Gender and Well-Being

Interactions between Work, Family and Public Policies

COST ACTION A 34

Second Symposium:

The Transmission of Well-Being: Marriage Strategies and Inheritance Systems in Europe (17th-20th Centuries)

25th -28th April 2007

University of Minho
Guimarães-Portugal

Please, do not quote without author's permission

Comparing and classifying personal life courses: From time to event methods to sequence analysis*

Gilbert Ritschard
Matthias Studer, Nicolas Müller and Alexis Gabadinho

Dept of Econometrics and Laboratory of Demography and Family Studies
University of Geneva, 40 bd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
gilbert.ritschard@metri.unige.ch

Abstract

This paper is mainly methodological. It is concerned with the different ways me may analyse personal life course data. Personal life courses are defined by a succession of events regarding living arrangement, familial life, education, professional career, health, etc. We may focus on one of these events — leaving home, marriage, first job, divorce, becoming disabled — and examine how the hazard of experiencing it evolves with time and may be affected by other factors or events. Alternatively, we may be interested, in a more holistic way, in analysing and comparing whole sequences. The paper surveys the main available methods and classifies them into a typology that distinguishes between the nature of data — time stamped events or sequences — they need and the kind of questions — descriptive or causal — they address. Both classical statistical methods and promising but less known data-mining-based approaches are discussed. The aim of the paper is to put these approaches into perspective by focusing on their specificity and the complementary views they bring on life courses. Three illustrations using data from the Swiss Household Panel will show the kind of results we may expect from some of the less known methods: The first concerns sex differences in working status mobility, the second is a survival tree analysis of the risk to divorce, while the third focuses in sex differences in the sequencing of a selection of young adults life events.

Keywords: Life course data, sequence analysis, event history analysis, induction trees, survival trees, mining frequent sequences and associations rules, social mobility.

*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.

1 Introduction

Well-Being is neither static, nor macro, but rather relies strongly on the dynamics of life courses. Individual factors (sex, birth cohort, parent's social status, ...), earlier events (leaving home, first union, completed education, first job, ...) as well as the sequencing of these earlier events undoubtedly jointly influence the chances of accessing Well-Being. Hence, we need to resort to suited techniques for discovering related 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 clustering. The aim of the paper is to make an overview of these methods with a special emphasize on less known non parametric heuristical data-mining-based approaches.

Our presentation will be organized as follows. We start in Section 2 with a life course data representation issue: Survival methods consider mainly time stamped events while sequence methods require indeed sequences as input. We show that these are indeed just alternative representations of the same life courses. 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. We then illustrate in Section 4 the scope of three less known heuristical methods. The illustrations are based on data from the Swiss Household Panel (SHP). The first one is an analysis of the working status dynamics by means of a mobility tree, the second one is an analysis of the risk to divorce with a survival tree, and the third is concerned with sex differences in the sequencing of a selection of young adult life events. Finally, we conclude in Section 5 by stressing, among others, limits of the tree approaches discussed.

2 Time to Event and Sequence Views

Methods for event histories data depend on the nature of the data and the questions we are interested in. So let us first clarify these points.

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 histories 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 events of each individual are represented by the sequence of states of the variables of interest. Panel data are special cases of state sequences where the states are observed at periodic time.

Table 1: Time stamped event view

ending secondary school in 1970	first job in 1971	marriage in 1973

Table 2: State sequence view

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

Table 3: Spell view

id	from	to	civil status	education	job
id1	1969	1969	single	primary	no
id1	1970	1970	single	secondary	no
id1	1971	1972	single	secondary	first
id1	1973	1973	married	secondary	first

Basically it is always possible to transform time stamped data into sequences of either events or states, and reciprocally sequences — at least state sequences — may be transformed into time stamped events. It is sometimes also useful to put the data into spell view (Table 3) or in person-period form.

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 for time stamped data are mainly concerned with the duration between two specific events, birth and leaving home, first union and first child, for example.

Table 4: 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 patterns*
causality	- Hazard regression models - *Survival trees*	- Markov models, *Mobility trees* - *Association rules* between subsequences

They try to answer questions about the distribution of the "survival" probabilities, i.e. the probabilities of not experiencing the 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 4. 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 Allison (1984), Yamaguchi (1991), Courgeau and Lelièvre (1993) 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 softwares 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 Abbott's optimal matching criteria (Abbott and Forrest, 1986; Abbott and Hrycak, 1990) has known some success and was for instance exploited by Malo and Munoz (2003), Widmer et al. (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 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 results.

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), 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. For a general introduction to data mining, see for instance Hand, Mannila, and Smyth (2001) for a statistically oriented presentation, or Han and Kamber (2001) for a database perspective.

Data-mining-based approaches have recently been 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 the survival trees (Segal, 1988; Leblanc and Crowley, 1992, 1993; Ahn and Loh, 1994; Ciampi et al., 1995; Huang et al., 1998). Their principle is based on that of classification and regression trees (Breiman et al., 1984; Quinlan, 1993; Kass, 1980) 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 in order 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 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 4. 1) Within the logic of "survival" methods, survival trees should complement regression like models by helping at discovering interaction effects between covariates. They will clearly exhibit differential effects like, for example, that 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. the typical subsequences, may then be used either as dependent or independent 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 like, 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 Some illustrations

We have clearly not the place to detail here all these possible approaches. We have chosen therefore to limit ourselves to illustrate the outcome of some of the less known data-mining-based methods only. We use data from the Swiss Household Panel Survey (SHP). We start with the idea of mobility tree that is the easiest to explain, then present results from survival trees and finally an analysis of subsequences of length two.

4.1 Mobility Trees

We present here a working status mobility analyses based on the individual records of the six first SHP waves (1999 to 2004). More specifically, we are interested in how the chances to be working active, unemployed or not in the labor force in 2004 depend on the working status of the previous years, and in how these relationships differ by sex and may be mediated by other covariates. For the target variable, i.e. the 2004 working status, we consider just three states, namely working active, unemployed and inactive. Active includes full time as well as part time employed, but also working in family enterprises, while inactive means not in the labor force and includes among others at home, student and retired.

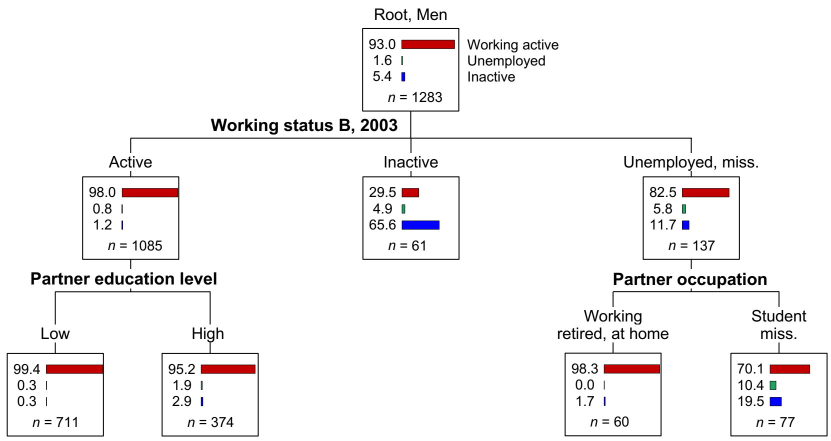

Figure 1: Mobility tree, Men

The mobility tree approach consists in growing a decision tree using the 2004 working status as target variable and to include the statuses of the previous years into the set of potential predictors, additional predictors being attained education level, partner working statuses in 2003 and 2004, and partner education level. Note that predictive working status variables are somewhat more disaggregated than the target variable in that the active category is split into full time (>80%), long part time and short time (<50%) active. A separate tree is grown for men and women.

The tree growing principle is as follows. Starting with the root node where all cases are grouped together, we look for the most discriminating factor among the predictor list and split the node according to its values. For example in Figure 1 the most discriminating factor has been found to be the working status of the previous year, 2003. We may notice here that despite we distinguished between full time and part time active in 2003, there was no significant distinction between these groups for predicting the 2004 working status. Hence we have only a split in 3 groups instead of 5 possible, full time, long and short part time statuses having been merged together. The procedure continues then iteratively by trying the same way to split the previously obtained nodes until some stopping rule is reached. We used here the CHAID (Kass, 1980) algorithm with a 5% significance limit and a minimal parent node size of 100.

Comparing the trees obtained for men and women (Figures 1 and 2), it appears clearly that this mobility process is strongly gendered. We see that men that are inactive in 2003 have a probability of about 66% to remain inactive the next year, while this probability is 77% for women. Likewise, men unemployed in 2003 have about 83% chances to be working active the next year against only 73% for women. These percentages however could have been obtained through simple cross tabulations. The most interesting aspects of trees, is their ability to detect relevant interactions, which we observe when we go to deeper levels. For instance, there is an interaction between the 2002 and 2003 status for women. Consider for example women that are not in the labor force in 2003. Their probability to remain inactive in 2004 largely depends on their 2002 status. Those who belonged to the labor force in 2002 have almost 50% chances to return to the labor force in 2004, while those who were already inactive in 2002 have 86% chances to remain inactive. This is most probably a maternity effect. Another interesting interaction effect is, in the tree for men, the one between being active in 2003 and partner's education. The chances to become inactive are about 3% when the partner has high education, while they are ten times less when the partner has low education.

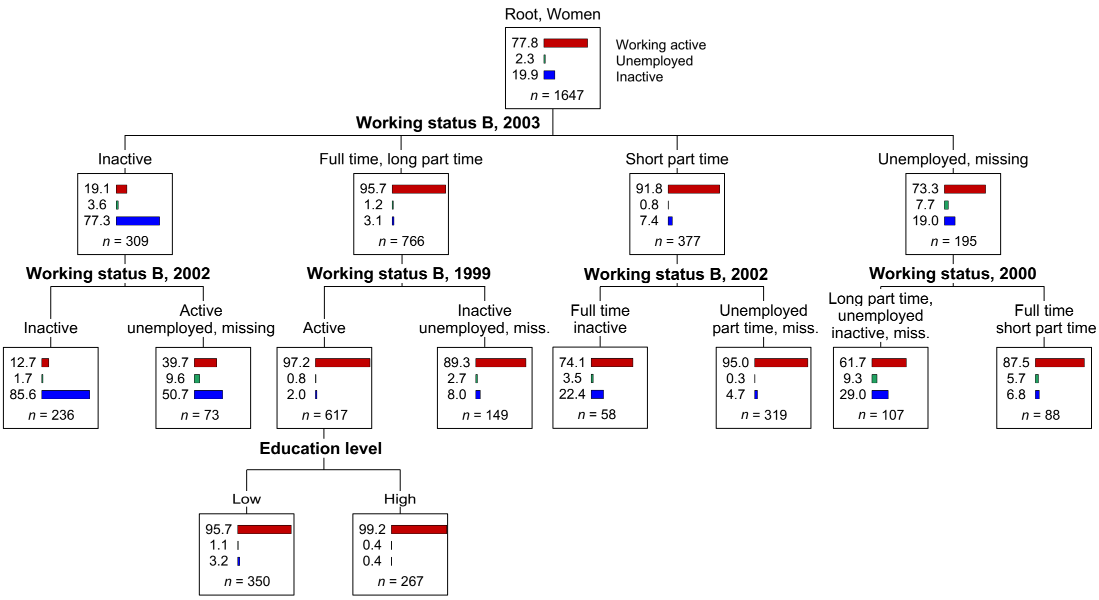

Figure 2: Mobility tree, Women

4.2 Survival Trees

The principle of survival tree is quite similar to that of classification trees used in the previous mobility analysis. The mere difference is that we are here interested in the survival curve of the groups rather than in the distribution of a categorical status variable. The aim is thus to segment the population into groups that differ as much as possible in their survival characteristics, rather than in a categorical distribution. Figure 3 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 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 cohorts, 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 4.

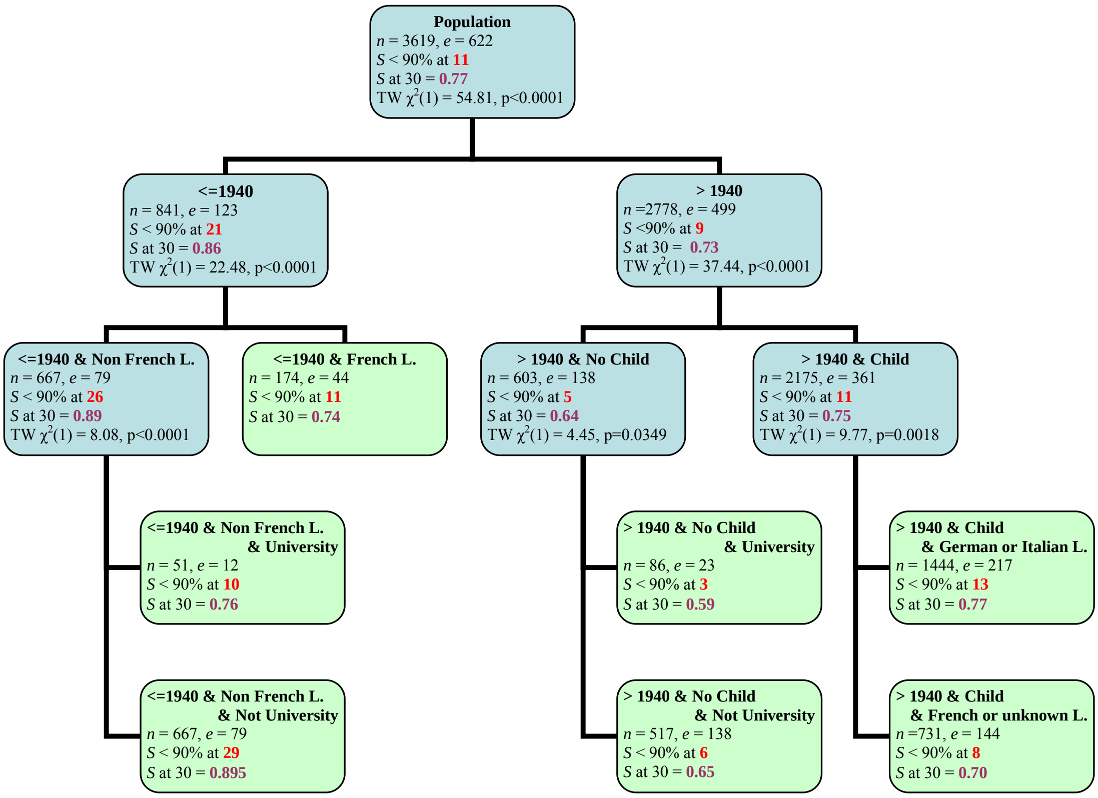

Figure 3: Survival tree for marriage duration until Divorce/Separation

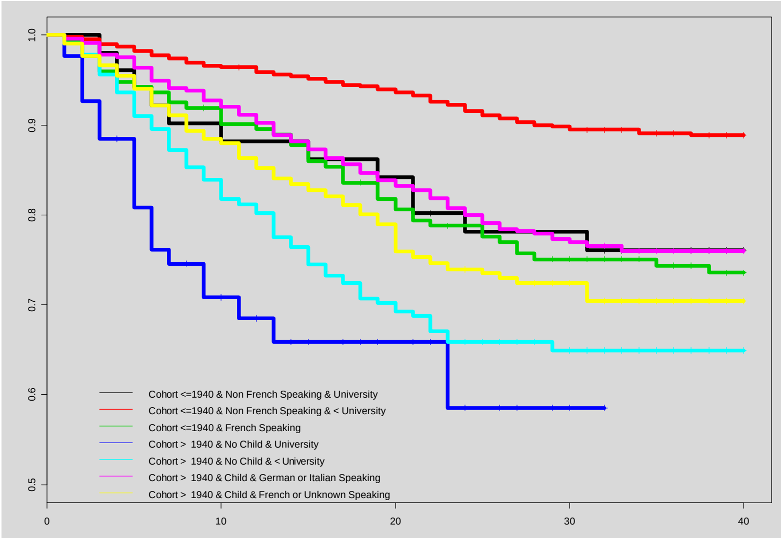

Figure 4: Survival tree, the 7 resulting survival curves

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 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.

4.3 Discriminating subsequences

We turn now to an analysis of the sequence order. We are interested in which pairs of events have the most significant different order between women and men.

Considering pairs of events $\{x, y\}$ such as for example leaving home and first job, we may distinguish three states for characterizing their sequencing: Event x happens before y, x and y happen at the same time, x happens after y. To these three states we have to add the case where the pair is not observed, i.e. when at least one of the event was not experienced by the concerned individual. This leaves us with the four states depicted in Table 5.

We assigned such a 4 state variable to each of the following couples of events {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}. Figure 5 exhibits the distributions of the non missing values for a selection of these variables. It shows that it is really exceptional — in 20th century Swiss life courses — to have a

Table 5: Four states for the sequencing of two events x and y

State for the couple (x, y)	condition	notation
x happens before y	$t_x < t_y$	$<$
x and y happen simultaneously	$t_x = t_y$	$=$
x happens after y	$t_x > t_y$	$>$
not observed	x and/or y missing	miss.

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.

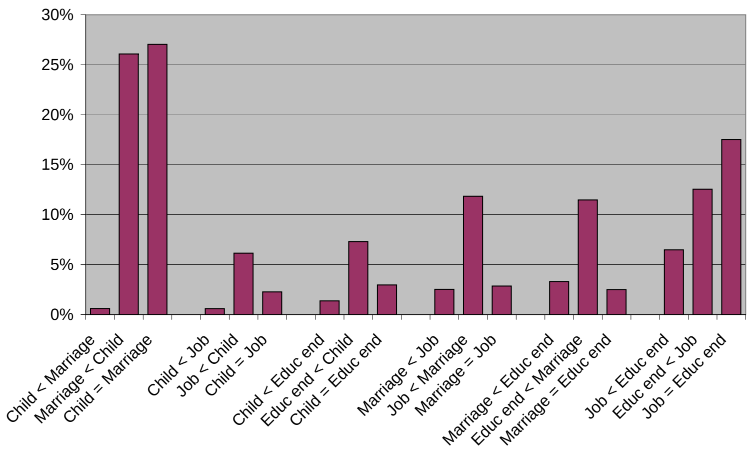

Figure 5: Frequencies of characteristic 2-event sequences

To find out the major differences in these sequencings that distinguish women from men, we may look at the classification tree in Figure 6. The target variable is sex and the predictors were the state variables associated to the pairs of events. The most discriminating pair is {Education End, 1st Job}, men having higher chances than women to start working before having ended education. Among those who start working before education has been completed, the proportion of men is even higher when we consider only those who start working before leaving their parent's home. If we look now at those who have their first job after ending school, we may note the large over representation of women among those who have a child before the first job. This — ending school and having a child before the first job — is indeed the profile with the highest women proportion. For men, the most characteristic profile is starting the first job before ending school and leaving home, and getting married when or after having completed education.

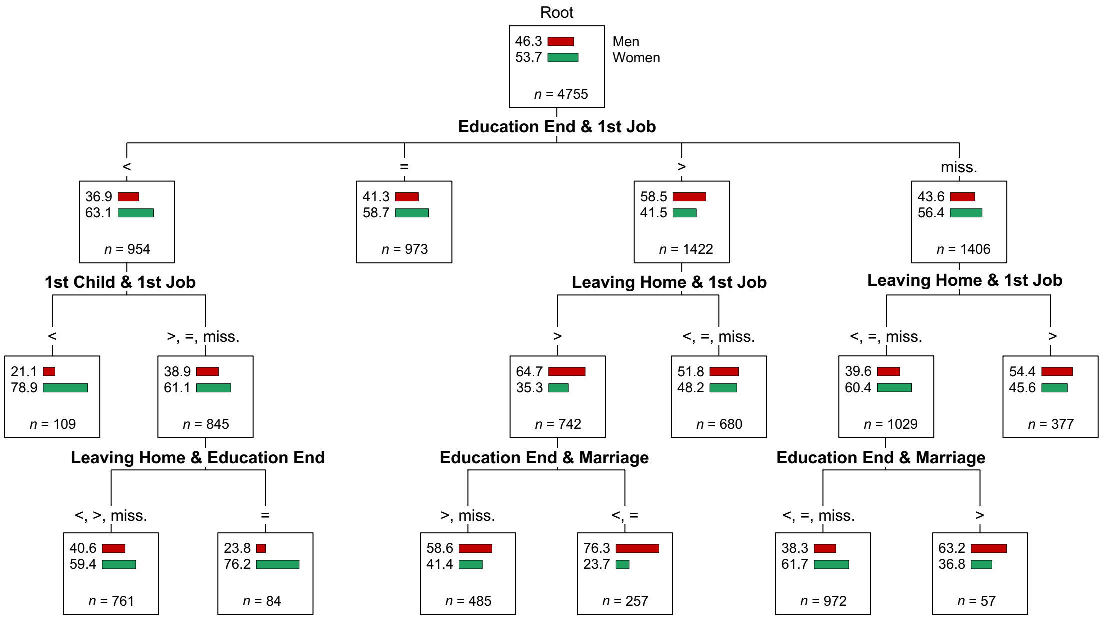

Figure 6: Discriminating sex with two event sequencing

5 Conclusion

We have seen that there are plenty of ways to look at individual longitudinal data, each way having its own advantages. The aim of the paper was to give a synthesized view of the available methods and to illustrate the kind of outcome we may expect from some less known data-mining-based techniques. We have especially put emphasize on tree methods which have two major advantages: First, their recursive splitting mechanism produce a tree structured comprehensive 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 — discrete or 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 already been investigated by (Dannegger, 2000). Another important issue is the validation of the tree. Classification trees are usually validated by measuring their classification performance. In our settings however, we are not interested in classification, but in the description provided by the tree. There is here also a need for better suited criteria. The deviance we proposed in (Ritschard, 2006) is a first solution. Beside trees, methods for mining sequence analyses, especially for finding the most typical subsequences of events and relationships between such subsequences are perhaps those from which me may expect the most highlighting holistic views on life courses. We have not yet experimented them, but hope to be able to demonstrate their usefulness for exploring individual life course data within the next months.

References

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

ABBOTT, Andrew and FORREST, John (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

ABBOTT, Andrew and HRYCAK, Alexandra (1990). Measuring resemblance in sequence data: An optimal matching analaysis of musician's carrers. *American Journal of Sociolgy 96*(1), 144–185.

ABBOTT, Andrew and TSAY, Angela (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, Hongshik and LOH, Wei-Yin (1994). Tree-structured proportional hazards regression modeling. *Biometrics 50*, 471–485.

ALLISON, Paul D. (1984). *Event history analysis, Regression for longitudinal event data*. QASS 46. Beverly Hills and London: Sage.

BARBER, Jennifer S., MURPHY, Susan A., AXINN, William G., and MAPLES, Jerry (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, André and BERCHTOLD, Annick (2004). MARCH 2.02: Markovian model computation and analysis. User's guide, www.andreberchtold.com/march.html.

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

BILLARI, Francesco 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, Francesco C., FÜRNKRANZ, Johannes, and PRSKAWETZ, Alexia (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population 22*(1), 37–65.

BLOCKEEL, H., FÜRNKRANZ, J., PRSKAWETZ, A., and BILLARI, F.C. (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, Hans-Peter and Rohwer, Götz (2002). *Techniques of Event History Modeling, New Approaches to Causal Analysis* (2nd ed.). Mahwah NJ: Lawrence Erlbaum.

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

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

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

Courgeau, Daniel and Lelièvre, Éva (1993). *Event History Analysis in Demography*. Oxford: Clarendon Press.

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

Dannegger, Felix (2000). Tree stability diagnostics and some remedies for instability. *Statistics In Medicine 19*(4), 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 13*, 223–241.

Ghisletta, Paolo, Le Goff, Jean-Marie, Levy, René, Spini, Dario, and Widmer, Eric (Eds.) (2005). *Towards an Interdisciplinary Perspective on the Life Course*. Advancements in Life Course Research, Vol. 10. Amsterdam: Elsevier.

Han, Jiawei and Kamber, Micheline (2001). *Data Mining: Concept and Techniques*. San Francisco: Morgan Kaufmann.

Hand, David J., Mannila, Heikki, and Smyth, Padhraic (2001). *Principles of Data Mining*. Adaptive Computation and Machine Learning. Cambridge MA: MIT Press.

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

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

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

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

Joye, Dominique et Bergman, Manfred Max (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 Crowley, J. (1992). Relative risk trees for censored survival data.

Biometrics 48, 411–425.

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

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

LEVY, René, GAUTHIER, Jacques-Antoine, et WIDMER, Eric D. (2006). Entre contraintes institutionnelle et domestique : les parcours de vie masculins et féminins en suisse. *Revue canadienne de sociologie*. (forthcoming).

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

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

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

RITSCHARD, Gilbert (2006). Computing and using the deviance with classification trees. In A. Rizzi and M. Vichi (Eds.), *COMPSTAT 2006 - Proceedings in Computational Statistics*, pp. 55–66. Berlin: Springer.

RITSCHARD, Gilbert and ORIS, Michel (2005). Life course data in demography and social sciences: Statistical and data mining approaches. In Ghisletta et al. (2005), pp. 289–320.

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

THERNEAU, Terry M. and GRAMBSCH, Patricia M. (2000). *Modeling Survival Data*. New York: Springer.

WIDMER, Eric D., LEVY, René, POLLIEN, Alexandre, and HAMMER, Raphaël (2003). Between standardisation, individualisation and gendering: An analysis of personal life courses in Switzerland. *Swiss Journal of Sociology 29*(1), 35–65.

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