Mining event histories: a social science perspective

Gilbert Ritschard*, Alexis Gabadinho, Nicolas S. Müller and Matthias Studer

Department of Econometrics and Laboratory of Demography,
Faculty of Economics and Social Sciences,
University of Geneva, 40, bd du Pont-d'Arve,
CH-1211 Geneva, Switzerland
Fax: +41 22 3798299
E-mail: gilbert.ritschard@unige.ch
E-mail: alexis.gabadinho@unige.ch
E-mail: nicolas.muller@unige.ch
E-mail: matthias.studer@unige.ch
*Corresponding author

Abstract: We explore how recent data mining-based tools developed in domains such as biomedicine or text mining for extracting interesting knowledge from sequence data could be applied to personal life course data. We focus on two types of approaches: 'survival' trees that attempt to partition the data into homogeneous groups regarding their survival characteristics, i.e., the duration until a given event occurs and the mining of typical discriminating episodes. We show how these approaches may fruitfully complement the outcome of more classical event history analyses and single out some specific issues raised by their application to socio-demographic data.

Keywords: event histories; state sequences; event sequences; mining frequent episodes; discriminating subsequences; survival trees; social science; life course; longitudinal data; data mining; data modelling; data management.

Reference to this paper should be made as follows: Ritschard, G., Gabadinho, A., Müller, N.S. and Studer, M. (2008) 'Mining event histories: a social science perspective', *Int. J. Data Mining, Modelling and Management*, Vol. 1, No. 1, pp.68–90.

Biographical notes: Gilbert Ritschard is a Professor of Statistics for the Social Sciences at the University of Geneva. His present research interests are in the use of classification trees and other data mining techniques within social sciences. He is one of the authors of the TraMineR package for sequence analysis in R, which is freely available from the CRAN (cran.r-project.org).

Alexis Gabadinho is a Doctoral student and received his MSc in Demography. He is one of the authors of the TraMineR package.

Nicolas S. Müller is a Doctoral student. He received his BA in Sociology and an MSc in Information Science. He is one of the authors of the TraMineR package.

Matthias Studer is a Doctoral student. He earned his BA in Economics and MA in Sociology. He is one of the authors of the TraMineR package.

1 Introduction

An individual life course paradigm emerged during the '80s from disciplines such as sociology and population studies. It states that analysing the time evolution of aggregated quantities such as the average age of women who married each year, the ratio of the number of new births on the number of women in age of pro-creating, or the proportion of unemployment is not sufficient and that we have to look at individual trajectories for understanding the social forces behind the way people organise their personal life courses. Much effort has been put for collecting individual longitudinal data. Many countries conduct nowadays large panel surveys which permit to follow sampled individuals during a great number of years. Retrospective biographical surveys such as the family and fertility survey (FFS) have also been conducted. The statistical match between censuses, population registers and possibly other administrative data sources permits also to create very rich databases of individual longitudinal data. All these data collection efforts would, nevertheless, be worthless without suitable tools for discovering interesting knowledge from life course data.

Personal life courses are defined by a succession of events regarding living arrangement, familial life, education, professional career, health, etc. Methods for analysing them are of mainly two sorts.

1 Methods that focus on a specific event – leaving home, marriage, childbirth, first job – and examine how the hazard of experiencing it evolves with time (since a specified starting event) and may be affected by other factors. We refer to them as survival methods since they are concerned with how long the subject survives in a given state, married for instance. They include the well known Kaplan-Meier (KM) survival curves and Cox proportional hazard model.

2 Methods for sequence analysis that are primarily concerned by the order in which events occur and the transition mechanism between successive states. These include among others discrete Markov models and optimal-matching-based clustering.

After an overview of these methods, our aim in this article is to show how life course analysis could benefit from non-parametric heuristic data mining-based approaches. Data mining of sequence data has proven its relevance in fields such as automatic text or web log analysis, biostatistics, medicine or marketing. Despite this increasing interest for sequence data or event histories, it received, however, only little attention until now within the social sciences. We put stress on the original insight that we may expect from such methods and discuss specific issues related with their application on socio-demographic data, specially the handling of time varying covariates and multilevel effects.

The paper is organised as follows. We begin in Section 2 by shortly discussing alternative representations of life course data. In Section 3, we make an overview of the most common methods used by social scientists for life course analysis and propose a typology distinguishing between survival and sequence methods, but also between descriptive and causal approaches, between parametric and non-parametric models. We then focus on two promising data mining-based approaches for life course data. In Section 4, we present survival trees and discuss their interest for detecting interactions among covariates, while Section 5 is devoted to the mining of typical sequential patterns

and the identification of those patterns that most discriminate given groups. Finally, we make some concluding remarks in Section 6.

2 Time to event and state sequence views

There are different ways of organising event histories data and each method may require a specific organisation. A life event can be seen as the change of state of some discrete variable such as the marital status, the number of children, the job, or the place of residence. Such life history data are collected in mainly two ways: as a collection of time stamped events (Table 1) or as state sequences (Table 2). In the former case, each individual is described by the realisation of each event of interest (e.g., being married, birth of a child, end of job, moving) mentioned together with the time at which it occurred. In the second case, the life history of each individual is represented by the sequence of states of the variables of interest, each state being given in regard of the corresponding period. Panel data are special cases of state sequences where the states are observed at periodic time.

Table 1 Time stamped event view, record for person id1

Ending secondary school in 1970	First job in 1971	Marriage in 1973

Table 2 State sequence view, person id1

Year	*1969*	*1970*	*1971*	*1972*	*1973*
Marital status	Single	Single	Single	Single	Married
Education level	Primary	Secondary	Secondary	Secondary	Secondary
Job	No	No	First	First	First

It is always possible to transform time stamped data into state sequences and reciprocally. It is sometimes also useful to put the data into spell view with a new line each time a change occurs in the state of any variable or in person-period form with one line for each period where the person is under observation. The latter form is almost the transpose of the state sequence view. The only difference is that periods where a person is not under observation give rise to missing values in the state sequence view, while the concerned lines would simply be dropped in the person-period presentation.

3 Methods for life events analysis

The aim of this section is to shortly survey the main methods available for dealing with individual life course data. We first recall classical statistical methods and then present promising data mining-based approaches. In each case, we distinguish between methods intended for time stamped event data and those that deal with sequences.

3.1 Classical statistical and data analysis methods

Methods most often used by social scientists are concerned with the duration T between two specific events, birth and leaving home, first union and first child, for example. They assume data in time stamped form and try to answer questions about the distribution of the 'survival' probabilities, i.e., the probabilities $S(t) = p(T > t)$ of not experiencing the second event before a duration t. We can distinguish descriptive methods that just attempt to describe the survival function $S(t)$, and causal or explanatory methods used to investigate the factors that may influence the survival curves $S(t \mid \mathbf{x})$, with $\mathbf{x}$ the vector of factors. The latter methods consist in regression-like models that express, for instance, the hazard rate $h(t) = p(T = t \mid T \geq t)$ as a transform of a linear function of the predictors $h(t \mid \mathbf{x}) = g(\mathbf{x}' \beta)$.

As for state or event sequence data, Abbott (1990, p.377) distinguishes three kinds of questions.

1 Are there typical sequence patterns, for instance, does the first job typically follow the end of education and precede leaving home, and if yes, what are their frequencies?

2 Given a set of sequence patterns, why are they the way they are? Which independent variables determine which pattern is observed? Does the socio-professional status, for example, influence the familial life course (time of marriage, number and timing of children)?

3 What are the effects of a given sequence patterns on some variables of interest? For example, does the specific pattern of the successive educational, professional and familial events influence the chances to be in good health at retirement time?

The first kind of questions has a descriptive concern, while the other two are issues of causality.

The previous discussion suggests the typology shown in Table 3. This table summarises the main methods that are used in the literature for analysing life events data. The survival analysis methods used with time stamped events are shared with biomedicine and industrial quality control where the concern is just the death of a patient or of a device, hence the term 'survival'. These 'survival' methods are perhaps the most widely used for event history analysis. They are well explained in several excellent textbooks, for instance in Yamaguchi (1991) and in Blossfeld and Rohwer (2002) with a social science perspective, and in Hosmer and Lemeshow (1999) from a biomedical point of view. The main feature of these methods is the handling of censored data, i.e., cases that run out of observation while at risk of experiencing the studied event. Hazard regression models, with discrete or continuous time, especially the semi parametric Cox (1972) model, are well suited for analysing the causes of events. Their success is largely attributable to their availability in standard statistical packages and to the ease of interpretation of the regression like coefficients, they produce. Advanced issues regarding these models include the simultaneous analysis of several events (Lillard, 1993; Hougaard, 2000) and the handling of variables shared by members of a same group, i.e.,

multilevel analysis (Courgeau and Baccaïni, 1998; Barber et al., 2000; Therneau and Grambsch, 2000; Ritschard and Oris, 2005).

Methods for sequence analysis, though best suited for analysing trajectories in a holistic perspective (Billari, 2005) are less popular. This is certainly due to the lack of friendly software for dealing with sequence data. A first simple approach consists just in counting the occurrences of predefined subsequences. This leads indeed to consider the predefined subsequences of interest as categorical variables, which may then be analysed with tools for such variables, log-linear models (Hogan, 1978) or classification trees (Billari et al., 2006) for instance.

Table 3 A typology of methods for life course data

Questions	*Nature of data*	
	Time stamped event	*State/event sequences*
Descriptive	• Survival curves: Parametic (Weibull, Gompertz) and non-parametic (KM, Nelson-Aalen) estimators	• Frequencies of typical patterns • Optimal matching clustering • *Discovering typical episodes*
Causality	• Hazard regression models • *Survival trees*	• Markov models, *mobility trees* • *Finding discriminating* episodes among groups

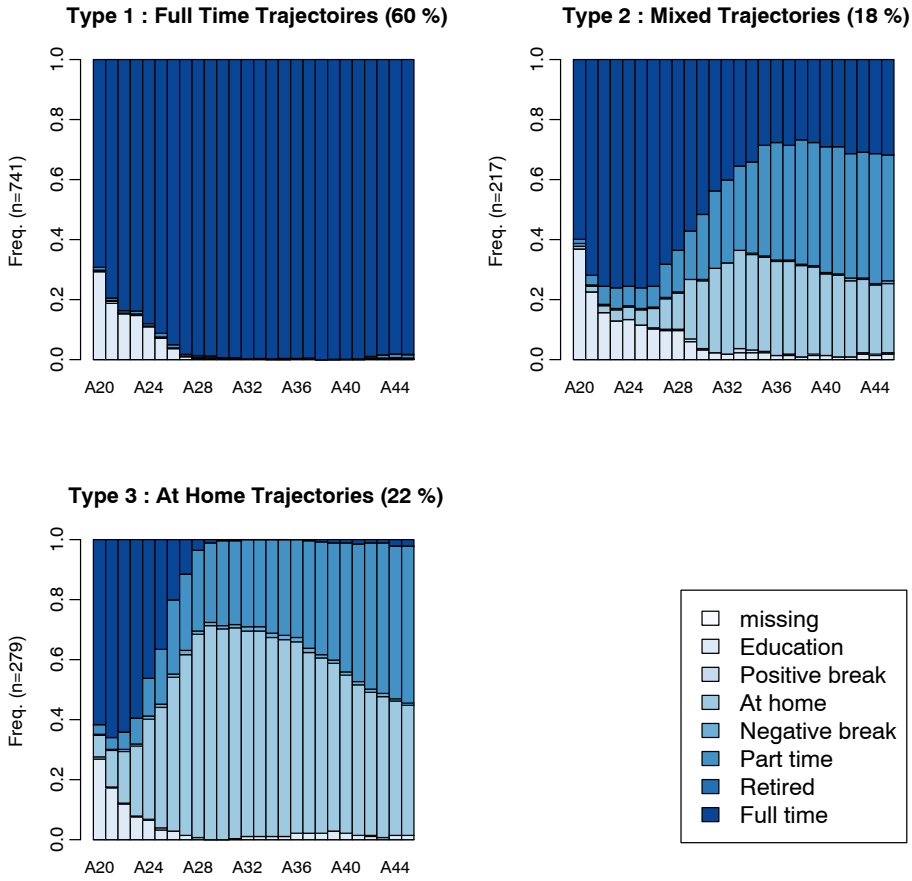
Note: Italic is used for data mining-based methods addressed in this article.

Clustering based on the edit distance (Levenshtein, 1966; Needleman and Wunsch, 1970; Sankoff and Kruskal, 1983) between each pair of sequences has been popularised in social sciences by Abbott (see Abbott and Tsay, 2000) under the name of optimal matching and was for example exploited by McVicar and Anyadike-Danes (2002), Malo and Munoz (2003) and Levy et al. (2006). See Abbott and Tsay (2000) for a survey of earlier social science works carried out in this field and the accompanying discussion for criticisms. The method is mainly descriptive. It consists in making a typology of the population by grouping together individuals with similar life course patterns. The life course associated to each class of the typology is then analysed by looking at how the probabilities to be in the different possible states change over the age scale. This produces nicely interpretable aggregated results. Figure 1 shows for instance the aggregated profile for the optimal matching clusters obtained for Swiss occupational sequences between age 20 and 45. Such representations are judiciously complemented with index plots (Scherer, 2001) depicting the variability of individual trajectories inside each cluster. Optimal matching clustering can be realised, for instance, with free software such as TDA (Rohwer and Pötter, 2002), with the SQ package (Brzinsky-Fay et al., 2006) for Stata or in R with our TraMineR package (Gabadinho et al., 2008). Recent developments regarding optimal matching include training procedures for learning 'optimal' state substitution costs (Gauthier et al., 2008). Similar techniques based on non-aligning similarity measures have also been recently considered for instance by Elzinga (2003).

Another useful method for sequence data is discrete Markov modelling that focuses on the state transition probabilities between two successive time points. They are often used for mobility analyses. Advances in this area include the modelling of high order process (Raftery and Tavaré, 1994; Berchtold and Raftery, 2002), hidden Markov models (HMM) (Rabiner, 1989) and their generalisation as double chain Markov models (DCMM) (Paliwal, 1993; Berchtold, 2002) and Markov models with covariates

[Berchtold and Berchtold, (2004), p.50]. Despite these advances, the estimation of Markov models lacks often reliability due to large standard errors of the estimated transition rates and the results provided remain hard to interpret when we departure from very simple specifications.

Figure 1 Aggregated view of Swiss occupational trajectories between 20 and 45 for each group of a three cluster solution obtained through optimal matching (see online version for colours)



Source: Based on data of the 2002 biographical survey of the Swiss Household Panel (SHP)

3.2 Data mining-based approaches

Data mining is mainly concerned with the characterisation of interesting patterns, either per se (unsupervised learning) or for a classification or prediction purpose (supervised learning). Unlike the statistical modelling approach, it makes no assumptions about an underlying process generating the data and proceeds mainly heuristically.

Data mining-based approaches were recently considered for analysing individual life courses from a socio-demographic point of view. Blockeel et al. (2001) showed how mining frequent item sets might be used to detect temporal changes in event sequences frequency from the Austrian FFS data. In Billari et al. (2006), three of the same authors also experienced an induction tree approach for exploring differences in Austrian and Italian life event sequences. Ritschard and Oris (2005) initiated social mobility analysis with induction trees.

A lot of works has also been done within the field of biomedicine. Of special interest for discriminating life courses are survival trees (Segal, 1988; Leblanc and Crowley, 1992, 1993; Ahn and Loh, 1994; Ciampi et al., 1995; Huang et al., 1998; Su and Tsai, 2005). Their principle is based on that of classification and regression trees (Kass, 1980; Breiman et al., 1984; Quinlan, 1993) that are especially good at discovering interactions effects of explanatory variables. They recursively seek the best way to partition the population according to values of the predictors so as to get survival probability curves or hazard functions that differ as much as possible from one group to the other. De Rose and Pallara (1997) have demonstrated the usefulness of this approach for socio-demographical analyses.

From this short survey, we may distinguish mainly three data mining techniques that seem promising for discovering interesting knowledge from life event data. We have reported them in italic in Table 3.

1 Within the spirit of 'survival' methods, survival trees should complement regression like models by helping at discovering interaction effects between covariates. They will clearly exhibit differential effects such as, for example, the consequence of having a first child on the activity rate that differs between women and men, but may also vary with cultural origin and other factors.

2 Methods for seeking typical subsequences are by their very nature well suited for the analysis of sequence data. Their outcome, i.e., typical subsequences, may then be used as either response or predictive variables for causal analysis.

3 The mining of interesting association rules between frequent subsequences is clearly of interest in the causal perspective. It will lead to statements such as, for example, having experienced the subsequence first job, first union, first child, is most likely to be followed by a sequence marriage, second child.

4 Survival trees

Survival trees are based on the principle of recursive partitioning algorithms. They require, however, specific splitting criteria adapted for the survival concern. In addition, since such trees are intended for dealing with censored data, the usual minimal node size constraints should be completed with additional constraints on the minimal number of events occurring in each node. We briefly explain hereafter the main splitting criteria used for survival trees.

4.1 Splitting criteria

As for classification trees, there are two main groups of splitting criteria: those that attempt to maximise the group difference in the spirit of CHAID (Kass, 1980) and other earlier tree growing methods, and those that maximise group homogeneity such as CART (Breiman et al., 1984) or C4.5 (Quinlan, 1993) for instance.

Between group survival curve divergence

A first idea considered for instance by Segal (1988) is to split each node so as to obtain Kaplan-Meier (KM) estimates of the survival curve that differ as much as possible between the two resulting nodes. The divergence between KM curves is measured with a chi-square statistic of the general Tarone-Ware family

$$TW = \sum_i \frac{w_i \left(d_{i1} - \mathrm{E}(D_i)\right)}{\left(w_i^2 \ \mathrm{var}(D_i)\right)^{1/2}}, \tag{1}$$

where the sum is over the time t_i at which at least one event occurs, and d_{i1} is the number of events (deaths) observed in the first group (node) at time t_i, D_i the random number of events that would occur in the first group according to the distribution in the node we want to split and w_i weight parameters. Special cases of this statistic are the Log-rank statistic (using $w_i = 1$), Gehan's statistic ($w_i = n_i$) and the one ($w_i = \sqrt{n_i}$) advocated by Tarone and Ware (1977), n_i standing for the number of cases at risk at time t_i. A more elaborated approach based on the same maximal separation principle can be found in Leblanc and Crowley (1993).

Group homogeneity: maximal likelihood relative risk

Leblanc and Crowley (1992) proposed to estimate for each node the maximal likelihood hazard proportionality factor (relative risk) and to select the split that maximises the gain in likelihood, or equivalently the reduction in deviance. The approach supposes that the hazard $\lambda_h(t)$ in each node h is proportional to a reference hazard (the overall hazard for the root node): $\lambda_h(t) = \theta_h \lambda_0(t)$. Estimations of the θ_h parameters are based on a full likelihood that can be derived assuming a known cumulative hazard function $\Lambda_0(t)$. Practically, since the cumulative hazard is not known, the authors rely on an iterative estimation process in which $\hat{\theta}_h$ and $\hat{\Lambda}_0(t)$ are estimated in turn. Notice that maximising the reduction in deviance amounts to maximise group homogeneity. Hence, this approach is more in line with classical tree growing algorithms such as CART or C4.5, which attempt to maximise some measure of node purity. It is available, for instance, in the *rpart* package (Therneau and Atkinson, 1997) for S-plus and R.

A related approach is that of Ciampi et al. (1995) who attempt to maximise Cox's partial likelihood of semi-parametric proportional hazard models. Their method is an instantiation of a general regression tree method (Ciampi, 1991) based on likelihood

maximisation. The method parallels CART but considers a pruning criterion in terms of loss of information – deterioration of deviance – with respect to a first large grown tree. A similar principle is adopted by Leblanc and Crowley (1992).

Ahn and Loh (1994) consider an approach based on the martingale residuals of a Cox model. Plotting at each node, these residuals against each covariate, they select as splitting variable the one for which the residuals look the less random, i.e., produce the most homogenous groups. For measuring randomness, residuals are split into those above their median values and the other ones. The randomness measure is then the p-value of the Levene test for the difference in variances between the two groups. The method can be seen as a special case of a more general method implemented in GUIDE (Loh, 2007). The method uses a deviance-based goodness-of-fit to determine whether the selected split is worth enough to continue growing the tree.

A well-known issue with recursive partitioning is that predictors with many different values are more likely to be selected as splitting variable than predictors offering fewer splitting possibilities (Kim and Loh, 2001; Hothorn et al., 2006b). Proposed unbiased strategies consist in separating the predictor selection process from the determination of the optimal split for a given predictor. Loh (2007)'s GUIDE program as well as the *party* package (Hothorn et al., 2006a) for R, which is based on a flexible linear statistic, both permit to grow survival trees using such strategies.

4.2 Illustration

Figure 2 shows a survival tree grown for the risk of divorce or more specifically for the duration of the marriage until divorce. Data come from the retrospective biographical survey carried out by the SHP in 2002. The criterion used consisted in maximising the differences between KM survival curves using the significance of the Tarone-Ware test. A 5% significance limit was retained as stopping rule. Explanatory factors considered include among others birth cohort, education level, whether ego had a child or not, language of the questionnaire and religious practice, the latter two being cultural indicators. In the nodes of the trees, we have indicated the number n of concerned cases, the number e of events (divorces), the 90% percentile of the survival probability S, and the survival probability at 30. The KM survival curves corresponding to the seven leaves (terminal nodes) of the tree are depicted in Figure 3.

It results clearly from the tree that the risk of divorce increases dramatically between those who are born before 1940 and younger generations, the 90% percentile falling from 21 to nine. We notice also that, while for the older generation there was a significant distinction between the French speaking population and the rest of the Swiss population – divorce being more common in the French speaking region – this distinction is for the younger generations limited to those who had a child. Non-French speaking people born before 1940 with education below university level are the less exposed to divorce. On the other side, those born after 1940 without child but with high education level are the most exposed.

Growing the tree with Leblanc and Crowley's (1992) approach, we obtain a somewhat simpler tree (Figure 4) corresponding to the first two levels of the tree obtained with the Tarone-Ware criterion. From the relative risks provided by the method, we learn, for instance, that the risk of divorcing for non-French speaking people born before 1940

is only about 48% of the risk for the whole population, while it is almost 1.9 greater for younger generations with no child.

Figure 2 Survival tree for marriage duration until divorce/separation (Tarone-Ware criterion) (see online version for colours)

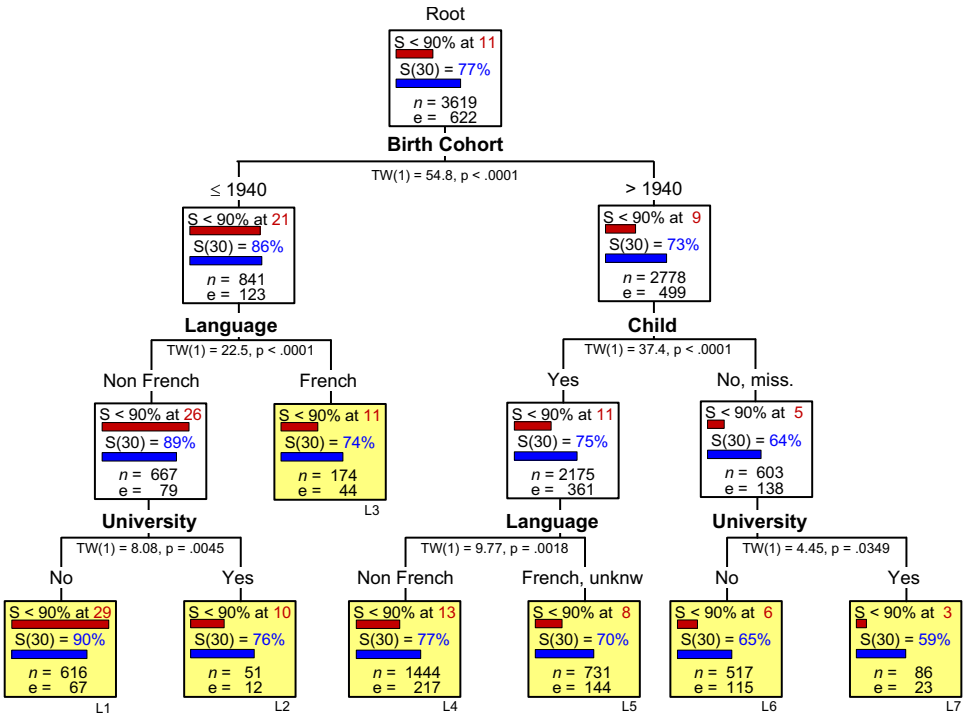


Figure 3 Survival curves associated to the seven leaves of the tree in Figure 2 (see online version for colours)

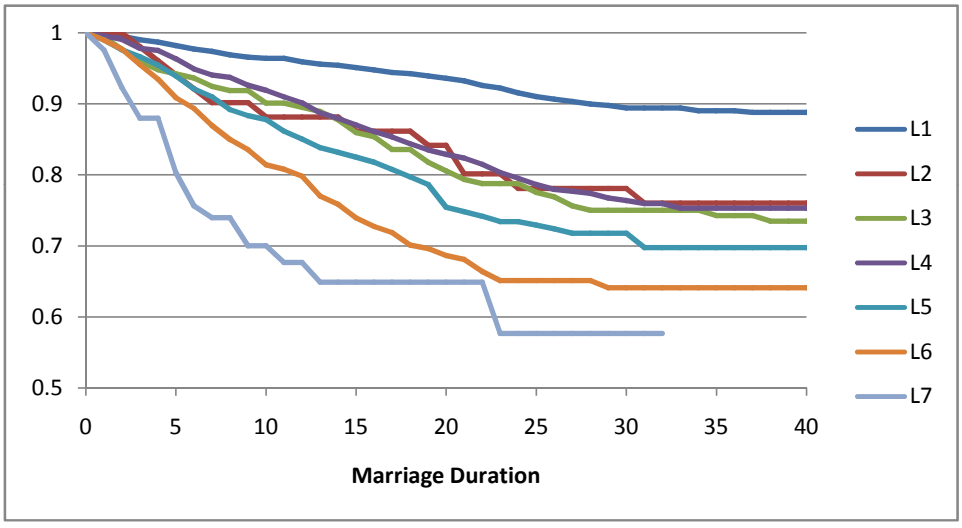


Figure 4 Proportional risk tree for marriage duration until divorce/separation (see online version for colours)

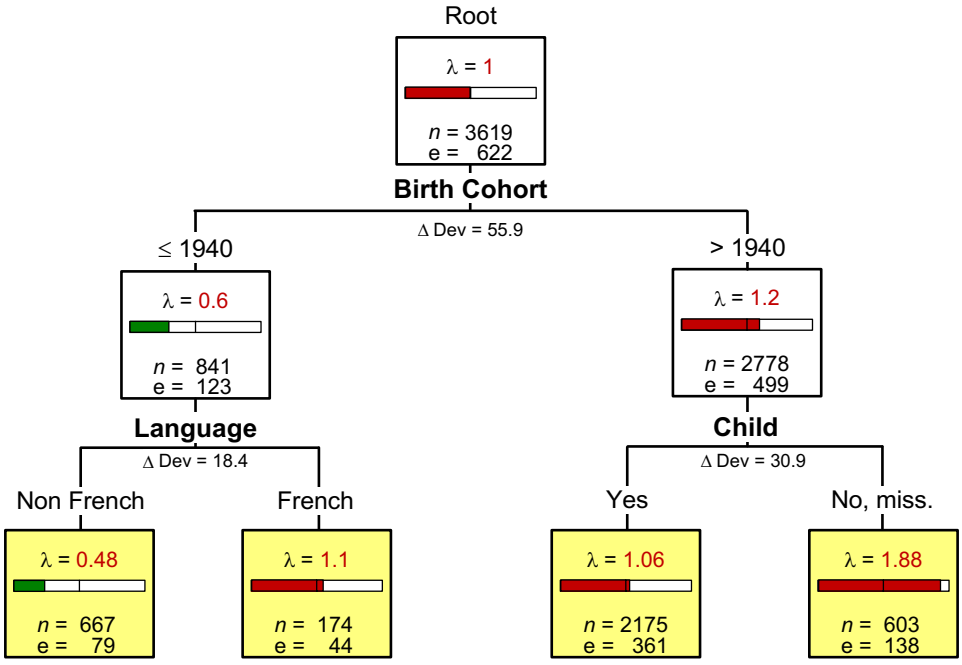


Table 4 Logistic regression on individual-year data with and without interaction effects

		exp(B)	*Sig.*	*exp(B)*	*Sig.*
Born after 1940		1.34	0.000	1.78	0.000
University		1.21	0.053	1.22	0.049
Child		0.73	0.000	0.94	0.619
Language	Unknown	1.49	0.000	1.50	0.000
	French	1.27	0.006	1.12	0.282
	German	1	Ref	1	Ref
	Italian	0.91	0.625	0.92	0.677
b_before_40*French				1.46	0.028
b_before_40*Child				0.68	0.010
Constant		0.01	0.000	0.01	0.000
Model $\mathcal{X}^2$		53.7	($d = 6$)	64.6	$d = 8$
$\Delta\mathcal{X}^2$			10.9, $d = 2$, sig = .004		

Notes: The exp(B) values are the exponentials of the regression coefficients and correspond to odds ratios for an increase of one unit of the associated predictor.

Survival trees advantageously complement classical regression like approaches by identifying important interaction effects. From the trees grown above, we learn for instance that the language effect primarily concerns people born before 1940, while the

presence of a child is a concern for those born after 1940. As can be seen in Table 4, adding these interaction effects in a discrete time logistic regression fitted on person-year data reduces significantly the deviance of the model.

4.3 Issues with survival trees

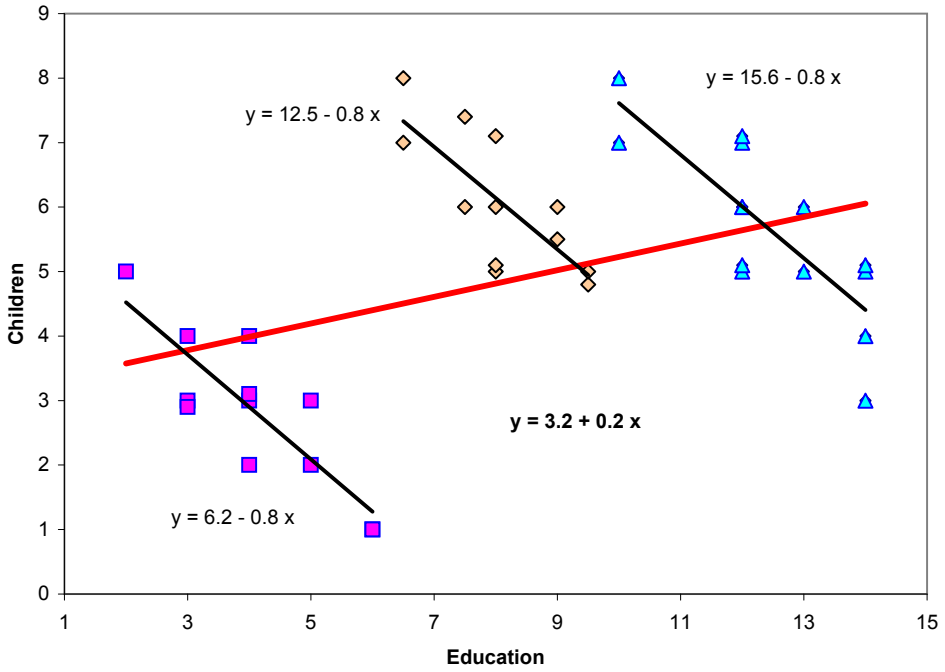
The methods just described were developed in the field of biostatistics. They are, however, also of interest for social sciences as shown by our illustration. When applying them in sociology, socio-demographic history or population studies, we have to take account of specificities of data we may encounter in these domains. We see two major issues.

First, predictors are most often time varying. Education level or income, for instance, changes with the age of each considered individual. Likewise, for the divorce example, the first childbirth may well happen the same year as the marriage for some individuals and only after some years of marriage for other ones. We should then explicitly consider the history of the values taken by such predictor when growing the tree. This is a difficult issue because it requires to defining interpretable splits preserving simultaneously ordering with respect to both the time and the history of values of the predictor. Segal (1992) discusses a few possibilities concluding, however, that none is really satisfactory. Huang et al. (1998) propose a piecewise constant approach that may be suitable for discrete time varying predictor that change values at only a limited number of time points. There is obviously room for development on this aspect.

A second important issue is related to the multilevel organisation of the data. In social science, though it is true in other domains too, data may often be grouped into small units whose members share common characteristics. For example, in the data collected by the SHP, we have small groups of individuals belonging to a same household. The variability among individuals comprises thus a part shared by members of a same unit. Ignoring it may lead to strongly biased results. Figure 5 taken from Ritschard and Oris (2005) illustrates, for instance, what happens in the case of a simple regression. Data are supposed representing the number of children by woman in regard to the education level, and the women are supposed coming from three different villages. Ignoring the village shared effect; regression provides a slightly positive line, indicating a positive relationship: the higher education, the higher the number of children. If we allow for a shared random discrepancy between villages, we would fit the piecewise lines with negative slopes indicating that the number of children decreases with education. Thus, ignoring the discrepancy among villages we fit indeed the village effect rather than individual effects.

It is clear that similar fallacious effects will happen with tree partitioning methods and it is a real challenge to find a way to incorporate such multilevel effects in tree growing procedures.

Figure 5 Multi-level: a simple linear regression example with three clusters (see online version for colours)



Survival trees and more generally survival analysis is very useful when we are interested in one specific event such as the divorce in the illustration shown. It is of poor help, however, if the concern is to gain insights on the individual life course described by the whole collection of events that characterise it. Methods that deal with such whole sequences without privileging one given event are better suited for this unitary, holistic, perspective on life courses (Billari, 2005). This leads us to the second broad class of methods: The mining of typical sequences.

5 Mining typical sequences and sequential relationships

It is worth distinguishing here between state sequences and event sequences. If we look at life courses as state sequences, interesting knowledge may be obtained by seeking patterns in transitions between states. With that perspective, we have shown in Ritschard and Oris (2005) that so called 'mobility trees' provide interesting alternatives to Markov transition models. Such mobility trees are classification trees in which the states of a variable of interest, the working status at time t for instance is taken as response variable, predictors being the same variable at $t - 1$, $t - 2$, ... and other possible covariates. This approach, however, focuses again on a given variable and does not provide the expected holistic view. To state sequences, we may also apply techniques developed for analysing DNA sequences or texts considered as letter sequences. Among those methods, optimal-matching based clustering, which we already discussed in Section 3.2, provide

valuable holistic knowledge in the form of categorisation of whole life courses. Indeed, once we have a matrix of proximities between sequences, whether we obtained it through optimal matching or for instance by using some of Elzinga's (2003) metrics, we may cluster the sequences or visualise their relative positions by means of some multidimensional scaling methods (Müller et al., 2008).

5.1 Mining episodes

If we represent life courses as sequences of time stamped events, we may consider using techniques that have been developed for mining interesting event subsequences or episodes, i.e., collection of events occurring frequently together. Such methods have been developed for instance for discovering customer buying sequence patterns (Srikant and Agrawal, 1996), detecting signal patterns that would announce a device or telecommunication network breakdown (Mannila et al., 1997) or finding sequences of most frequently accessed pages at a website (Zaki, 2000). Different approaches for characterising interesting sequences were considered in the literature among which prominent approaches are those of Bettini et al. (1996), Srikant and Agrawal (1996) and Mannila et al. (1997) for which Joshi et al. (2001) proposed a nice unifying and flexible formulation.

Though mining typical event sequences is in some sense a specialised case of the mining of frequent item sets (Agrawal et al., 1995), it is much more complex and requires the user to specify time constraints and select a counting method. Indeed, if there is general agreement about how to count occurrences of item sets in the classical unordered framework, there is no such agreement for episodes. In the latter case, the additional time dimension raises such questions as: What is the maximal time span, i.e., sequence length we want to analyse? Until which time gap should events be considered to occur simultaneously? For instance, regarding the first of these two questions, if we were interested only in active life, we would exclude events happening say before 15 and after the legal retirement age. Likewise for the second one, ending an education cycle in June and starting a first job in December of the same year could be considered either as simultaneous or parallel events since they occur the same year or as successive events. Moreover, we may consider that two or more events form a relevant sequence only if they occur within a given maximal time span or window length. Leaving home and having a child next year, is not the same as leaving home and having a child ten years later. In case of repeating events, we have also to specify how to count multiple ways of forming similar episodes, i.e., subsequences of types of events. For example, assuming a girl starts a job (J) in 1980 and has children (C) in 1985 and 1987. Should we count the episode (J, C) once or twice? For a rigorous enumeration of all these issues, see Joshi et al. (2001). Clearly, there is no universal answer to all of them. The choice depends largely on the application domain and may be specific to each situation and to what the user is expecting. We briefly comment hereafter about the nature of episode constraints we may want to impose and alternative counting methods.

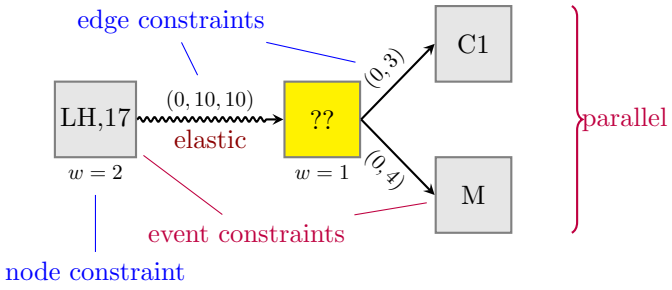
5.2 Sequential relationships

Besides finding frequent episodes, it is interesting to look at the structure of the episodes. Mannila et al. (1997) for instance distinguish between serial (strict sequential order

between events), parallel (no strict order) episodes and possible combination between these two forms. More generally, Joshi et al. (2001) represent episode structure in the form of a directed acyclic graph (DAG), in which nodes may contain simultaneous events, an edge between two nodes indicating that the concerned events are present in that order in the episode.

Such representations provide a convenient way of designing various nodes, windows and overall span time constraints. We may also set node constraints regarding the events they should contain so as to focus the analysis on situations that matter for the problem at hand. To illustrate, assume we are interested in the typical events that occur until people who leave their parental home (LH) within two years of their 17 get married (M) and have a first child (C1). We could then impose, for example, the episode structure depicted in Figure 6, in which we specify the starting and ending events and leave only the intermediate events free. In this structure, getting married (M) and having a child (C1) are shown as parallel events meaning that we do not matter about the order in which they occur. The graph specifies in addition a series of time constraints. The $w = 2$ below the first node, specifies that the events in that node should occur within two years. This is a node constraint. Edge constraints are given by a couple (a, b) meaning that events in the destination node should occur at least after a years, and at most after b years. Thus, the first childbirth should occur at least zero years, at most three years after the last observed event and marriage at least zero and at most four years after. The graph contains an 'elastic edge' between the starting node and the sought event. Such an edge means that though only one free node is represented, we would also cope with more intermediate nodes. The associated elastic edge constraints is in the form (a, b, c), where a and b refer to the minimal and maximal time gap between successive nodes, and the last c specifies the maximal allowed extension. In our example, we do not specify the number of intermediate events, but restrict the search to events occurring at most ten years after the last event (LH or 17) in the starting node.

Figure 6 Time and node constraints (see online version for colours)

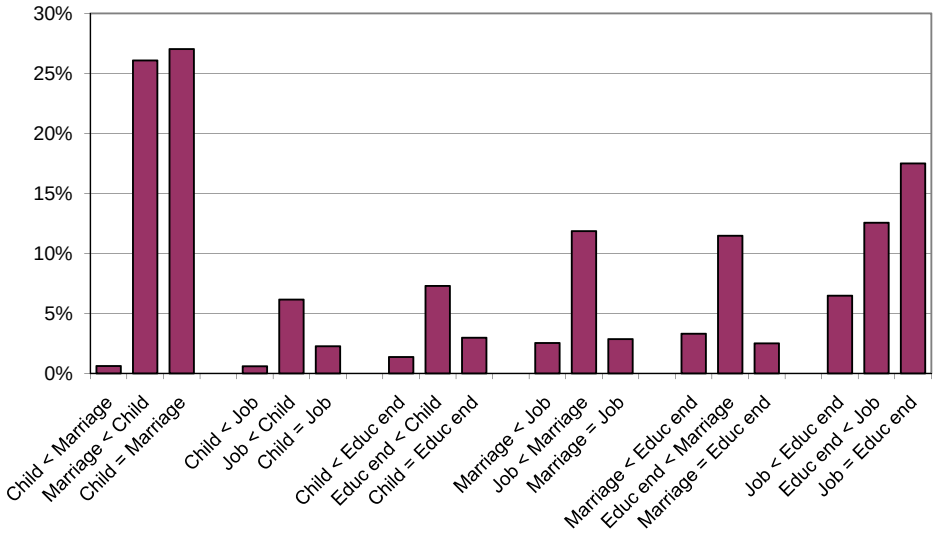


Note: Example for searching typical events occurring until people who leave home (LH) within two years from their 17 both get married (M) and have a first child (C1).

Leaving no free node, we characterise a priori fixed episodes. This may be useful when one is interested in comparing the distribution among different possible structures of a set of episodes. For instance, considering the SHP biographical data, Figure 7 shows the distribution among three alternative structures for the following pairs of event types (education end, 1st job), (education end, marriage), (education end, 1st child), (1st job, 1st child), (1st job, marriage), (marriage, 1st child), (leaving home, 1st job), (leaving

home, education end). The alternative sequencing structures considered are for each pair (x, y): event x happens before y (noted $x < y$), x and y happen the same year ($x = y$), x happens after y ($x > y$). From Figure 7, we learn that it is really exceptional – in 20th century Swiss life courses – to have a child before being married and also before having a first job. The most common situation is to have the first child after ending education and after having found a first job. It is also quite common to start the first job the same year as when we end education.

Figure 7 Distribution of alternative structures of two-event episodes (see online version for colours)

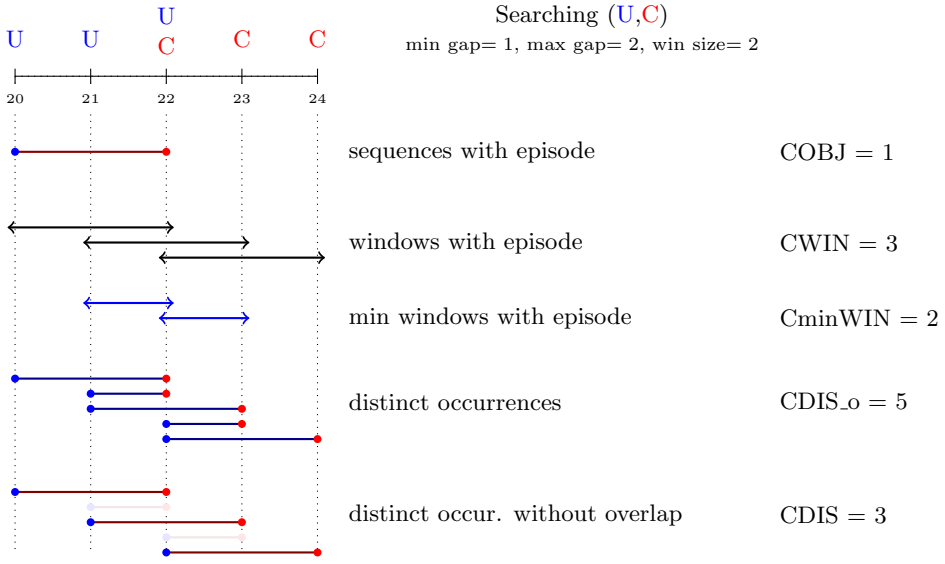


5.3 Counting methods

The choice of the counting method is another important concern for the social scientist. In case of repeating events, there are different ways of forming similar episodes and hence of counting them. For instance, Figure 8 summarises the methods distinguished by Joshi et al. (2001). The result of each method is illustrated for the example sequence depicted at the top of the figure. This example represents a woman who started a first union at 20, a second union at 21 and a third union at 22, and who had a first child at 22, a second child at 23 and a third child at 24. There are thus three different unions and three childbirths. We are interested in sequences (U, C) where the child arrives within one (min gap) or two (max gap) years after the start of the union. The first method (COBJ) consists in counting the objects, i.e., life courses in our case that contain at least one occurrence of the relevant episode. The second way (CWIN) is to slide a window (of length at most two) over the sequence and count how many times the window covers the searched episode. The third way (CminWIN) proceeds similarly but considers only windows of minimal size. On our example, only two windows of minimal size contain the searched episode instead of three windows of length two. The last two methods count the number of distinct occurrences of the episode. With CDIS_o two occurrences are considered as

distinct when they differ in at least one event, while with CDIS, each event can belong to at most one occurrence. For the latter, there is clearly some arbitrariness in the forming of the non-overlapping occurrences. For instance, if we start by forming the two occurrences with length one of the episode (U, C), there remains then no other non-overlapping possibility. Hence, we count two instead of three non-overlapping occurrences.

Figure 8 Different counting methods (see online version for colours)



Note: Figure inspired from Joshi et al. (2001).

A difficulty in life course analysis is that a counting method suited for searching a given episode does not necessarily make sense for some other episode. For instance, counting occurrences without overlap is well suited for an episode such as (union, child) since it does not make sense to associate a same child with different unions. On the other hand, taking overlap into account may be admissible for a sequence (end education, new job) where it is reasonable to associate successive new job starts to a same education cycle. Mixing counting methods in a same frequent episode mining algorithm looks quite difficult because of the need of comparing supports based on different counting schemes.

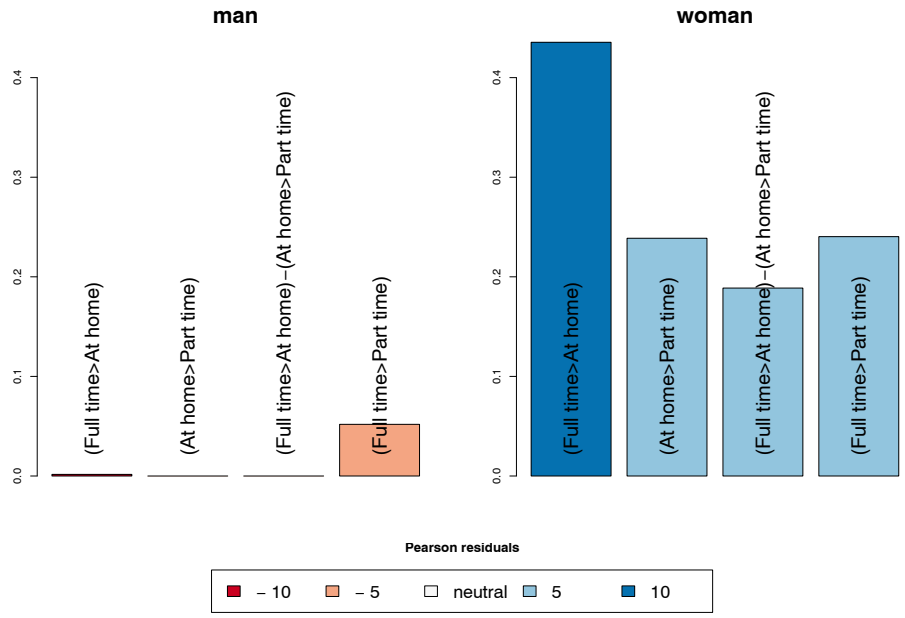
Another difficulty concerns episodes included in longer subsequences. Clearly, an episode included in a longer one cannot be less frequent than the latter. It may in such situation be informative to count the number of sequences that contain the shorter episode but no other frequent subsequences containing it. This would permit for instance to distinguish between people who left home and married from those who left home, married and had a child. This option is available in TraMineR.

The above discussion shows, nevertheless, that mining frequent episodes in life event sequences requires flexible algorithms that can cope with constraints on episode structure, with episode length as well as with different counting methods.

5.4 Discriminating episodes

A frequent episode is not necessarily interesting per se. For instance, for the occupational sequences from the SHP biographical survey, the two most frequent episodes according to the COBJ counting method are non-surprisingly (ending education – starting a full time job) and (ending education – starting a part time job). The social scientist would like to know more, namely, which one among the frequent episodes is the most helpful for distinguishing groups, for instance women from men, or older birth cohorts from more recent ones. The discriminating power of each frequent episode can be measured for instance by the independence chi-square statistic for the cross tabulation of the episode indicator variable by the group variable. Hence, the most discriminating episode is that for which we get the largest chi-square, or alternatively the lowest p-value. This algorithm has been implemented in TraMineR and Figure 9 exhibits the men-women contrast for the four more discriminating episodes found for the Swiss occupational trajectories. We learn thus that stopping a full time job for staying at home is the most discriminating episode between women and men.

Figure 9 Most discriminating occupational episodes for sex (see online version for colours)



Notes: The darker the bar, the larger the deviance from the independence expected count.

Source: Occupational trajectories built from the 2002 SHP biographical survey.

5.5 Episode rules

Social scientists are primarily interested in understanding and explaining social processes rather than in making prediction or classification. They most often formulate their

theories in causal form saying for instance that given characteristics such as being a woman with low education would favour given behaviours (e.g., low activity rate). In that respect, rules stemming from empirical evidence of some implication between two typical episodes will undoubtedly be valuable material for building causal explanations. Though the main aim of mining frequent item sets is to derive such association rules, this aspect has not received special attention in the case of sequentially ordered patterns. For deriving rules, we need indeed some suited criterion such as the confidence or some other interestingness measure. We may indeed use measures similar to those used with unordered item sets. Each of them will, however, result in variants depending on the counting method and various time constraints retained. The multilevel problem that we raised for tree approaches is also a concern for association rules. Indeed, assuming data were collected by clusters we should account of it when validating the rule. A challenge would then be to define interestingness measures able to account for effects shared by members of each cluster.

An interesting issue for the social scientist is to derive association rules between relevant episodes each found in one of two parallel sequences such as the sequence of family events and the professional life course, or the sequence of life events of a woman and that of her partner. One solution could be searching frequent episodes in a mix of the two sequences and then restrict the search of rules among candidates in which the premise and the consequent belong each to a different sequence. Alternatively, we could search frequent episodes in each type of sequence and then search rules among candidates obtained by combining frequent episodes from each sequence.

6 Conclusions

We have seen that there are plenty of ways to look at individual history data, each way having its own advantages. The aim of this presentation was to give a synthesised view of the available methods and especially of the kind of outcome, we may expect from some data-mining-based techniques. We have especially put emphasis on survival tree methods and sequence mining techniques. The former have two major advantages: first, their recursive splitting mechanism produce a tree structured comprehensible output that can be straightforwardly interpreted. Secondly, they automatically detect relevant interaction effects between explanatory factors. Following a branch of the tree, we read how states of different variables combine themselves for defining profiles of homogeneous group regarding the target survival distribution. By thus highlighting interactions, trees complement regression like methods in which the effect of an explanatory factor is – except when an interaction is specifically specified – assumed to be independent of the values taken by the other factors. These tree approaches have, however, also drawbacks. The most important criticism formulated against trees is their potential instability. Indeed, when two predictors have at one node almost the same discriminating power, small changes in the data may lead to change the one that is selected as splitting variable. There is undoubtedly a need for stability criteria, an issue that has for instance been investigated for classification trees by Dannegger (2000). Methods for mining typical event sequences and relationships between such subsequences are perhaps those from which we may expect the most highlighting holistic views on life courses. Unlike survival trees and more generally survival methods, which by their very nature have to focus on a given type of event, extracting typical episodes from life course sequences does not privilege

any type of event and are best suited for discovering prominent characteristics of complete life trajectories. Available techniques, at least those flexible enough for allowing a great number of time and node constraints, are directly applicable to life course data. Nevertheless, for both survival trees and association rules involving episodes, further developments (time varying covariates in survival tree, multilevel effects, mix of counting methods,...) are still necessary to cope with the most exigent needs of the social scientist.

Acknowledgements

This study was realised within a research project supported by the Swiss National Foundation for scientific research under grant SNF-100012-113998. The experimental results reported are based on data collected by the SHP.

References

Abbott, A. (1990) 'A primer on sequence methods', *Organization Science*, Vol. 1, No. 4, pp.375–392.

Abbott, A. and Tsay, A. (2000) 'Sequence analysis and optimal matching methods in sociology, review and prospect', *Sociological Methods and Research*, Vol. 29, No. 1, pp.3–33 (with discussion, pp.34–76).

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I. (1995) 'Fast discovery of association rules', in Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.): *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA, pp.307–328.

Ahn, H. and Loh, W-Y. (1994) 'Tree-structured proportional hazards regression modeling', *Biometrics*, Vol. 50, pp.471–485.

Barber, J.S., Murphy, S.A., Axinn, W.G. and Maples, J. (2000) 'Discrete-time multilevel hazard analysis', in Sobel, M.E. and Becker, M.P. (Eds.): *Sociological Methodology*, The American Sociological Association, New York, Vol. 30, pp.201–235.

Berchtold, A. (2002) 'High-order extensions of the double chain Markov model', *Stochastic Models*, Vol. 18, No. 2, pp.193–227.

Berchtold, A. and Berchtold, A. (2004) 'Markovian model computation and analysis', User's guide, available at http://www.andreberchtold.com/march.html (accessed on 2 March 2002).

Berchtold, A. and Raftery, A.E. (2002) The mixture transition distribution model for high-order Markov chains and non-gaussian time series', *Statistical Science*, Vol. 17, No. 3, pp.328–356.

Bettini, C., Wang, X.S. and Jajodia, S. (1996) 'Testing complex temporal relationships involving multiple granularities and its application to data mining (extended abstract)', in *PODS '96: Proceedings of the 15th ACM SIGACT SIGMOD-SIGART Symposium on Principles of Database Systems*, ACM Press, New York, pp.68–78.

Billari, F.C. (2005) 'Life course analysis: two (complementary) cultures? Some reflections with examples from the analysis of transition to adulthood, in Levy, R., Ghisletta, P., Le Goff, J-M., Spini, D. and Widmer, E. (Eds.) (2005) *Towards an Interdisciplinary Perspective on the Life Course*, Advances in Life Course Research, Elsevier, Amsterdam, Vol. 10, pp.267–288.

Billari, F.C., Fürnkranz, J., and Prskawetz, A. (2006) 'Timing, sequencing and quantum of life course events: a machine learning approach', *European Journal of Population*, Vol. 22, No. 1, pp.37–65.

Blockeel, H., Fürnkranz, J., Prskawetz, A. and Billari, F. (2001) 'Detecting temporal change in event sequences: an application to demographic data', in De Raedt, L. and Siebes, A. (Eds.): *Principles of Data Mining and Knowledge Discovery: 5th European Conference, PKDD 2001*, Springer, Freiburg in Brisgau, Vol. LNCS 2168, pp.29–41.

Blossfeld, H-P. and Rohwer, G. (2002) *Techniques of Event History Modeling, New Approaches to Causal Analysis*, 2nd ed., Lawrence Erlbaum, Mahwah, NJ.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*, Chapman and Hall, New York.

Brzinsky-Fay, C., Kohler, U. and Luniak, M. (2006) 'Sequence analysis with Stata', *The Stata Journal*, Vol. 6, No, 4, pp.435–460.

Ciampi, A. (1991) 'Generalized regression trees', *Computational Statistics and Data Analysis*, Vol. 12, No. 1, pp.57–78.

Ciampi, A., Negassa, A. and Lou, Z. (1995) 'Tree-structured prediction for censored survival data and the Cox model', *Journal of Clinical Epidemiology*, Vol. 48, No. 5, pp.675–689.

Courgeau, D. and Baccaïni, B. (1998) 'Multilevel analysis in the social sciences', *Population: An English Selection*, Vol. 10, No. 1, pp.39–71.

Cox, D.R. (1972) 'Regression models and life-tables', *Journal of the Royal Statistical Society, Series B*, Vol. 34, No. 2, pp.187–220.

Dannegger, F. (2000) 'Tree stability diagnostics and some remedies for instability', *Statistics in Medicine*, Vol. 19, No. 4, pp.475–491.

De Rose, A. and Pallara, A. (1997) 'Survival trees: an alternative non-parametric multivariate technique for life history analysis', *European Journal of Population*, Vol. 13, pp.223–241.

Elzinga, C.H. (2003) 'Sequence similarity: a non-aligning technique', *Sociological Methods and Research*, Vol. 31, pp.214–231.

Gabadinho, A., Ritschard, G., Studer, M., and Müller, N.S. (2008) *Mining Sequence Data in R with TraMineR*, User's guide, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva (released on CRAN the comprehensive R archive network).

Gauthier, J-A., Widmer, E.D., Bucher, P. and Notredame, C. (2008) 'How much does it cost? Optimization of costs in sequence analysis of social science data', *Sociological Methods and Research*, (forthcoming).

Hogan, D.P. (1978) 'The variable order of events in the life course', *American Sociological Review*, Vol. 43, pp.573–586.

Hosmer, D.W. and Lemeshow, S. (1999) *Applied Survival Analysis, Regression Modeling of Time to Event Data*, Wiley, New York.

Hothorn, T., Hornik, K. and Zeileis, A. (2006a) 'Party: a laboratory for recursive part(y)itioning', User's manual.

Hothorn, T., Hornik, K. and Zeileis, A. (2006b) 'Unbiased recursive partitioning: a conditional inference framework', *Journal of Computational and Graphical Statistics*, Vol. 15, No. 24, pp.651–674.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data*, Springer, New York.

Huang, X., Chen, S. and Soong, S. (1998) 'Piecewise exponential survival trees with time-dependent covariates', *Biometrics*, Vol. 54, pp.1420–1433.

Joshi, M.V., Karypis, G. and Kumar, V. (2001) 'A universal formulation of sequential patterns', in *Proceedings of the KDD'2001 Workshop on Temporal Data Mining*, August 2001, San Fransisco.

Kass, G.V. (1980) 'An exploratory technique for investigating large quantities of categorical data', *Applied Statistics*, Vol. 29, No. 2, pp.119–127.

Kim, H. and Loh, W-Y. (2001) 'Classification trees with unbiased multiway splits', *Journal of the American Statistical Association*, Vol. 96, No. 454, pp.589–604.

Leblanc, M. and Crowley, J. (1992) 'Relative risk trees for censored survival data', *Biometrics*, Vol. 48, pp.411–425.

Leblanc, M. and Crowley, J. (1993) 'Survival trees by goodness of split', *Journal of the American Statistical Association*, Vol. 88, No. 422, pp.457–467.

Levenshtein, V. (1966) 'Binary codes capable of correcting deletions, insertions and reversals', *Soviet Physics Doklady*, Vol. 10, pp.707–710.

Levy, R., Gauthier, J-A. and Widmer, E.D. (2006) 'Entre contraintes institutionnelle et domestique: les parcours de vie masculins et féminins en suisse', *Revue Canadienne de Sociologie*, Vol. 31, No. 4, pp.461–489.

Lillard, L.A. (1993) 'Simultaneous equations for hazards: marriage duration and fertility timing', *Journal of Econometrics*, Vol. 56, pp.189–217.

Loh, W-Y. (2007) GUIDE (version 5) User manual, Technical report, Department of Statistics, University of Wisconsin, Madison.

Malo, M.A. and Munoz, F. (2003) 'Employment status mobility from a lifecycle perspective: a sequence analysis of work-histories in the BHPS', *Demographic Research*, Vol. 9, No. 7, pp.471–494.

Mannila, H., Toivonen, H. and Verkamo, A.I. (1997) 'Discovery of frequent episodes in event sequences', *Data Mining and Knowledge Discovery*, Vol. 1, No. 3, pp.259–289.

McVicar, D. and Anyadike-Danes, M. (2002) 'Predicting successful and unsuccessful transitions from school to work using sequence methods', *Journal of the Royal Statistical Society A*, Vol. 165, No. 2, pp.317–334.

Müller, N.S., Lespinats, S., Ritschard, G., Studer, M. and Gabadinho, A. (2008) 'Visualisation et classification des parcours de vie', *Revue des Nouvelles Technologies de l'Information RNTI*, Vol. 11, pp.499–510.

Needleman, S. and Wunsch, C. (1970) 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of Molecular Biology*, Vol. 48, pp.443–453.

Paliwal, K.K. (1993) 'Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer', *Proceedings ICASSP*, Vol. 2, pp.215–218.

Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo.

Rabiner, L.R. (1989) 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE*, Vol. 77, No. 2, pp.257–286.

Raftery, A.E. and Tavaré, S. (1994) 'Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model', *Applied Statistics*, Vol. 43, pp.179–199.

Ritschard, G. and Oris, M. (2005) 'Life course data in demography and social sciences: statistical and data mining approaches', in Levy, R., Ghisletta, P., Le Goff, J-M., Spini, D. and Widmer, E. (Eds.) (2005) *Towards an Interdisciplinary Perspective on the Life Course*, Advances in Life Course Research, Elsevier, Amsterdam, Vol. 10, pp.289–320.

Rohwer, G. and Pötter, U. (2002) TDA user's manual, Software, Ruhr-Universität Bochum, Fakultät für Sozialwissenschaften, Bochum.

Sankoff, D. and Kruskal, J.B. (Eds.) (1983) *Time Warps, String Edits and Macro-Molecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, Reading.

Scherer, S. (2001) 'Early career patterns: a comparison of Great Britain and West Germany', *European Sociological Review*, Vol. 17, No. 2, pp.119–144.

Segal, M.R. (1988) 'Regression trees for censored data', *Biometrics*, Vol. 44, pp.35–47.

Segal, M.R. (1992) 'Tree-structured methods for longitudinal data', *Journal of the American Statistical Association*, Vol. 87, No. 418, pp.407–418.

Srikant, R. and Agrawal, R. (1996) 'Mining sequential patterns: generalizations and performance improvements', in Apers, P.M.G., Bouzeghoub, M. and Gardarin, G. (Eds.): *Advances in Database Technologies – 5th International Conference on Extending Database Technology (EDBT'96)*, Springer-Verlag, Avignon, France, Vol. 1057, pp.3–17.

Su, X. and Tsai, C-L. (2005) 'Tree-augmented Cox proportional hazards models', *Biostat*, Vol. 6, No. 3, pp.486–499.

Tarone, R.E. and Ware, J. (1977) 'On distribution-free tests for equality of survival distributions', *Biometrika*, Vol. 64, No. 1, pp.156–160.

Therneau, T.M. and Atkinson, E.J. (1997) 'An introduction to recursive partitioning using the RPART routines', Technical Report Series 61, Mayo Clinic, Section of Statistics, Rochester, Minnesota.

Therneau, T.M. and Grambsch, P.M. (2000) *Modeling Survival Data*, Springer, New York.

Yamaguchi, K. (1991) *Event History Analysis*, ASRM 28, Sage, Newbury Park and London.

Zaki, M.J. (2000) 'Sequence mining in categorical domains: incorporating constraints', in *9th International Conference on Information and Knowledge Management (CIKM 2000)*, 9–11 November, McLean, VA, ACM Press, New York, pp.422–429.