Mining event histories: A social scientist view

Gilbert Ritschard[1]

[1] Dept of Econometrics, University of Geneva, 40 bd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland

Keywords: Event histories, Sequences, Mining frequent episodes, Survival trees.

Abstract

Individual longitudinal or sequence data are common to many fields. For instance, they are essential for understanding and predicting the evolution of a patient's disease after it has been diagnosed (survival analysis), the behavior of a visitor of a web site (web log mining), but also for categorizing or clustering signal sequences in domains such as telecommunication. This paper focuses on the analysis of individual longitudinal data within social sciences, especially in population science where we are interested in describing and understanding life courses. A life event can be seen as the change of state of some discrete variable, e.g. the marital status, the number of children, the job, the place of residence. Such life history data are collected in mainly two ways: As a collection of time stamped events or as state sequences. The former is used for instance by survival analysis that focuses on a given type of event and is concerned with its hazard rate or equivalently the duration until it happens. Sequence analysis on the other hand is concerned with the sequencing of the events and is best suited for characterizing whole life trajectories. We consider using data-mining-based approaches borrowed from other fields for analysing life courses with both a survival and a sequence perspective. We put stress on the social scientist's expectations and address some of the statistical challenges they raise.

1 Introduction

An individual life course paradigm emerged during the 80's 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 procreating, 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 organize their personal life courses. Much effort has been put for collecting individual longitudinal data. Many countries conduct today 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 and may be affected by other factors. We shall call them the survival methods. They include the well known Kaplan-Meier 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 for instance discrete Markov models and optimal-matching-based clustering. The aim of the paper is to make an overview of these methods with a special emphasize on non parametric heuristic data-mining-based approaches.

Our presentation will be organized as follows. We start in Section 2 by shortly discussing alternative representations of life course data. In Section 3 we propose a typology of methods for life course data distinguishing between survival and sequence methods, but also between descriptive and causal approaches, between parametric and non parametric models. In Section 4 we present survival trees in some details while Section 5 is devoted to the mining of typical sequential patterns and rules connecting episodes. Finally, we make some concluding remarks in Section 6.

2 Time to Event and State Sequence Views

There are different ways of organizing event histories data and each method may require a specific organization. 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 realization 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
civil 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 data and those that deal with sequences.

3.1 Statistical and data analysis methods

Methods most often used by social scientists are concerned with the duration 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 of not experiencing the second event before a duration t. We can distinguish

descriptive methods that just attempt to describe the survival function, and causal or explanatory methods used to investigate the factors that may influence the survival curves.

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 socioprofessional status, for example, influence the familial life course (time of marriage, number and timing of children)? 3) What are the effects of 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 summarizes 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 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 semiparametric 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.

Clustering based on the edit distance (Levenshtein, 1966; Needleman and Wunsch, 1970; Sankoff and Kruskal, 1983) between each pair of sequences has been popularized in social sciences by Abbott (see Abbott and Tsay, 2000) under the name of optimal matching and was for example exploited by Malo and Munoz (2003), Levy et al. (2006), Joye and Bergman (2004) and Lesnard (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

TABLE 3. A typology of methods for life course data

	nature of data	
questions	time stamped event	state/event sequences
descriptive	- Survival curves: Parametric (Weibull, Gompertz) and non parametric (Kaplan-Meier, Nelson-Aalen) estimators.	- Optimal matching clustering - Frequencies of typical patterns - *Discovering typical episodes*
causality	- Hazard regression models - *Survival trees*	- Markov models, *Mobility trees* - *Association rules* between episodes

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. Such representations are judiciously complemented with Index plots (Brzinsky-Fay et al., 2006) depicting the variability of individual trajectories inside each cluster. Optimal matching clustering can be realized for instance with free softwares such as TDA (Rohwer and Pötter, 2002) and SALTT (Notredame et al., 2006), or with the SQ package (Brzinsky-Fay et al., 2006) for Stata. Recent developments regarding optimal matching include training procedures for learning 'optimal' state substitution costs (Gauthier et al., 2007a) and multichannel approaches (Gauthier et al., 2007b). 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 modeling 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 modeling of high order process (Raftery and Tavaré, 1994; Berchtold and Raftery, 2002), Hidden Markov Models, HMM, (Rabiner, 1989) and their generalization 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 and the results provided remain hard to interpret when we departure from very simple specifications.

3.2 Data-mining-based approaches

Data mining is mainly concerned with the characterization of interesting patterns, either per se (unsupervised learning) or for a classification or prediction purpose (supervised learning). Unlike the statistical modeling 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 itemsets may be used to detect temporal changes in event sequences frequency from the Austrian Family and Fertility Survey (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. We initiated ourselves (Ritschard and Oris, 2005) 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 either as 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 statement 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

We briefly explain hereafter the main splitting criteria used for survival trees. Note that since such trees are intended for dealing with censored data, the usual minimal node size constraints may be completed with additional constraints on the minimal number of events that should occur in each node.

4.1 Splitting criteria

As for classical classification trees, there are two main groups of splitting criteria: Those that attempt to maximize the group difference in the spirit of CHAID (Kass, 1980) and other earlier tree growing methods, and those that maximize 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

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

where d_{i1} is number of events (death) observed in the first group (node) at each time t_i where at least one event occurs, 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 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 maximizes 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): $\lambda_h(t) = \theta_h\lambda_0(t)$. Estimation 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 maximizing the reduction in deviance amounts to maximize group homogeneity. Hence this approach is more in line with classical tree growing algorithms such as CART or C4.5, which attempt to maximize 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 maximize 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 maximization. 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).

Martingale-based residuals.

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

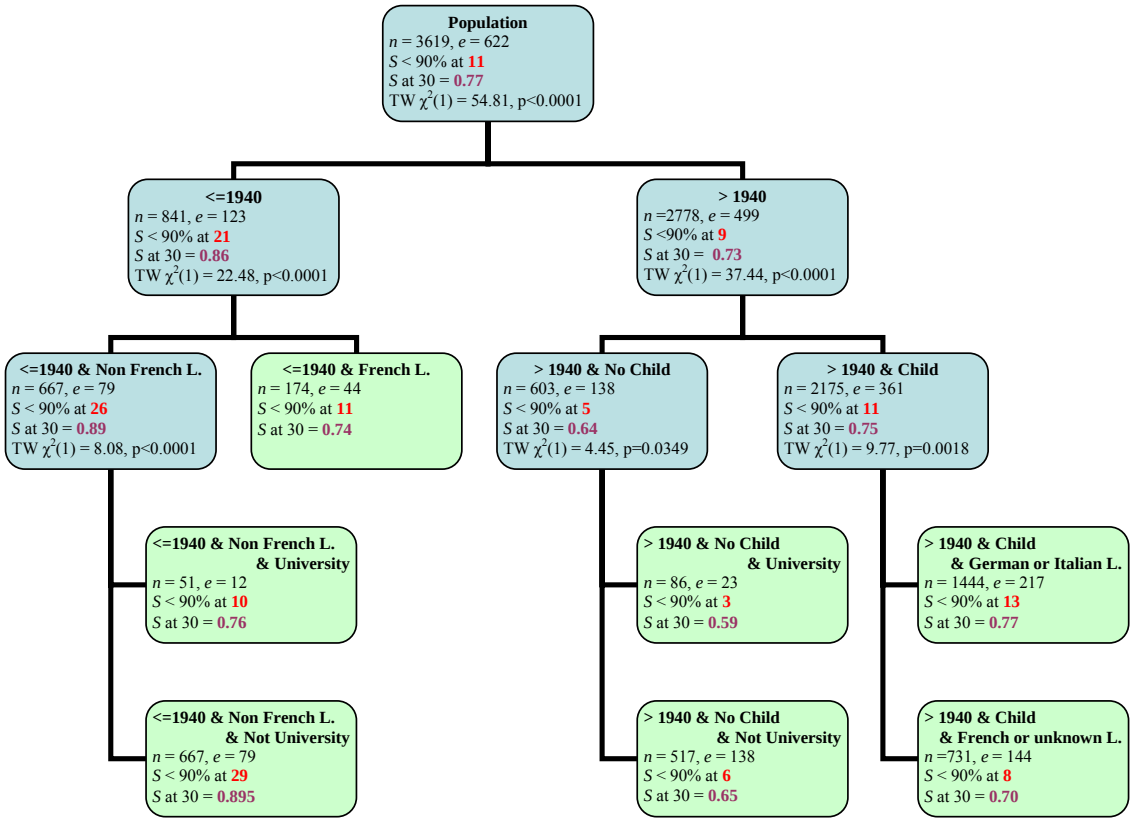

FIGURE 1. Survival tree for marriage duration until Divorce/Separation (Tarone-Ware criterion)

one for which the residuals look the less random. For measuring randomness, residuals are split into those above their median values and the other ones. The randomness measure is then just 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.

4.2 Illustration

Figure 1 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 Swiss Household Panel (SHP) in 2002. The criterion used consisted in maximizing the differences between Kaplan-Meier survival curves using the significance of the Tarone-Ware Test. A 5% significance limit was used 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 Kaplan-Meier survival curves corresponding to the 7 leaves (terminal nodes) of the tree are depicted in Figure 2.

It results clearly from this 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 9. We notice also that if 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

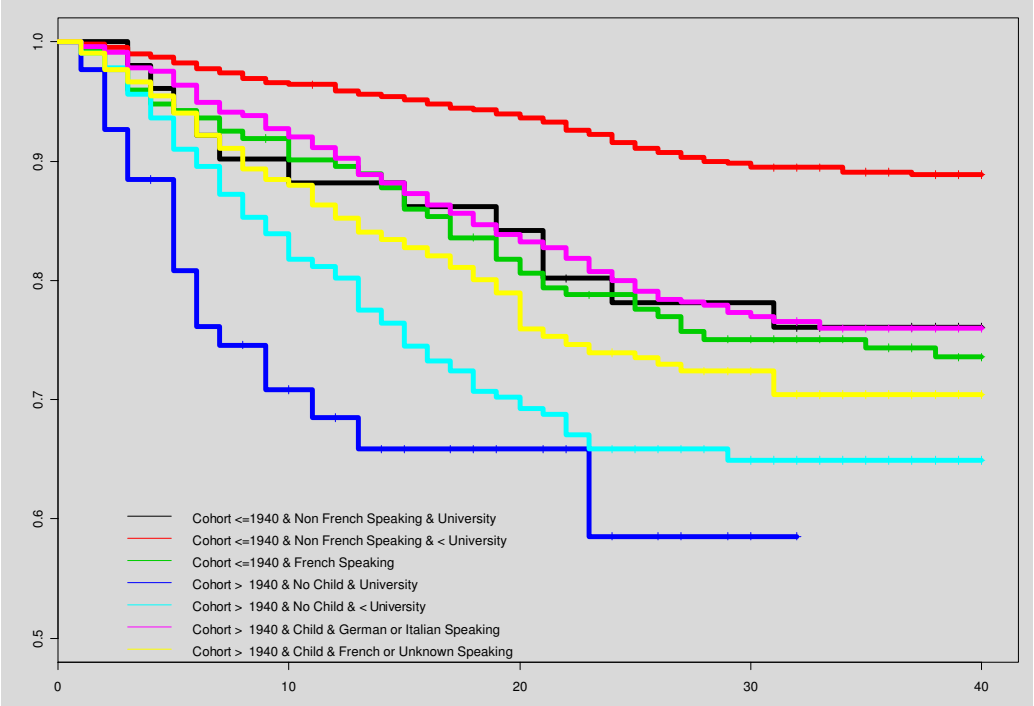

FIGURE 2. Survival tree, the 7 resulting KM survival curves

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 (1992)'s approach, we obtain a somewhat simpler tree corresponding to the first two levels of the tree obtained with the Tarone-Ware criterion. From the (not shown) 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.

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 child birth 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 of the difficulty in formulating interpretable splits preserving simultaneously ordering with respect to both the time and the predictor itself. Segal (1992) discusses a few possibilities, concluding, however, than 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 organization 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 Swiss Household Panel, 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 3 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.

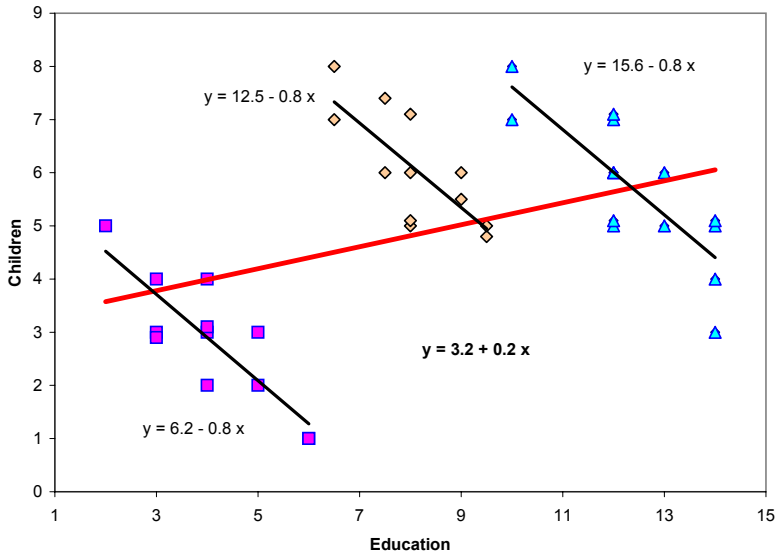

FIGURE 3. Multi-level: A simple linear regression example with 3 clusters

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.

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 characterize 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 under point 3.2, provide valuable holistic knowledge in the form of categorization of whole life courses.

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 web site (Zaki, 2000). Different approaches for characterizing what 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 specialized case of the mining of frequent itemsets, 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 itemsets 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 are 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 10 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, 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.

5.2 Sequential relationships

Beside 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) and 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 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 node, 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. For instance, assume we are interested in finding typical episodes of professional and education events that occur between leaving home (LH) and the first childbirth (C1). We would then set the first and last node as depicted in Figure 4 and look for the possible content of the in between nodes. The dashed edge in Figure 4 indicates an *elastic edge*, i.e. one that can be extended by adding nodes according to the discovery process.

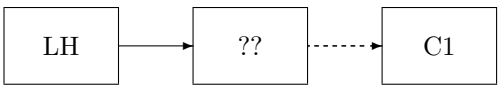

FIGURE 4. A sequential structure with node constraints

Leaving no free node, we characterize 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 5 shows the distribution among three alternatives structures for the following couples 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 couple (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 5 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.

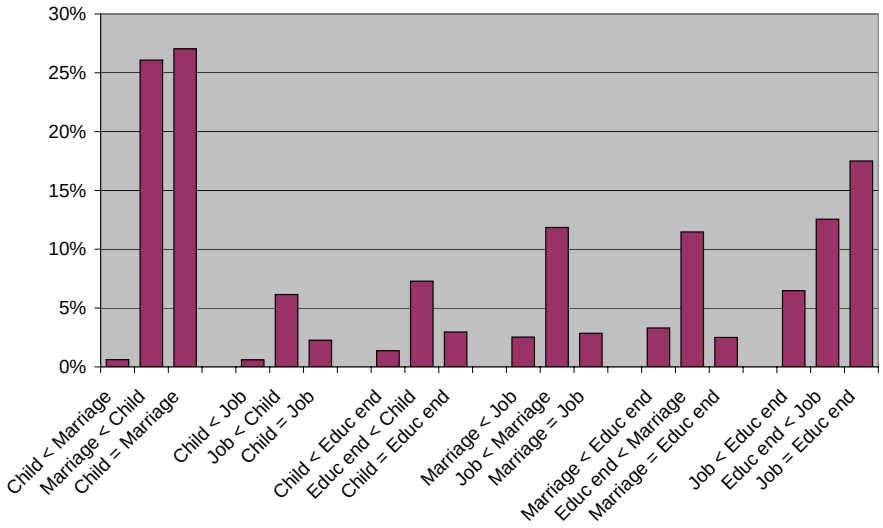

FIGURE 5. Distribution of alternative structures of 2-event episodes

5.3 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 favor given behaviors (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 itemsets 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 itemsets. Each of them will, however, result in variants depending on the counting method and various time constraints retained.

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 the one of her husband. 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 Conclusion

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 synthesized 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 emphasize 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 me 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, should be directly applicable to life course data.

Acknowledgments: This study has been realized within the Swiss National Science Foundation project SNSF 100012-113998/1. The empirical results are based on data collected within the "Living in Switzerland: 1999-2020" project steered by the Swiss Household Panel (www.swisspanel.ch) of the University of Neuchâtel and the Swiss Statistical Office.

References

Abbott, A. (1990). A primer on sequence methods. *Organization Science 1*(4), 375–392.

Abbott, A. and A. Tsay (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research 29*(1), 3–33. (With discussion, pp 34-76).

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

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

Berchtold, A. (2002). High-order extensions of the double chain Markov model. *Stochastic Models 18*(2), 193–227.

Berchtold, A. and A. Berchtold (2004). MARCH 2.02: Markovian model computation and analysis. User's guide, www.andreberchtold.com/march.html.

Berchtold, A. and A. E. Raftery (2002). The mixture transition distribution model for high-order Markov chains and non-gaussian time series. *Statistical Science 17*(3), 328–356.

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

Billari, F. C. (2005). Life course analysis: Two (complementary) cultures? Some reflections with examples from the analysis of transition to adulthood. In Ghisletta et al. (2005), pp. 267–288.

Billari, F. C., J. Fürnkranz, and A. Prskawetz (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population 22*(1), 37–65.

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

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

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

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

Ciampi, A. (1991). Generalized regression trees. *Computational Statistics and Data Analysis 12*(1), 57–78.

Ciampi, A., A. Negassa, and Z. Lou (1995). Tree-structured prediction for censored survival data and the Cox model. *Journal of Clinical Epidemiology 48*(5), 675–689.

Courgeau, D. and B. Baccaïni (1998). Multilevel analysis in the social sciences. *Population: An English Selection 10*(1), 39–71.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B 34*(2), 187–220.

Dannegger, F. (2000). Tree stability diagnostics and some remedies for instability. *Statistics In Medicine 19*(4), 475–491.

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

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

Gauthier, J.-A., E. D. Widmer, P. Bucher, and C. Notredame (2007a). How much does it cost? Optimization of costs in sequence analysis of social science data. Manuscript, University of Lausanne. (Under review).

Gauthier, J.-A., E. D. Widmer, P. Bucher, and C. Notredame (2007b). Multichannel sequence analysis applied to social science data. Manuscript, University of Lausanne. (Under review).

Ghisletta, P., J.-M. Le Goff, R. Levy, D. Spini, and E. Widmer (Eds.) (2005). *Towards an Interdisciplinary Perspective on the Life Course*. Advancements in Life Course Research, Vol. 10. Amsterdam: Elsevier.

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

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

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

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

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

Joye, D. et M. Bergman (2004). Carrières professionnelles : une analyse biographique. In E. Zimmermann et R. Tillmann (Eds.), *Vivre en Suisse 1999-2000 : une année dans la vie des ménages et familles en Suisse*, Volume 3 of *Collection Population, famille et société*, pp. 77–92. Bern : Peter Lang.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

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

Leblanc, M. and J. Crowley (1993). Survival trees by goodness of split. *Journal of the American Statistical Association 88*(422), 457–467.

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

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

Levy, R., J.-A. Gauthier, et E. D. Widmer (2006). Entre contraintes institutionnelle et domestique : les parcours de vie masculins et féminins en suisse. *Revue canadienne de sociologie 31*(4), 461–489.

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

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

Malo, M. A. and F. Munoz (2003). Employment status mobility from a lifecycle perspective: A sequence analysis of work-histories in the BHPS. *Demographic Research 9*(7), 471–494.

Mannila, H., H. Toivonen, et A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*(3), 259–289.

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

Notredame, C., P. Bucher, J.-A. Gauthier, and E. D. Widmer (2006). T-COFFEE/SALTT: User guide and reference manual. Technical report, CNRS Marseille and PAVIE University of Lausanne. (available at http://www.tcoffee.org/saltt/).

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

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

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

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

Ritschard, G. and M. Oris (2005). Life course data in demography and social sciences: Statistical and data mining approaches. In Ghisletta et al. (2005), pp. 289–320.

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

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

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

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association 87*(418), 407–418.

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

Su, X. and C.-L. Tsai (2005). Tree-augmented Cox proportional hazards models. *Biostat 6*(3), 486–499.

Tarone, R. E. and J. Ware (1977). On distribution-free tests for equality of survival distributions. *Biometrika 64*(1), 156–160.

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

Yamaguchi, K. (1991). *Event history analysis*. ASRM 28. Newbury Park and London: Sage.

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