Pr(\sigma_t = l | \sigma_{t-1} = k)$, assumed to be constant over time (which results in a *homogeneous* HMM). By the first-order Markov assumption, $(\boldsymbol{\pi}_0, \boldsymbol{\Pi})$ fully characterize the distribution of $\boldsymbol{\sigma}$ over time. The sequence of the observed states, $\boldsymbol{s} = (s_1, \ldots, s_T)$, with $s_t \in (1, \ldots, M)$, is regarded as a realization of the unobservable underlying latent process. It is assumed that the state observed at a given time point only depends on the concomitant latent state and not on the past ones. The emission probabilities are arranged in a matrix, $\boldsymbol{\Psi}$, whose (k, m)-th element is the probability of emission of the m-th observed state from the k-th hidden state. The observed states are assumed to be independent conditionally on the sequence of the hidden states. Both baseline and time-varying covariates can be inserted in the model by assuming that the elements of $\boldsymbol{\Psi}$ and possibly those of $\boldsymbol{\pi}_0$ and $\boldsymbol{\Pi}$ depend (also) on covariates, so that the probability of an observed sequence $\boldsymbol{s}$ is obtained by integrating over the latent sequences as

$$\begin{aligned} \mathrm{P}(\boldsymbol{s}|\mathbf{x}) &= \sum_{\sigma_1=1}^{K} \sum_{\sigma_2=1}^{K} \ldots \sum_{\sigma_T=1}^{K} \mathrm{P}(\sigma_1, \ldots, \sigma_T|\mathbf{x}) \cdot \mathrm{P}(\boldsymbol{s}|\sigma_1, \ldots, \sigma_T, \mathbf{x}) \\ &= \sum_{\sigma_1=1}^{K} \sum_{\sigma_2=1}^{K} \ldots \sum_{\sigma_T=1}^{K} \left[\mathrm{P}(\sigma_1|\mathbf{x}_1) \prod_{t=2}^{T} \mathrm{P}(\sigma_t|\sigma_{t-1}, \mathbf{x}_t) \right] \cdot \left[\prod_{t=1}^{T} \mathrm{P}(s_t|\sigma_t, \mathbf{x}_t) \right], \end{aligned} \tag{5}$$

where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_t)$ denotes the set of baseline and time-varying covariates, respectively. Note that the introduction of time-varying covariates may move the latent process away from the first-order Markov assumption. As remarked in Vermunt et al. (2008), HMMs allow one to account for autocorrelation through the latent process, as well as for measurement error or misclassification through the imperfect relationship allowed between latent and observed states. Importantly, such models can be extended to the case when multiple sequences are observed for each statistical unit.

To also account for additional heterogeneity, mixture Hidden Markov models can be constructed, that include an additional latent variable indicating membership to one of several latent classes. Both the initial probabilities and the probabilities of transitioning from one latent state to another may be allowed to differ across latent classes.

The models described above can all be estimated via likelihood or Bayesian inference, and they have different features. The event history models, MSLT and SCM, have the advantage of

accounting for the permanences in the states. Nonetheless, in their original formulation they cannot be extended to the case when multiple sequences per individual are considered (which is however not the case in our illustration). In addition, they do not allow for the possible partitioning of cases in groups not defined by observed covariates. On the other hand, HMMs and MHMMs introduce an additional layer of complexity that makes the covariate effects more difficult to interpret. Such additional layer may also contribute to making the parameters weakly identified when data are sparse.

Our goal here is not to discuss the merits of a specific model or to contrast alternative proposals from a theoretical point of view. Indeed, our main point is that many possible approaches can be followed and different models can be built, each with desirable characteristics and possible drawbacks. The relevant issue is that since alternative models generally rely upon different assumptions, a fair comparison of their performance is not easy, particularly with respect to their ability to reproduce and/or predict the complete trajectories. This calls for the use of tools and criteria to assess the "holistic" quality of the models.

Typically, it is not possible to obtain a univocally defined *prediction* of the entire trajectory based on the estimated models' parameters. However, the models usually allow one to generate, via micro-simulation, event histories based on the estimated parameters, and conditionally on the covariates' values. Our proposal is therefore to evaluate models by comparing these simulated trajectories with the observed ones, using the methods described in the next section.

4. Comparing model performances: some distance-based proposals

In this section, we offer some criteria to compare the performance of two competing models with respect to their ability to reproduce or to predict sequences.

Consider the generation via simulation of one or more trajectories for each statistical unit from a fitted model. Indeed, it seems useful to try and take into account the variability of the trajectories generated for each subject by generating more sequences, say G, for each individual. In the following, we indicate by $\mathbf{S} = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_n)$ the set of the n observed sequences, by $\widehat{\boldsymbol{S}}_i = (\hat{\boldsymbol{s}}_{i1}, \dots, \hat{\boldsymbol{s}}_{iG})$ the set of G sequences generated from a fitted model for the i-th subject, and by $\hat{\mathbf{S}} = (\widehat{\boldsymbol{S}}_1, \dots, \widehat{\boldsymbol{S}}_n)$ the whole set of the nG generated sequences.

In the following, we describe some criteria that can be used to compare observed and predicted sequences. This points to the problem of assessing a model's *goodness of fit*. Nonetheless, it is also possible to design simulations to evaluate the *predictive ability* of a model. Indeed, one may

split the sample into a training set and a validation set, and use only the former to fit the model. Based on the estimated model, sequences can then be generated for each case in the validation set.

We will evaluate a model's goodness of fit (or predictive ability) by assessing the *similarity* between the observed and the predicted trajectories.

A first step along this direction is to compare the features of the observed and predicted sequences are with respect to the observed frequencies and durations of states in the two groups. This may be done qualitatively, and possibly conditionally on the values of the explanatory variables or marginally for a specific level of just one of the covariates.

A more in-depth analysis can be based on assessing the "error" incurred when using a model to predict/explain sequences. This can be done by evaluating, for each unit, how similar the corresponding G generated trajectories are to the observed one. To do so, it is necessary to properly define a measure of the dissimilarity between two sequences. This is a standard problem in the area of Sequence Analysis, and a variety of proposals have been introduced, which are extensively discussed and compared elsewhere (see e.g., Studer and Ritschard, 2016). Clearly, the dissimilarity criterion should be chosen accurately, and sensitivity analyses may be performed to assess the extent to which the choice of the dissimilarity measure affects the goodness of fit assessment.

The dissimilarities between the sequence observed for a given case and the corresponding set of generated sequences can be summarized at the individual level by defining, for example the quantity

$$\hat{\varepsilon}_i = \frac{1}{G}\sum_{g=1}^{G} \Delta(\boldsymbol{s}_i, \hat{\boldsymbol{s}}_{ig}) \tag{6}$$

where $\Delta(\boldsymbol{s}, \boldsymbol{w})$ denotes the dissimilarity between the two sequences $\boldsymbol{s}$ and $\boldsymbol{w}$. Note that the collection of dissimilarities $\Delta(\boldsymbol{s}_i, \hat{\boldsymbol{s}}_{ig})$, for a given observed sequence $\boldsymbol{s}_i$ and for $g = 1, \dots, G$, depends both on the heterogeneity of the generated sequences and on the position of $\boldsymbol{s}_i$ relatively to them. One may therefore choose to alternatively consider $\tilde{\varepsilon}_i = \Delta(\boldsymbol{s}_i, \tilde{s}_i)$, that is the individual dissimilarity between $\boldsymbol{s}_i$ and a *summary* of the generated sequences, $\tilde{s}_i$, such as, for example their *medoid.* The medoid of a set of sequences is the sequence with the smallest average dissimilarity from all the others in the group, and it is usually regarded as a reliable synthesis (see Sheikh et al., 2007; Aassve et al., 2007).

Two models can be compared based on such prediction errors – $\hat{\varepsilon}_i$ and/or $\tilde{\varepsilon}_i$ – typically through summaries of their distributions (i.e., the mean, the standard deviation, the range), possibly conditionally on the covariates' levels.

Alternative approaches can be envisioned to exploit the dissimilarities. Indeed, information is available on the conditional *heterogeneity* both of the original and of the generated sequences, and one may wish to evaluate if and to what extent the two *sets* of sequences are similar.

To do so, one may refer to criteria used in the context of cluster analysis to measure the dissimilarity between two *clusters*. For example, one could consider the *average-linkage*, i.e. the average of *all* the pairwise dissimilarities between sequences in the two groups of observed and generated sequences, or *Ward's-linkage*, that is the difference between the heterogeneity within the whole set of sequences (observed and generated) and the heterogeneities within the groups of sequences.

In particular, Ward's criterion is strongly connected to ANOVA-like measures, as discussed in Piccarreta and Billari (2007) and in Studer et al. (2011). Specifically, consider two groups of cases, C_1 and C_2, with sizes n_1 and n_2 respectively (in our case the group of observed and generated sequences). Extending the notions of *total sum of squares*, T, and *within sum of squares*, W, to the case when only pairwise dissimilarities are available, Piccarreta and Billari (2007) consider:

$$\mathrm{T} = \frac{1}{2(n_1+n_2)} \sum_{i,\ell \in (C_1 \cup C_2)} \delta_{i\ell}^2$$

$$\mathrm{W} = \mathrm{W}_1 + \mathrm{W}_2 = \frac{1}{2n_1} \sum_{i,\ell \in C_1} \delta_{i\ell}^2 + \frac{1}{2n_2} \sum_{i,\ell \in C_2} \delta_{i\ell}^2,$$

where $\delta_{i\ell} = \Delta(\boldsymbol{s}_i, \boldsymbol{s}_\ell)$ is the dissimilarity between two sequences. Since 0 ≤ W ≤ T, the well-known R-square and $\mathcal{F}$-statistic can be easily extended to this more general case. When comparing two groups one obtains:

$$R^2 = 1 - \frac{\mathrm{W}}{\mathrm{T}}$$

$$\mathcal{F} = \frac{(\mathrm{T} - \mathrm{W})}{\mathrm{W}/(n-2)}. \qquad (7)$$

The distribution of $\mathcal{F}$ is not known, and permutation tests (see, e.g., Edington and Onghena, 2007; Pesarin 2001) can be used to verify the differences between dissimilarities in the two groups. Studer at al. (2011) study and discuss in depth this extension of the $\mathcal{F}$-test to the

dissimilarity case. In particular, they conclude that when the chosen dissimilarity criterion cannot be regarded as a Euclidean distance, then it is convenient to refer to unsquared rather than squared dissimilarities in the expressions above.

To generalize these ideas, we suggest two additional criteria that refer to the study of the *complete distributions of dissimilarities.* Indeed, a first possibility is to consider the entire distributions of between-sequence dissimilarities, also called *interpoint distance distributions* (IDDs). This approach is based on the estimated cumulative distribution function of the dissimilarities between sequences within a group, or across two groups. Specifically, consider a distribution $\mathcal{F}_{\boldsymbol{Y}}$ taking values in a possibly highly dimensional (as in this case) space $\mathcal{Y}$. Define the random variable $D = \Delta(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ as the dissimilarity between two *i.i.d.* elements $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ extracted from $\mathcal{F}_{\boldsymbol{Y}}$. The IDD is the distribution $\mathcal{F}_D(d) = P(D \leq d)$ of D (see, e.g., Bonetti, 2016). D indicates the "distance" between two randomly selected observations as measured by any symmetric (non-negative) function of the two observations, and in particular by any dissimilarity measure that may be relevant for the problem at hand. To estimate $\mathcal{F}_D(d)$, an *i.i.d.* sample, $(\mathbf{y}_1, \dots \mathbf{y}_n)$ can be drawn from $\mathcal{F}_{\boldsymbol{Y}}$, and inference can be based upon the set of the $\binom{n}{2}$ pairwise (dependent) dissimilarities between them. In particular, Bonetti and Pagano (2005) consider the empirical cumulative density function (ECDF):

$$F_n(d) = \frac{2}{n(n-1)} \sum_{1 \leq i < \ell \leq n} 1(\delta_{i\ell} \leq d),$$

where $\delta_{i\ell}$ indicates the distance or dissimilarity between the i-th and the ℓ-th sample observations $\mathbf{y}_i$ and $\mathbf{y}_\ell$. The ECDF of all the pairwise distances, evaluated at a finite number of values along the distance axis, has an asymptotic multivariate normal distribution. Indeed, if one considers a grid of bins-defining points $d_1, \dots, d_B$ along the distance axis, and the vector of the ECDF is evaluated at the end of each bin, then such vector may be written as $F_n(\mathbf{d}) = [F_n(d_1), \dots, F_n(d_B)]$. In particular, the comparison between $F_n(\mathbf{d})$ and a null hypothesis distribution for D (say, $F_0(\mathbf{d})$) can be based on the statistic:

$$\mathcal{M} = [F_n(\mathbf{d}) - F_0(\mathbf{d})]^{\mathrm{T}} \hat{\Sigma}^{-} [F_n(\mathbf{d}) - F_0(\mathbf{d})],$$

where $\hat{\Sigma}^{-}$ is a generalized inverse of the estimated variance-covariance matrix of $F_n(\mathbf{d})$. This statistic can be regarded as a Mahalanobis distance between the observed and the expected distribution of the distances discretized to the B bins. While $\mathcal{M}$ converges in distribution to a

chi-squared random variable as n tends to infinity, experience shows that convergence is slow. For this reason, it is often preferable to use empirical testing routines, such as Monte Carlo or permutation testing (Bonetti and Pagano, 2005).

Manjourides (2009) extends this approach to the two-samples case, that is the situation when one wants to test whether two groups of multivariate observations follow the same distribution by comparing their IDDs. This extension is most relevant here. For two groups C_1 and C_2 (of n_1 and n_2 cases respectively), let $F_n^{(c)}(\mathbf{d}) = [F_n^{(c)}(d_1), \ldots, F_n^{(c)}(d_B)]$, where $F_n^{(c)}(d)$ is the ECDF computed using only the subjects in the c-th group , with c=1, 2. The statistic to test the null hypothesis that the distribution of the distances is the same in the two groups is:

$$\widetilde{\mathcal{M}} = \left[F_n^{(1)}(\mathbf{d}) - F_n^{(2)}(\mathbf{d})\right]^{\mathrm{T}} \hat{\Sigma}^{-} \left[F_n^{(1)}(\mathbf{d}) - F_n^{(2)}(\mathbf{d})\right] \tag{8}$$

where $\hat{\Sigma}^{-}$ is the Moore-Penrose generalized inverse of the estimated variance-covariance matrix $\hat{\Sigma}$ of the vector $[F_n^{(1)}(\mathbf{d}) - F_n^{(2)}(\mathbf{d})]$. Inference can be based on the permutation distribution obtained by repeatedly permuting the group labels of the observations.

Another possibility consists of the application of a Wilcoxon-like test, as first suggested in Mosler (2002). The idea is to contrast the within-group dissimilarities (relative to sequences in the same group) to the between-group dissimilarities (computed between sequences belonging to different groups) with a rank-based test statistic. For the two groups C_1 and C_2, let $\delta_{i\ell}^{(c)}$ (with c=1, 2) denote the *intra*-sample dissimilarity calculated between two cases in the same group, and $\delta_{i\ell}^{(1,2)}$, with $i \in C_1$ and $\ell \in C_2$, be the *inter*-sample dissimilarity calculated for cases belonging to different groups. After sorting the set that includes *all* the $\binom{n_1}{2} + \binom{n_2}{2}$ *intra*-sample and *all* the $n_1 n_2$ inter-sample dissimilarities in ascending order, and after assigning ranks to these dissimilarities, the test statistic

$$\mathcal{T} = \sum_{i \in C_1} \sum_{\ell \in C_2} R\left(\delta_{i\ell}^{(1,2)}\right) \tag{9}$$

can be defined, where $R(\delta_{i\ell}^{(1,2)})$ denotes the rank of the dissimilarity $\delta_{i\ell}^{(1,2)}$. If both samples come from the same distribution, then the generic inter-sample distance $\delta_{i\ell}^{(1,2)}$ should follow the same marginal distribution as the generic intra-sample distances, $\delta_{i\ell}^{(1)}$ or $\delta_{i\ell}^{(2)}$. Mosler (2002) suggested to reject the null hypothesis when $\mathcal{T}$ is large. He derived the first two moments of $\mathcal{T}$, and

proposed an exact permutation-based non-parametric approach to testing with $\mathcal{T}$, which is also based on permuting the group labels of the observations.

The $\mathcal{F}$, $\widetilde{\mathcal{M}}$, and $\mathcal{T}$ statistics can all be used to compare the dissimilarities of the observed (in-sample or out-of-sample) and generated sequences (or of sequences generated using competing models).

Since the performance of a model may differ across the covariate space, it might be also interesting to evaluate the extent of dissimilarity between observed and generated sequences corresponding to specific combinations of covariate values –with high enough frequencies – thus keeping the assessments separated rather than pooled into an overall measure across covariates' values. Observe that the calculation of all the dissimilarities involved in the definition of the statistics can make their calculation cumbersome when the number of simulated sequences is high. Interestingly, with respect to this point one should consider that since no model truly holds, as the number of generated sequences increases one should possibly expect shrinking p-values for all tests when they are applied to the trajectories generated by *any* model.

It is also important to stress that when covariates are continuous (or discrete with many values) there may exist just one subject corresponding to a given covariate value. In that case, the only available dissimilarities for that covariate level become those between the observed trajectory and the generated trajectories. A possible summary of such a set of dissimilarities is then the individual prediction error, $\hat{\varepsilon}_i$ or $\tilde{\varepsilon}_i$, introduced before.

5. Results: Application to the FFS data

We refer to the criteria described in the previous section to compare models applied to the FFS data. Since our methods apply to any model that can produce trajectories, for the sake of simplicity below we focus our illustration on the two simpler models, MSLT and SCM, illustrated in Section 3.

In both cases, the sparseness of the data (within combinations of covariate values) did not allow fitting the models with the various "Children" states. Therefore, the three states **NC**, **MC**, and **NC**, were grouped into a unique absorbing state "**C**". Also, since transitions from states with children to states without children are not allowed, all the parameters regarding these transitions were set to zero. In the following, attention will be focused on the $n = 1852$ women whose initial state was different from "**C**."

The marginal frequencies of all transitions in the observed data are shown in Table 1.

The explanatory variables were *Cohort* (binary variable indicating whether a woman is in the younger cohort, 1958-1962), *Education* (coded with two binary variables, *Educ2* and *Educ3*, indicating 0-3 years and more than 3 years of additional education after the age of 15), *Religion*, and *Divorce*, as described in Section 2.

Table 1. Frequencies of transitions from row state to column state in the FFS data set

	N	**U**	**M**	**C**
N	0	912	920	43
U	178	0	554	48
M	32	8	0	1140

The transition from one state to another at a given moment was allowed to depend on the previously visited state, on the *Age* at the time of the transition (second-order polynomial), and on the time spent in the state visited before the transition (*TVal*). Note that the latter covariates change at each visited state.

Table A1 in Appendix A reports the maximum likelihood estimates for the MSLT model, whereas Tables A2 and A3 report results obtained for the duration and for the transition parts of the SCM model. In both cases, only the variables which turned out to be significant for at least one of the conditional transition probabilities are reported.

In particular, for the MSLT model a Wald backward elimination procedure led to selecting the entire set of regressors, including Age^2 and the duration *Tval*. For the SCM, the results of the variable selection procedure yielded for the duration part of the model the covariates *Age*, *Previous state*, and *Cohort*. For the transition component of the SCM the following covariates were significant for at least some of the conditional transitional probabilities: *Tval* (Time spent in the previous state), *Age*, *Education*, *Religion*, *Divorce* and *Cohort*. Note that the overall interpretation of the effects of the two time-varying covariates *Tval* and *Age* is rather complicated, since they enter both components of the SCM.

Both SCM and MSLT can be used to describe the trajectories' generating mechanism. We now contrast their performances by comparing the observed sequences with those generated from the two models by plugging in the estimated parameter values (reported in Tables A1, A2, and A3 in Appendix A). For both models, starting from one initial state and conditionally on the baseline covariate values, we sampled the subsequent evolution of the trajectory from the

estimated probability distributions implied by each model. This is further explained in Bonetti et al. (2013) and in Cai et al. (2010). For simplicity, we did not account for the parameter estimators' sampling variability.

For each observed sequence we generated $G = 100$ trajectories based on the individual's covariate values. The resulting sets of simulations will be indicated below as $\hat{\mathbf{S}}^{(\mathrm{SCM})}$ and $\hat{\mathbf{S}}^{(\mathrm{MSLT})}$.

To define the pairwise dissimilarity measure we used Optimal Matching (OM), an alignment technique which was originally introduced in molecular biology to study protein or DNA sequences (Sankoff and Kruskal, 1983), and which was later extended to the study of life courses in Sociology (Abbott, 1995). In OM attention is focused on the quantification of the *effort* needed to transform one sequence into another. Three elementary transformation operations are taken into account: (i) insertion of a state; (ii) deletion of a state; (iii) substitution of a state with another. Each operation is assigned a cost, and the dissimilarity is defined as the minimum total transformation cost from one sequence to the other. Substitution costs may be assigned subjectively on the basis of a priori knowledge or considerations (see, among the others, McVicar and Anyadike-Danes, 2002). Alternatively, one may follow a data-driven approach and relate substitution costs to transition frequencies, so that frequent transitions are less costly than rare transitions (Rohwer and Pötter, 2004). Even if sometimes criticized (see Aisenbrey and Fasang 2010, for an in-depth review of the most relevant criticisms and of alternative proposals), OM remains the criterion most widely used to measure dissimilarity between trajectories in Sequence Analysis. Following a standard approach, we set the insertion and the deletion costs equal to 1, while we chose the substitution cost between two states to be inversely related to the transition's frequency (Rohwer and Pötter, 2004).

We start by comparing the features of the observed sequences ($\mathbf{S}$) with those of the simulated ones ($\hat{\mathbf{S}}^{(\mathrm{SCM})}$ and $\hat{\mathbf{S}}^{(\mathrm{MSLT})}$). Specifically, we focus on the state sequences ($\boldsymbol{v}$) in the three sets and compare the frequencies of the most frequent state sequences, the distribution of the visited states (irrespective of their duration), and the average duration of each visited state. Results are shown in Figure 1.

Overall, 5732 states were visited (recall that one state can be visited more than once). The most visited state is **N**, followed by **M** and **C**, and a similar order is also observed for the states' durations. The traditional family formation pattern, (**N**,**M**,**C**), is the most frequent state sequence, even if a relatively high proportion of women experienced cohabitation before marriage. Moving to the simulated sequences, the MSLT-based sequences appear to be more

similar to the observed ones than the SCM-based sequences. Actually, in the latter case, the frequency and the average duration of the state **N** are larger than for the observed ones, and the reverse holds for state **C**. As for the visited states, the most frequent observed sequence is also the most frequent simulated sequence, (**N**,**M,C**). Nonetheless, in the simulated sets some sequences have a different relevance compared to the sample. For example, SCM overestimates the relevance of the sequence **N**, and in general of the sequences including **N**, whereas it underestimates the frequency of the sequences including state **C**. As for MSLT, it slightly overestimates the relevance of the sequences containing state **U**.

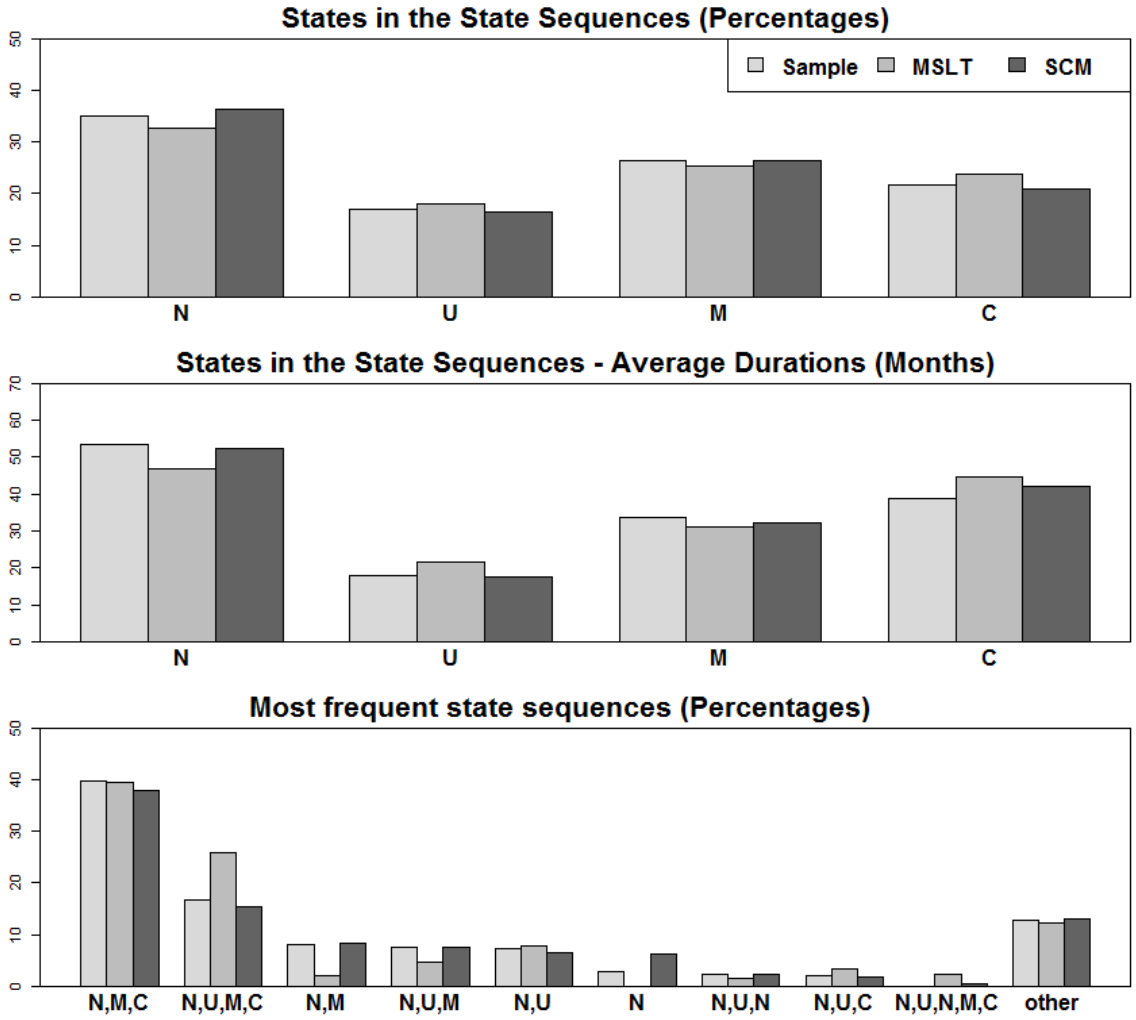


Figure 1. Description of sequences in the FFS dataset and in the two simulated datasets

It can also be useful to consider the plot of the transversal state distributions. Figure 2 shows the distribution of the states for each of the 144 months of observation (note that these plots do not

describe the transitions from one state to another). The previous considerations are confirmed: MSLT better resembles the distributions of the states observed in the sample, whereas for SCM one may observe an overestimation of the relevance of state **N**, and some underestimation of state **C.**

We now move to quantifying the differences in performance of the two models using the dissimilarities-based criteria described in the previous section. We start by considering the individual prediction errors, $\hat{\varepsilon}_i$ and $\tilde{\varepsilon}_i$. To compare MSLT and SMC with respect to these quantities, in Figure 3 we report the scatterplots of the errors for the two models, together with selected summaries. From the output it can be noted that SMC appears to perform slightly worse than MSLT, even if the differences are not so dramatic, particularly as concerns the distance from the medoid of the simulated sequences. Note that one may also monitor the behavior of the prediction errors for different levels of covariates, or of their combinations, so as to determine whether the model/s have specific prediction problems corresponding to specific input values (see, e.g., Figure B1 in Appendix B, showing the distributions of the errors corresponding to different covariate levels).

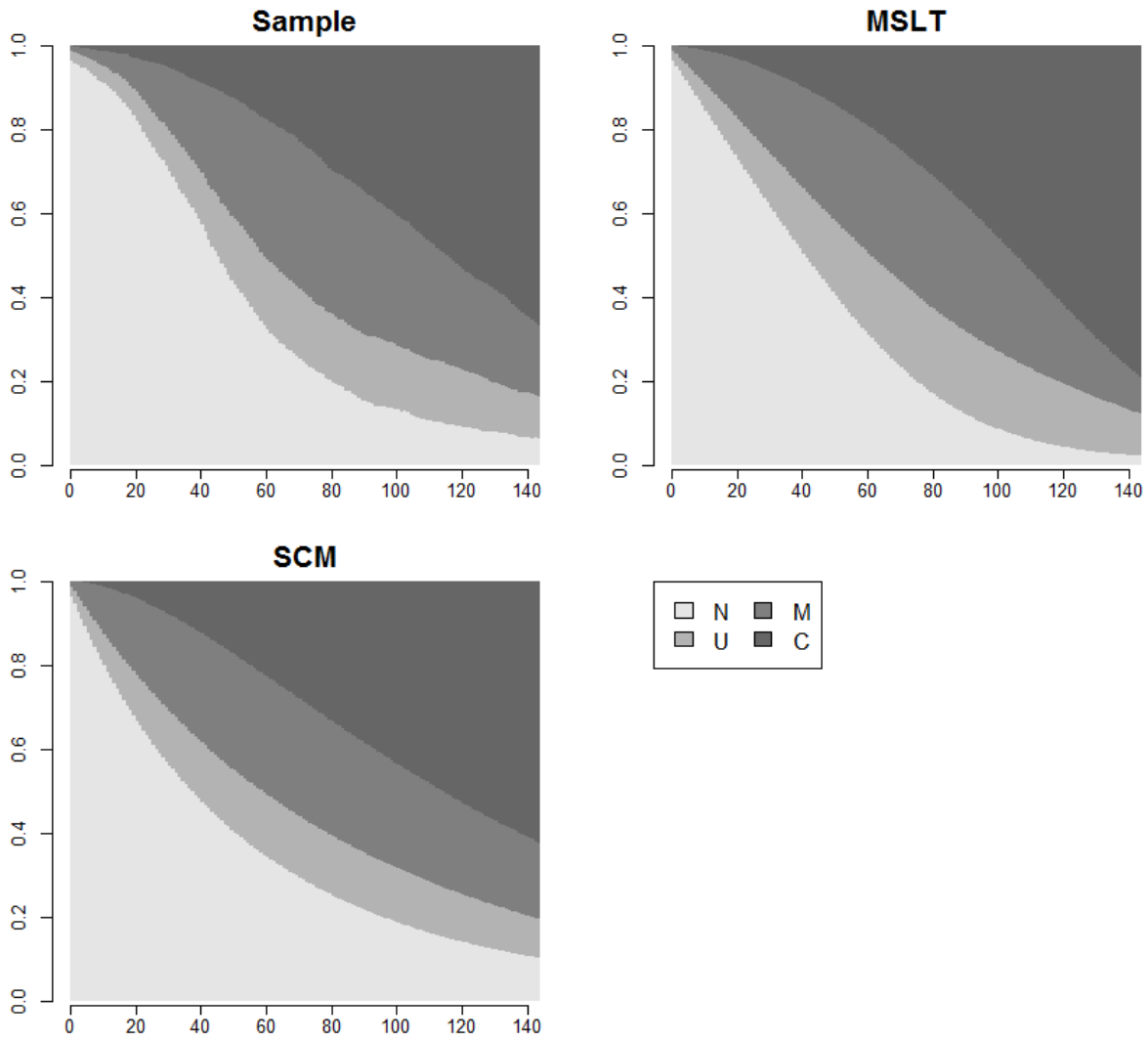
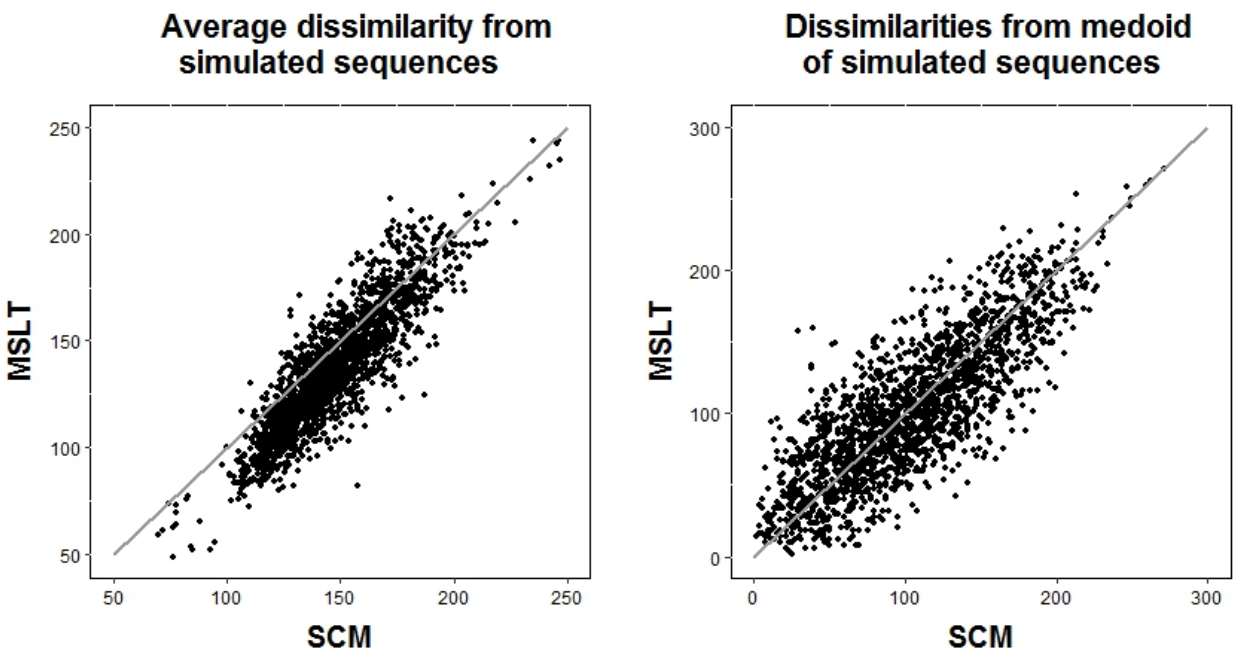


Figure 2. Transversal state distributions in the FFS dataset and in the two simulated datasets



	SCM	MSLT
Min	70.031	48.93
Mean	147.223	136.217
Std.Dev	23.421	28.882
Max	246.737	243.767

	SCM	MSLT
Min	2	2
Mean	105.76	103.627
Std.Dev	50.618	48.911
Max	270.772	270.663

Figure 3. Comparison and summaries of the individual prediction errors $\hat{\varepsilon}_i$, the average dissimilarity of each observed sequence from the simulated ones corresponding to it (left panel) and $\tilde{\varepsilon}_i$, the dissimilarity of each observed sequence from the medoid of the simulated sequences (right panel).

Monitoring and evaluating the behavior of individual predictions offers insights about possible extreme differences, and it allows one to draw conclusions about the models' performances at the *descriptive level*. When, as in our case, all the covariates are discrete, a more in depth analysis can be conducted, aiming at testing the extent of dissimilarity between all the observed sequences characterized by specific combinations of covariates and the corresponding sets of simulated sequences. Clearly, this procedure is only suitable when the frequency of the combination of covariates' levels is reasonably high, so that inferential results may have sufficient power. Table 2 reports the 5 most frequent combinations of covariate values in our dataset, characterizing a total of 1175 individuals (63.4% of the sample size).

Table 2. Most frequent combinations of covariates values

$\mathbf{x}$	Cohort	Education	Religion	Divorce	Initial state	Frequency
$\mathbf{x}_1$	53-57	0-3 Yrs	Yes	No	**N**	227
$\mathbf{x}_2$	53-57	>3 Yrs	Yes	No	**N**	247
$\mathbf{x}_3$	58-62	0-3 Yrs	Yes	No	**N**	195
$\mathbf{x}_4$	58-62	>3 Yrs	No	No	**N**	177
$\mathbf{x}_5$	58-62	>3 Yrs	Yes	No	**N**	329

Given the relatively large number of sequences available for each selected combination of covariates values, for each combination $\mathbf{x}$ of covariate values in Table 2 we chose a number of generated sequences coinciding with the number of observed frequencies (reported in the last column of Table 2), and compared the observed sequences $\mathbf{S}(\mathbf{x})$ with the model-based simulated sequences $\hat{\mathbf{S}}^{(\mathrm{SCM})}(\mathbf{x})$ and $\hat{\mathbf{S}}^{(\mathrm{MSLT})}(\mathbf{x})$.

We assess the significance of the observed differences by using the statistics $\mathcal{F}$, $\widetilde{\mathcal{M}}$, and $\mathcal{T}$ introduced in Section 4. Following the considerations in Studer et al. (2011), the $\mathcal{F}$ statistic (Equation 7) was calculated based on the dissimilarities rather than on their squared values. As for the $\widetilde{\mathcal{M}}$ statistic, the interpoint distance distributions (IDDs) in each set were estimated using $B = 20$ bins. To estimate the SCM and to generate sequences from it we used the **R** programming language. OMA and sequence analysis were applied using package `TraMineR` in R (Gabadinho et al., 2011). For the calculation of the $\widetilde{\mathcal{M}}$ statistic, the `Stata` functions `mstat` and `mtest` were used (Tebaldi et al., 2011). The p-values characterizing the test statistics, based on 1000 permuted samples, are reported in Table 3.

The results in the table confirm that the trajectories generated using the MSLT model tend to be more similar to the trajectories in the data when compared to the trajectories generated by the SCM model. Similar conclusions can also be drawn when the Mosler-Wilcoxon $\mathcal{T}$ test is used. Those results are reported in Table 3, which shows that the $\mathcal{F}$ statistic is highly sensitive, and leads to rejection in all cases. As such, this example suggests that it might be less suitable for goodness-of-fit assessments (when one is interested in comparing generated sequences to observed sequences) than for model building – when one is interested specifically in comparing the trajectories generated by two (possibly nested) models.

Table 3 also shows that the $\widetilde{\mathcal{M}}$ and the $\mathcal{T}$ test produce somewhat complementary results, since they return (non) rejections and different assessments of the evidence against the null hypothesis for different sets of covariate values.

It is interesting to observe that the MSLT model uses parameters specific for the different arrival states, whereas the SCM model uses parameters specific both for the arrival and for the departure states. In general, one might wonder about the reasons for the better fit of the first model (at least with respect to some covariate levels). This might be due to the different use of the durations in the two models. In the MSLT these affect the transitions probabilities at each time point. Instead, the SCM assumes a simple geometric regression model for the time-to-next transition, and includes the duration in the part of the model that refers to the conditional transition probabilities at the time when a transition occurs. Even if this is outside the scope of this paper, this suggests that evaluating the goodness of fit of the fitted models with our methods can also be useful to identify possible directions to modify the models to improve their performance.

Table 3. Tests on the differences between observed and model-based sequences and dissimilarities

Discrepancy analysis: $\mathcal{F}$ statistic test results (two-sided permutation p-values)

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{S}(\mathbf{x})$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{SCM})}(\mathbf{x})$	0.001	0.001	0.001	0.001	0.001
$\mathbf{S}(\mathbf{x})$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{MSLT})}(\mathbf{x})$	0.032	0.001	0.023	0.003	0.001
$\hat{\mathbf{S}}^{(\mathrm{MSLT})}$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{SCM})}(\mathbf{x})$	0.001	0.001	0.001	0.001	0.001

Interpoint distances: $\widetilde{\mathcal{M}}$ statistic test results (two-sided permutation p-values)

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{S}(\mathbf{x})$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{SCM})}(\mathbf{x})$	0.0005	0.0049	0.0006	0.0114	<0.0001
$\mathbf{S}(\mathbf{x})$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{MSLT})}(\mathbf{x})$	0.0008	<0.0001	0.2969	0.1874	0.0004
$\hat{\mathbf{S}}^{(\mathrm{MSLT})}$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{SCM})}(\mathbf{x})$	<0.0001	<0.0001	<0.0001	0.0270	<0.0001

Mosler-Wilcoxon $\mathcal{T}$ test results (two-sided permutation p-values)

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{S}(\mathbf{x})$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{SCM})}(\mathbf{x})$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\mathbf{S}(\mathbf{x})$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{MSLT})}(\mathbf{x})$	0.164	0.002	0.136	0.004	<0.0001
$\hat{\mathbf{S}}^{(\mathrm{MSLT})}$ **vs.** $\hat{\mathbf{S}}^{(\mathrm{SCM})}(\mathbf{x})$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

6. Discussion

Predicting trajectories in their entirety is a difficult task. It requires accounting for the visited states, their durations, and their ordering, and it is further complicated by the fact that

trajectories are generally all different one from another, even if possibly only by a slight extent. In addition, the adoption of a holistic perspective does not allow including time-varying covariates among the explanatory variables. This has led to an increasing interest towards models focused on specific aspects of the trajectories, for example transitions and/or durations, and based upon simplifying assumptions (for example, the first-order Markov property).

We have suggested the use of criteria to understand and to assess whether the simplified and more readable/interpretable structure arising from such models satisfactorily and properly describes the primary object of interest, that is the observed trajectories (as it is implicitly assumed). Such criteria are also useful to compare competing models. Indeed, the assessment of the results and, possibly, model selection are usually based upon criteria (for example, AIC or BIC) that depend on the assumed data generating process, and the different assumptions underlying different models typically make the comparison of their results difficult.

We have introduced two measures of the prediction error incurred for each specific subject. Such errors can be analyzed graphically, or summarized to obtain a global indicator of performance, possibly conditionally on specific covariate values, or of their combinations. We have then discussed three criteria to compare and test the difference between sequences generated by the models and the observed data, within a combination of goodness of fit and prediction ideas. The methods may indeed be seen either as part of a strategy for model selection, or as pure prediction comparison tools.

In the first case, model-generated trajectories are compared to the trajectories that were used to estimate the models' parameters. Since the adopted permutation-based inference does not take that into account, the resulting p-values can be interpreted both as a metric for goodness of fit and as formal tests of hypotheses. Indeed, they can provide a relative assessment of the goodness of fit of competing models, and serve as a guide in the model selection process when non-nested models are used. Our illustration here was based on this in-sample prediction power approach, and the apparent ability of the proposed distance methods to detect differences between the two models in the FFS data seems quite encouraging. In the second case, the competing models may be used to produce trajectories that are compared to observed trajectories that were not used to estimate the models' parameters.

The proposed procedures could be used to compare the whole sample to the generated sequences. Here, we chose to keep the assessments separate rather than pooled across combinations of (baseline) discrete covariate values with relatively high frequency. The

comparison of the predictive power of two models for covariate values with low or unit frequency requires particular care. Indeed, for such cases one would reasonably want to generate a number of sequences larger than the number of observed sequences.

In general, the p-values of the comparison criteria will depend on such number of simulated sequences. Since no model truly holds, as the number of simulated sequences increases, one may expect the tests to become more likely to reject. This would not allow a good comparison of the performance of the models. Hence, the number of simulated sequences could also be used as a tuning parameter, that may be changed to allow differences between the performances of different models to emerge, since the goal is to assess the relative predictive performance of the models being compared.

As an open issue, one should consider that the proposed criteria provide insights about the possible discrepancy between observed and simulated sequences, but they do not allow a substantive "interpretation" or understanding of the possible prediction errors/problems, since they do not describe in what respects the model-generated sequences differ from the observed ones.

This can be a problem because, clearly, specific prediction errors might be more serious than others. It is true, however, that the measures of dissimilarity used in Sequence Analysis typically penalize some differences more than others, so that hopefully the "most serious" substantial deviations should be emphasized by a relatively large dissimilarity through the very definition of the dissimilarity used.

Relatedly, the proposed measures clearly depend upon the chosen measure of dissimilarity. On the one side, this can be regarded as a limitation, which could possibly be overcome by performing sensitivity analyses aimed at assessing if and to what extent the dissimilarity measure affects results and conclusions. The possibly different results that one may obtain with two different dissimilarities may actually emphasize the features of the sequences that matter the most when comparing groups of trajectories in a specific application.

On the other hand, if the choice of the dissimilarity does matter – because it captures the features of the sequences that are deemed to be most relevant by the researcher – then dissimilarity-based measures allow for a comparison that focuses on such features, thus incorporating *a priori* knowledge into the assessment of the performance of the models.

References

Aassve A., Billari F.C. and Piccarreta, R. 2007. “Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories.” European Journal of Population, 23: 369-388.

Abbott A. 1995. “Sequence Analysis: New Methods for Old Ideas.” Annual Review of Sociology. 21: 93-113.

Agresti A. 2002. Categorical Data Analysis. New York: John Wiley &.

Bartolucci F., Farcomeni A., and Pennoni F. 2013. Latent Markov Models for Longitudinal Data. Boca Raton: Chapman and Hall/CRC press.

Beyersmann J., Allignol A., and Schumacher M. 2012. Competing Risks and Multistate Models with R. New York: Springer.

Bonetti M. 2016. Interpoint Distance Distribution. Wiley StatsRef: Statistics Reference Online. 1-3.

Bonetti M. and Pagano M. 2005. “The Interpoint Distance Distribution as a Descriptor of Point Patterns, with an Application to Cluster Detection.” Statistics in Medicine. 24 (5): 753-73.

Bonetti M., Piccarreta R., and Salford G. 2013. “Parametric and Nonparametric Analysis of Life Courses: An Application to Family Formation Patterns. Demography. 50(3): 881-902.

Cai L., Hayward M.D., Saito Y., Lubitz J., Hagedorn A. and Crimmins E. 2010. “Estimation of Multi-State Life Table Functions and their Variability from Complex Survey Data Using the SPACE Program.” Demographic Research. 22(6): 129-158.

Cai L., Schenker N. and Lubitz J. 2006. “Analysis of Functional Status Transitions by Using a Semi-Markov Process Model in the Presence of Left-censored Spells.” Journal of the Royal Statistical Society, Series C. 55(4): 477-491.

Edington E.S., & Onghena P. 2007. Randomization Tests. London, UK: Chapman and Hall/CRC.

Gabadinho A., Ritschard G., Müller N.S., and Studer M. 2011. “Analyzing and Visualizing State Sequences in R with TraMineR.” Journal of Statistical Software. 40(4): pp. 1-37.

Latten J. and De Graaf A. 1997. Fertility and Family Surveys in Countries of the ECE Region. Standard Country Report. The Netherlands. New York, NY and Geneva: United Nations.

Lawless J. F. 2003. Statistical Models and Methods for Lifetime Data. New York: John Wiley & Sons.

Lombardi S. (2012). Multistate Models for Event-history Data: Methods and Applications to Women’s Childbearing and Family Formation Patterns. MSc Thesis. Milano: Bocconi University.

Manjourides J. 2009. Distance Based Methods for Space and Time Modelling of the Health of Populations. PhD diss. Harvard School of Public Health, Department of Biostatistics.

McVicar D., and Anyadike-Danes M. 2002. “Predicting Successful and Unsuccessful Transitions from School to Work by Using Sequence Methods.” Journal of the Royal Statistical Society Series A. 165: 317-334.

Mosler K. 2002. Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach. New York: Springer.

Piccarreta R. and Billari F.C. 2007. “Clustering Work and Family Trajectories using a Divisive Algorithm.” Journal of the Royal Statistical Society, Series A. 170: 1061-1078.

Pesarin F. 2001. Multivariate Permutation Tests with Applications in Biostatistics. Chichester, UK:Wiley.

Putter H., Fiocco M., and Geskus R.B. 2007. “Tutorial in Biostatistics: Competing Risks and Multi-state Models.” Statistics in Medicine. 26: 2389-2430.

Rohwer G. and Pötter U. 2004. TDA User’s Manual. Bochum: Ruhr-Universitat Bochum.

Sankoff D. and Kruskal J.B. 1983. Time Warps, String Edits and Macromolecules. Addison-Wesley, Reading.

Sheikh Y., Khan E., and Kanade T. 2007. “Mode-seeking by Medoid Shifts.” Pp. 1-8 in ICCV 2007: Proceedings of the 11th International Conference on Computer Vision IEEE Computer Society, Washington, DC.

Studer M., Ritschard G., Gabadinh, A., and Müller N.S. 2011. “Discrepancy Analysis of State Sequences.” Sociological Methods and Research. 40(3): 471-510.

Studer M. and Ritschard G. 2016. “What Matters in Differences between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures.” Journal of the Royal Statistical Society, Series A. 179: 481-511.

Tebaldi P., Bonetti M., and Pagano M. 2011. “M Statistic Commands: Interpoint Distance Distribution Analysis.” The Stata Journal. 11(2): 271-289.

Vermunt J.K., Tran B., and Magidson J. 2008. Latent Class Models in Longitudinal Research. Pp. 373-385 in *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, edited by S. Menard. Burlington, MA: Elsevier.

Appendix A

In this Appendix, we report the maximum likelihood estimates of the MSLT model and of the SCM model as obtained in Lombardi (2012) and in Bonetti et al. (2013) for the FFS data. In particular, Table A1 shows the maximum likelihood estimates for the MSLT model, and Tables A2 and A3 refer to the duration and for the transition components of the SCM model. Details of the data, the model selection procedures, and the interpretation of the results, can be found in the two references.

Table A1. MSLT estimates and p-values. Reprinted with permission from Lombardi (2012)

Transition to N			
Parameter	**Estimate**	**S.E.**	p**-values**
Intercept	7.847	0.2126	<0.0001
Previous state = **M**	–11.560	0.2672	<0.0001
Previous state = **U**	–6.735	0.2501	<0.0001
Tval	–0.015	0.0014	<0.0001
Age	0.019	0.0053	0.043
Age^2	–0.0002	0.00005	0.0004
Educ2	0.261	0.1361	0.055
Educ3	1.086	0.1359	<0.0001
Religion	–0.192	0.0833	0.021
Divorce	–0.247	0.1559	0.11
Cohort	0.139	0.0786	0.077
Transition to M			
Parameter	**Estimate**	**S.E.**	p**-values**
Intercept	3.072	0.1929	<0.0001
Previous state = **M**	0.637	0.1837	0.0005
Previous state = **U**	0.847	0.2335	0.0003
Tval	–0.006	0.0011	<0.0001
Age	0.012	0.0034	0.0004
Age^2	–0.0001	0.00003	<0.0001
Educ2	0.248	0.0985	0.012
Educ3	0.416	0.1003	0.0001
Religion	0.043	0.0619	0.48
Divorce	–0.310	0.1141	0.007
Cohort	0.042	0.0591	0.470
Transition to U			
Parameter	**Estimate**	**S.E.**	p**-values**
Intercept	2.496	0.2252	<0.0001
Previous state = **M**	–8.554	0.4064	<0.0001
Previous state = **U**	2.894	0.2386	<0.0001
Tval	–0.003	0.0015	0.071
Age	0.023	0.0047	<0.0001
Age^2	–0.0002	0.00004	<0.0001
Educ2	0.539	0.1510	<0.0004
Educ3	1.202	0.1493	<0.0001
Religion	–0.507	0.0855	<0.0001
Divorce	–0.036	0.1548	0.82
Cohort	0.354	0.0828	<0.0001

Table A2. Parameter estimates and p-values for the duration component of the SCM model. Reprinted from Bonetti et al. (2013; page 897) with permission from Springer

Parameter	Estimate	S.E.	p-value
Intercept	–4.329	0.0301	<0.00001
Age	–0.012	0.0006	<0.00001
Previous state = **M**	0.900	0.0410	<0.00001
Previous state = **U**	1.077	0.0458	<0.00001
Cohort	0.079	0.0300	0.009

Table A3. Parameter estimates and p-values for the transition component of the SCM model. Reprinted from Bonetti et al. (2013; pages 897 and 898) with permission from Springer

	Transition from N to M			Transition from N to U		
Parameter	**Estimate**	**S.E.**	p**-value**	**Estimate**	**S.E.**	p**-value**
Intercept	2.799	0.5736	<0.0001	–4.071	0.8806	<0.0001
Tval	–0.005	0.0064	0.22	0.013	0.0064	0.02
Age	–0.054	0.0080	<0.0001	–0.270	0.0086	0.0009
Educ2	0.319	0.5456	0.28	0.394	0.8540	0.32
Educ3	0.110	0.5274	0.42	1.546	0.8380	0.03
Religion	1.869	0.3674	<0.0001	–0.253	0.3769	0.25
Divorce	–2.259	0.4213	<0.0001	0.890	0.5493	0.05
Cohort	–0.145	0.3413	0.34	0.390	0.3711	0.15
	Transition from M to N			**Transition from M to U**		
Parameter	**Estimate**	**S.E.**	p**-value**	**Estimate**	**S.E.**	p**-value**
Intercept	2.553	0.7213	0.0002	1.652	0.5760	0.002
Tval	–0.013	0.0063	0.02	0.005	0.0063	0.22
Age	–0.015	0.0057	0.005	–0.010	0.0050	0.02
Educ2	–0.107	0.6485	0.43	0.647	0.5446	0.12
Educ3	0.695	0.6343	0.14	1.201	0.5248	0.01
Religion	–0.161	0.3416	0.32	0.672	0.3625	0.03
Divorce	–0.515	0.4290	0.11	–1.083}	0.3762	0.002
Cohort	–0.384	0.3557	0.14	0.782	0.3356	0.01
	Transition from U to N			**Transition from U to M**		
Parameter	**Estimate**	**S.E.**	p**-value**	**Estimate**	**S.E.**	p**-value**
Intercept	–4.657	1.3044	0.0002	2.932	0.6849	<0.0001
Tval	0.007	0.0136	0.31	–0.018	0.0059	0.001
Age	–0.0003	0.0130	0.49	–0.008	0.0052	0.06
Educ2	0.567	1.1027	0.30	0.398	0.6048	0.25
Educ3	–1.088	1.3913	0.22	0.819	0.5963	0.08
Religion	–0.400	0.7203	0.29	0.101	0.3168	0.38
Divorce	–1.086	2.3363	0.32	–1.076	0.4019	0.004
Cohort	–0.473	0.7413	0.26	0.062	0.3342	0.43

Appendix B

Figure B1 shows the distributions of the individuals' prediction errors corresponding to the different values of each covariate.

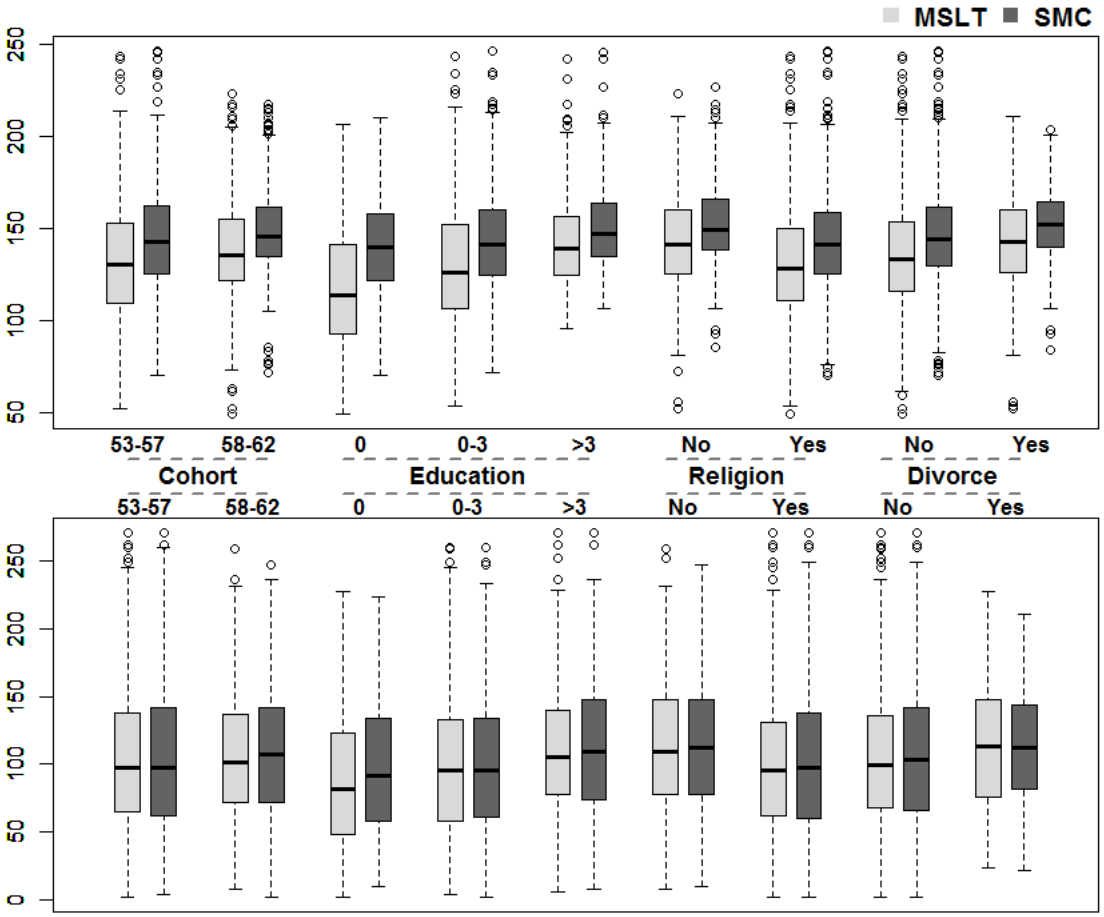


Figure B1. Distributions of the individual prediction errors, $\hat{\varepsilon}_i$, average dissimilarity of the observed sequence from the simulated ones (top panel) and $\tilde{\varepsilon}_i$, dissimilarity of the observed sequence from the medoid of the generated ones (bottom panel) for covariates' values.