Mining Event Histories: A Social Scientist View

Gilbert Ritschard

Department of Econometrics, University of Geneva
http://mephisto.unige.ch

IASC 2007, Aveiro, Portugal, August 30 - September 1

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES
Département d'économétrie

Outline

Outline

Motivation

- Individual life course paradigm.
  - Following macro quantities (e.g. #divorces, fertility rate, mean education level, ...) over time
    insufficient for understanding social behavior.
  - Need to follow individual life courses.
- Data availability
  - Large panel surveys in many countries
    (SHP, CHER, SILC, GGP, ...)
  - Biographical retrospective surveys (FFS, ...).
  - Statistical matching of censuses, population registers and other administrative data.

Motivation

- Individual life course paradigm.
  - Following macro quantities (e.g. #divorces, fertility rate, mean education level, ...) over time
    insufficient for understanding social behavior.
  - Need to follow individual life courses.
- Data availability
  - Large panel surveys in many countries
    (SHP, CHER, SILC, GGP, ...)
  - Biographical retrospective surveys (FFS, ...).
  - Statistical matching of censuses, population registers and other administrative data.

Motivation

- Need for suited methods for discovering interesting knowledge from these individual longitudinal data.
- Social scientists use
  - Essentially Survival analysis (Event History Analysis)
  - More rarely sequential data analysis (Optimal Matching, Markov Chain Models)
- Could social scientists benefit from data-mining approaches?
  - Which methods?
  - Are there specific issues with those methods for social scientists?

Motivation

- Need for suited methods for discovering interesting knowledge from these individual longitudinal data.
- Social scientists use
  - Essentially Survival analysis (Event History Analysis)
  - More rarely sequential data analysis (Optimal Matching, Markov Chain Models)
- Could social scientists benefit from data-mining approaches?
  - Which methods?
  - Are there specific issues with those methods for social scientists?

Motivation

- Need for suited methods for discovering interesting knowledge from these individual longitudinal data.
- Social scientists use
  - Essentially Survival analysis (Event History Analysis)
  - More rarely sequential data analysis (Optimal Matching, Markov Chain Models)
- Could social scientists benefit from data-mining approaches?
  - Which methods?
  - Are there specific issues with those methods for social scientists?

Alternative views of Individual Longitudinal Data

Table: Time stamped events, record for Sandra

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

Table: State sequence view, Sandra

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

Alternative views of Individual Longitudinal Data

Table: Time stamped events, record for Sandra

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

Table: State sequence view, Sandra

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

Issues with life course data

- Incomplete sequences
  - Censored and truncated data:
    Cases falling out of observation before experiencing an event of interest.
  - Sequences of varying length.
- Time varying predictors.
  - Example: When analysing time to divorce, presence of children is a time varying predictor.
- Data collected by clusters
  - Example: Household panel surveys.
  - Multi-level analysis to account for unobserved shared characteristics of members of a same cluster.

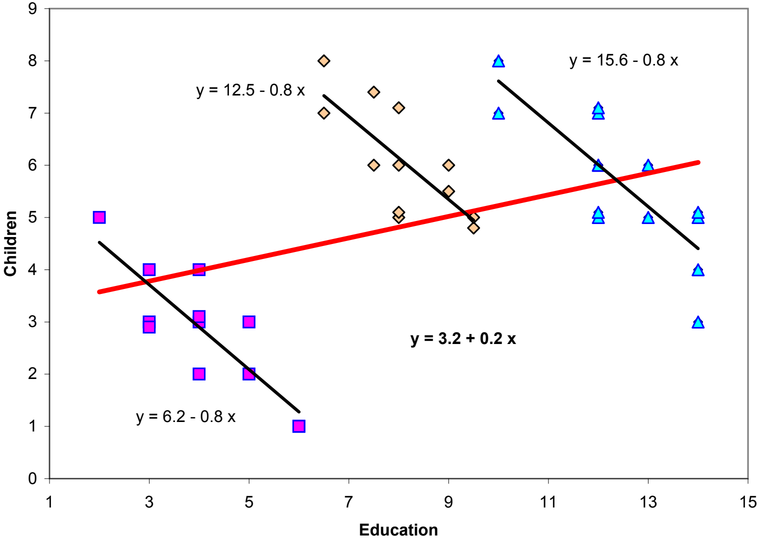
Issues with life course data

- Incomplete sequences
  - Censored and truncated data:
    Cases falling out of observation before experiencing an event of interest.
  - Sequences of varying length.
- Time varying predictors.
  - Example: When analysing time to divorce, presence of children is a time varying predictor.
- Data collected by clusters
  - Example: Household panel surveys.
  - Multi-level analysis to account for unobserved shared characteristics of members of a same cluster.

Issues with life course data

- Incomplete sequences
  - Censored and truncated data:
    Cases falling out of observation before experiencing an event of interest.
  - Sequences of varying length.
- Time varying predictors.
  - Example: When analysing time to divorce, presence of children is a time varying predictor.
- Data collected by clusters
  - Example: Household panel surveys.
  - Multi-level analysis to account for unobserved shared characteristics of members of a same cluster.

Multi-level: Simple linear regression example



Classical statistical approaches

Survival Approaches

- Survival or Event history analysis (Blossfeld and Rohwer, 2002)
  - Focuses on one event.
  - Concerned with duration until event occurs or with hazard of experiencing event.
- Survival curves: Distribution of duration until event occurs

$$S(t) = p(T \geq t) \ .$$

- Hazard models: Regression like models for $S(t, \mathbf{x})$ or hazard $h(t) = p(T = t \mid T \geq t)$

$$h(t, \mathbf{x}) = g\Big(t, \beta_0 + \beta_1 x_1 + \beta_2 x_2(t) + \cdots\Big) \ .$$

Classical statistical approaches

Survival Approaches

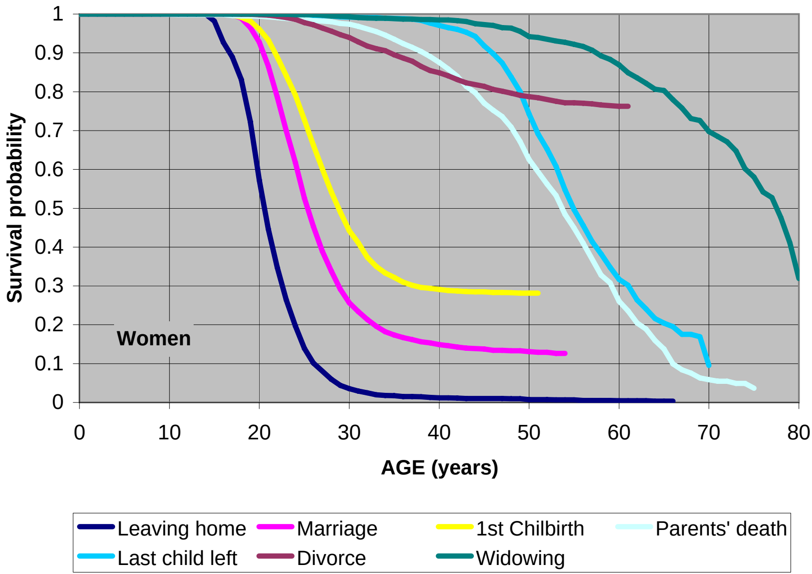
- Survival or Event history analysis (Blossfeld and Rohwer, 2002)
  - Focuses on one event.
  - Concerned with duration until event occurs
    or with hazard of experiencing event.
- Survival curves: Distribution of duration until event occurs

$$S(t) = p(T \geq t) \ .$$

- Hazard models: Regression like models for $S(t, \mathbf{x})$ or hazard
  $h(t) = p(T = t \mid T \geq t)$

$$h(t, \mathbf{x}) = g\Big(t, \beta_0 + \beta_1 x_1 + \beta_2 x_2(t) + \cdots\Big) \ .$$

Survival curves (Switzerland, SHP 2002 biographical survey)

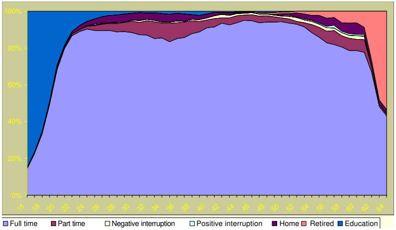


Analysis of sequences

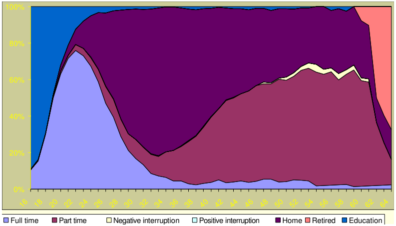
- Frequencies of given subsequences
  - Essentially event sequences.
  - Subsequences considered as categories $\Rightarrow$ Methods for categorical data apply (Frequencies, cross tables, log-linear models, logistic regression, ...).
- Markov chain models
  - State sequences.
  - Focuses on transition rates between states.
    Does the rate also depend on previous states?
    How many previous states are significant?
- Optimal Matching (Abbott and Forrest, 1986) .
  - State sequences.
  - Edit distance (Levenshtein, 1966; Needleman and Wunsch, 1970) between pairs of sequences.
  - Clustering of sequences.

Optimal Matching

- Example from (Gauthier, Widmer, Bucher, and Notredame, 2007)
  - Professional life course, age 16-64, Switzerland
  - SHP retrospective survey, ∼ 3000 cases
  - 5 clusters: Full Time, Part Time, Come Back, Home, Erratic



Full Time, 53%



Come Back, 16%

Typology of methods for life course data

	Issues	
Questions	duration/hazard	state/event sequencing
descriptive	• Survival curves: Parametric (Weibull, Gompertz, ...) and non parametric (Kaplan-Meier, Nelson-Aalen) estimators.	• Optimal matching clustering • Frequencies of given patterns • Discovering typical episodes
causality	• Hazard regression models (Cox, ...) • Survival trees	• Markov models • Mobility trees • Association rules among episodes

Outline

Survival trees: Principle

- Target is survival curve or some other survival characteristic.
- Aim: Partition data set into groups that
- differ as much as possible (max inter class variability)
  - Example: Segal (1988) maximizes difference in KM survival curves by selecting split with smallest p-value of Tarone-Ware Chi-square statistics

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

- are as homogeneous as possible (min intra class variability)
  - Example: Leblanc and Crowley (1992) maximize gain in deviance (-log-likelihood) of relative risk estimates.

Survival trees: Principle

- Target is survival curve or some other survival characteristic.
- Aim: Partition data set into groups that
- differ as much as possible (max inter class variability)
  - Example: Segal (1988) maximizes difference in KM survival curves by selecting split with smallest p-value of Tarone-Ware Chi-square statistics

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

- are as homogeneous as possible (min intra class variability)
  - Example: Leblanc and Crowley (1992) maximize gain in deviance (-log-likelihood) of relative risk estimates.

Survival trees: Principle

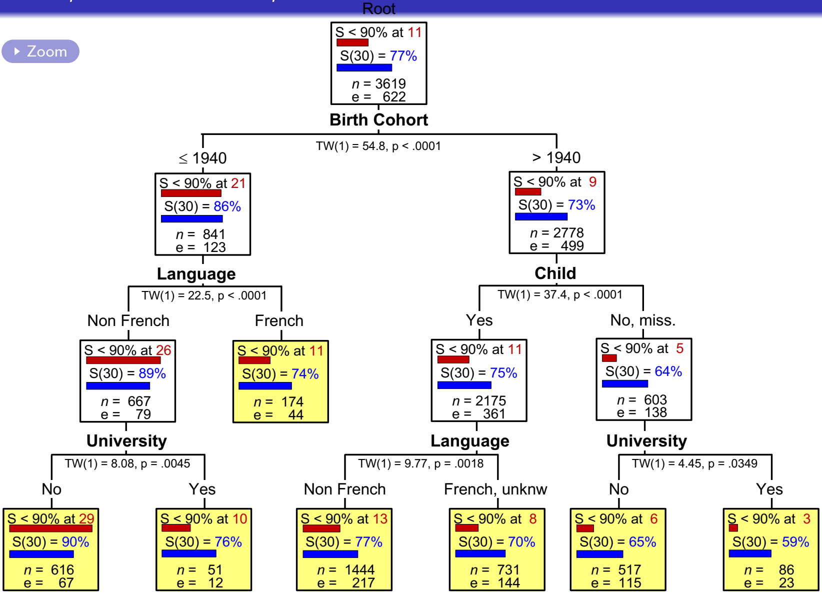
- Target is survival curve or some other survival characteristic.
- Aim: Partition data set into groups that
- differ as much as possible (max inter class variability)
  - Example: Segal (1988) maximizes difference in KM survival curves by selecting split with smallest p-value of Tarone-Ware Chi-square statistics

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

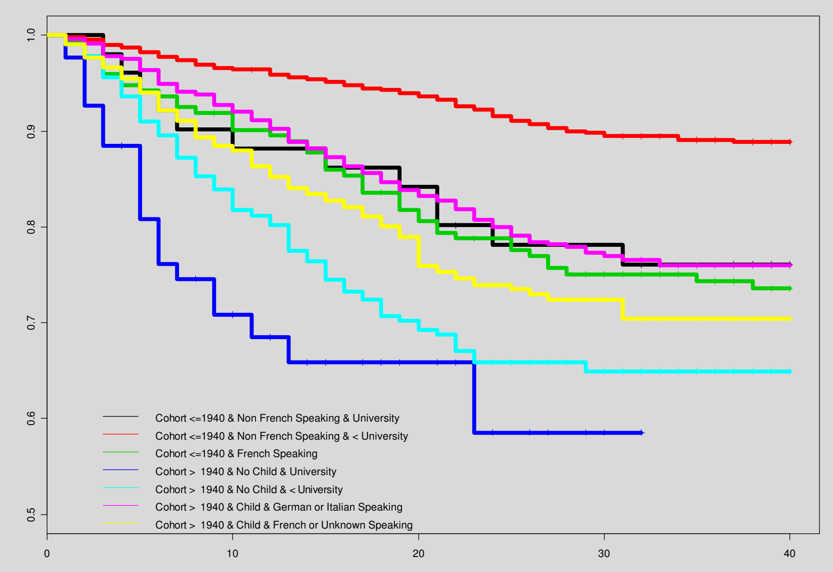
- are as homogeneous as possible (min intra class variability)
  - Example: Leblanc and Crowley (1992) maximize gain in deviance (-log-likelihood) of relative risk estimates.

Divorce, Switzerland, Differences in KM Survival Curves I

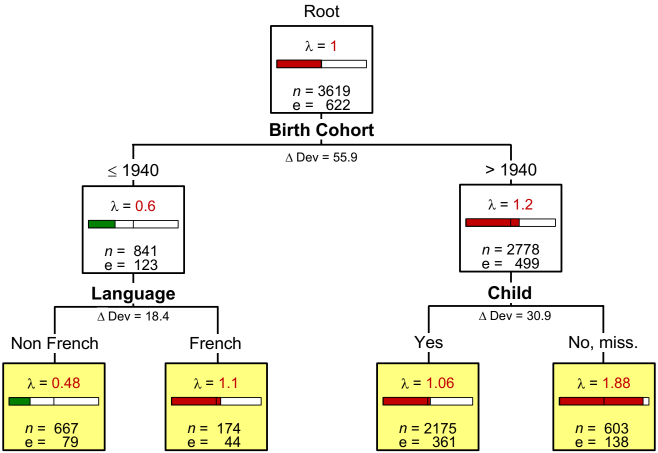
Zoom



Divorce, Switzerland, Differences in KM Survival Curves II



Divorce, Switzerland, Relative risk



Issues with survival trees in social sciences

1. Dealing with time varying predictors
   - Segal (1992) discusses few possibilities, none being really satisfactory.
   - Huang et al. (1998) propose a piecewise constant approach suitable for discrete variables and limited number of changes.
   - Room for development ...
2. Multi-level analysis
   - How can we account for multi-level effects in survival trees, and more generally in trees?
   - Conjecture: Should be possible to include unobserved shared effect in deviance-based splitting criteria.

Issues with survival trees in social sciences

1. Dealing with time varying predictors
   - Segal (1992) discusses few possibilities, none being really satisfactory.
   - Huang et al. (1998) propose a piecewise constant approach suitable for discrete variables and limited number of changes.
   - Room for development ...
2. Multi-level analysis
   - How can we account for multi-level effects in survival trees, and more generally in trees?
   - Conjecture: Should be possible to include unobserved shared effect in deviance-based splitting criteria.

Outline

1. Longitudinal Analysis
   - Motivation
   - Methods for Longitudinal Data
2. Survival Trees
   - Principle
   - Example
   - Social Science Issues
3. Mining Frequent Episodes
   - What Is It About?
   - Example: Counting Alternate Episode Structures
   - Issues Regarding Episode Rules

Mining Frequent Episodes

- Survival approaches not useful in a unitary (holistic) perspective of the whole life course.
- Sequence analysis of whole collection of life events better suited for such holistic approach (Billari, 2005).
- Popular methods in social sciences
  - Optimal Matching.
  - Markov Models.
- What can we expect from frequent episodes mining?
  - GSP (Srikant and Agrawal, 1996)
  - MINEPI, WINEPI (Mannila et al., 1997)
  - TCG, TAG (Bettini et al., 1996)
  - SPADE (Zaki, 2001)

Frequent episodes. What is it?

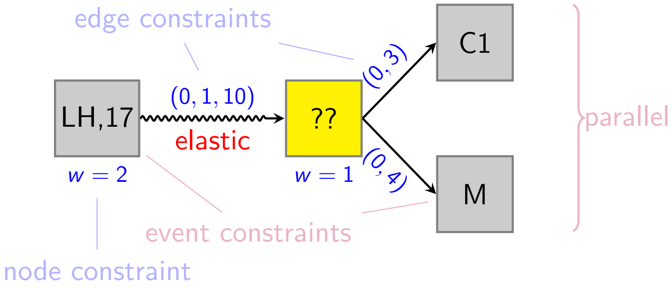
- Episode: Collection of events occurring frequently together.
- Mining typical episodes:
  - Specialized case of mining frequent itemsets.
  - Time dimension $\Rightarrow$ Partially ordered events.
- More complex than unordered itemsets: User must
  - specify time constraints (and episode structure constraints).
  - select a counting method.

Frequent episodes. What is it?

- Episode: Collection of events occurring frequently together.
- Mining typical episodes:
    - Specialized case of mining frequent itemsets.
    - Time dimension $\Rightarrow$ Partially ordered events.
- More complex than unordered itemsets: User must
    - specify time constraints (and episode structure constraints).
    - select a counting method.

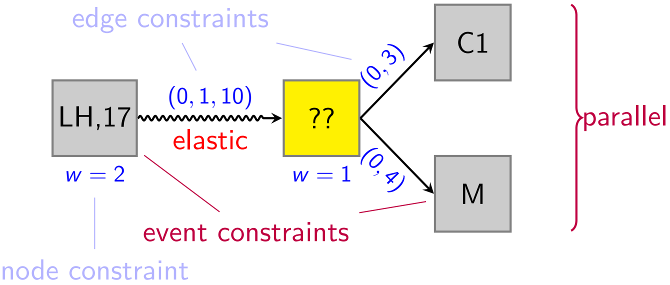
Episode structure constraints

For people who leave home within 2 years from their 17, what are typical events occurring until they get married and have a first child?



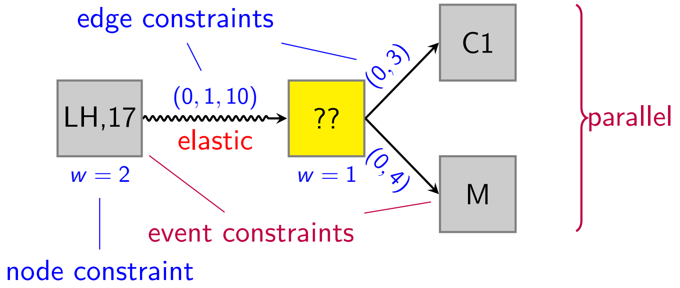
Episode structure constraints

For people who leave home within 2 years from their 17, what are typical events occurring until they get married and have a first child?

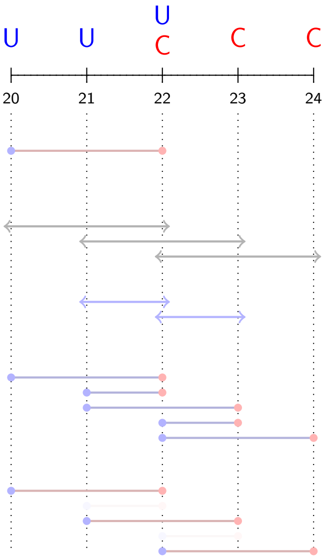


Episode structure constraints

For people who leave home within 2 years from their 17, what are typical events occurring until they get married and have a first child?



Counting methods (Joshi et al., 2001)



Searching (U,C)

min gap= 1, max gap= 2, win size= 2

indiv. with episode COBJ = 1

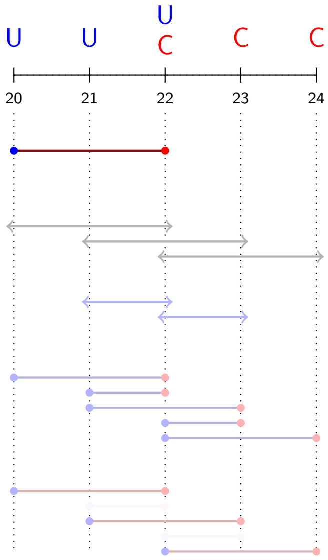
windows with episode CWIN = 3

min win. with episode CminWIN = 2

distinct occurrences CDIS_o = 5

dist. occ. without overlap CDIS = 3

Counting methods (Joshi et al., 2001)

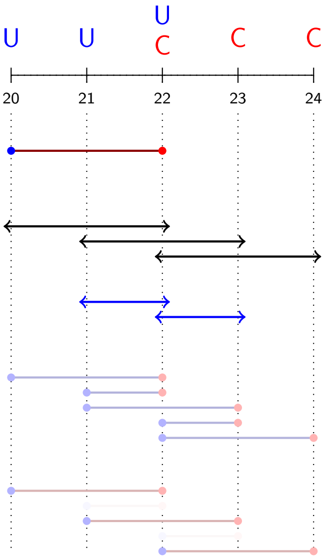


Searching (U,C)

min gap= 1, max gap= 2, win size= 2

indiv. with episode	COBJ = 1
windows with episode	CWIN = 3
min win. with episode	CminWIN = 2
distinct occurrences	CDIS_o = 5
dist. occ. without overlap	CDIS = 3

Counting methods (Joshi et al., 2001)

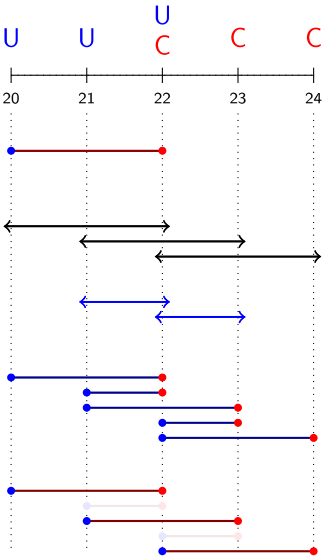


Searching (U,C)

min gap= 1, max gap= 2, win size= 2

indiv. with episode	COBJ = 1
windows with episode	CWIN = 3
min win. with episode	CminWIN = 2
distinct occurrences	CDIS_o = 5
dist. occ. without overlap	CDIS = 3

Counting methods (Joshi et al., 2001)



Searching (U,C)

min gap= 1, max gap= 2, win size= 2

indiv. with episode COBJ $=1$

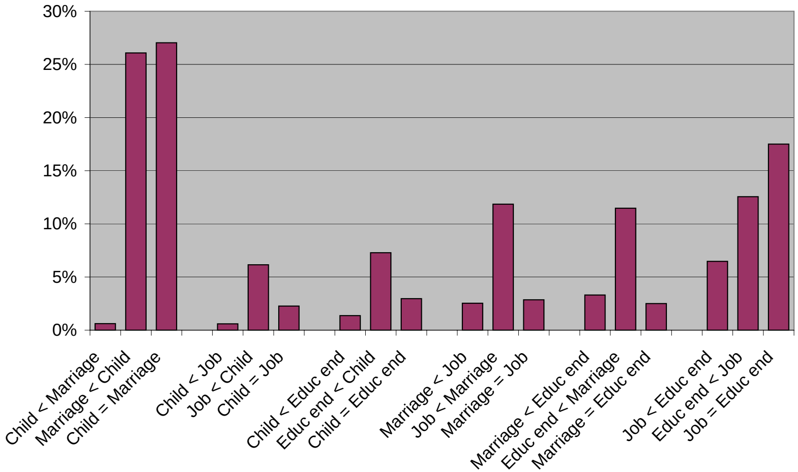
windows with episode CWIN $=3$

min win. with episode CminWIN $=2$

distinct occurrences CDIS_o $=5$

dist. occ. without overlap CDIS $=3$

Example: Counting alternate structures (COBJ, no max gap)



Switzerland, SHP 2002 biographical survey ($n = 5560$).

Rules between episodes

- Social scientists like causal explanations.
- Empirically assessed rules are valuable material in that respect.

- Little attention paid to this aspect in the literature on frequent subsequences.
  - Mined episodes are already structured: if (U,C) is a frequent episode, then we know that C often follows U.
  - Deriving association rules from frequent ordered patterns is similar to what is done with unordered itemsets.
- Rule relevance criteria: confidence, surprisingness, implication strength, ...
- Their value depends on the selected counting method.

Issues with episode rules in social sciences

- Parallel life courses:
  - Family events and professional life course.
  - Life courses of each partner of a couple.
- Mining associations between frequent episodes of a sequence with those of its parallel sequence.
  - Frequent episodes from mix of the 2 sequences, and then restrict search of rules among candidates with premise and consequence belonging to a different sequence.
  - Frequent episodes from each sequence, and then search rules among candidates obtained by combining frequent episodes from each sequence.
- Accounting for multi-level effects when validating rules.
  - Is rule relevant among groups, or within groups?

Issues with episode rules in social sciences

- Parallel life courses:
  - Family events and professional life course.
  - Life courses of each partner of a couple.
- Mining associations between frequent episodes of a sequence with those of its parallel sequence.
  - Frequent episodes from mix of the 2 sequences, and then restrict search of rules among candidates with premise and consequence belonging to a different sequence.
  - Frequent episodes from each sequence, and then search rules among candidates obtained by combining frequent episodes from each sequence.
- Accounting for multi-level effects when validating rules.
  - Is rule relevant among groups, or within groups?

Summary

- Data mining approaches (survival trees, frequent episodes) have promising future in life course analysis.
  - Complement classical statistical outcomes with new insights.
- Their use within social sciences raises specific issues:
  - Accounting for multi-level effects when growing survival tree or mining association rules.
  - Handling time varying predictors in survival trees.
  - Selecting relevant counting methods (event dependent)?
  - Suitable criteria for measuring association strength between frequent episodes.
  - ...

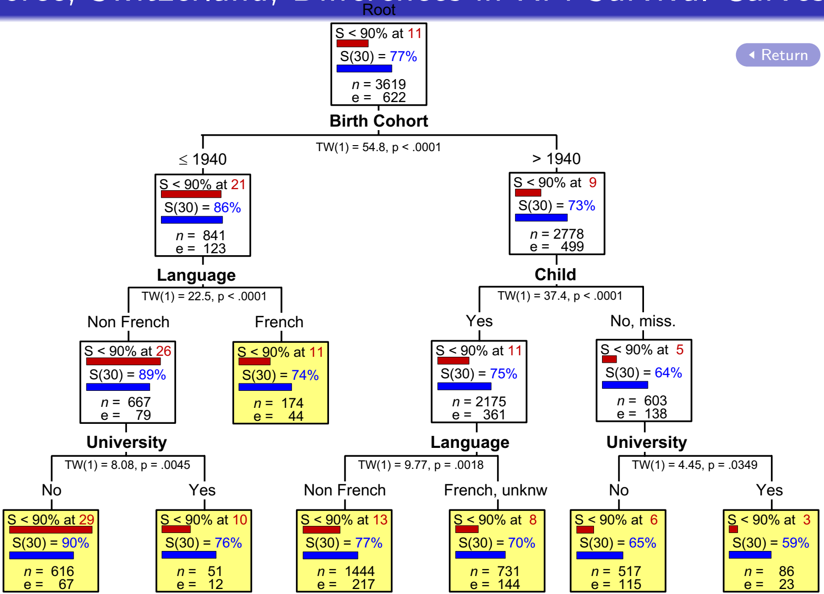
Summary

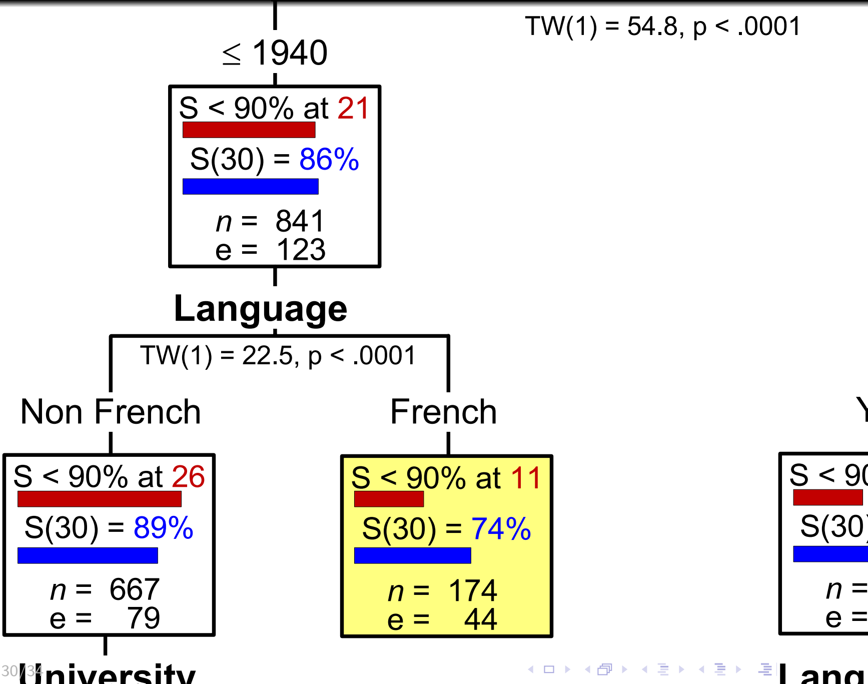
- Data mining approaches (survival trees, frequent episodes) have promising future in life course analysis.
  - Complement classical statistical outcomes with new insights.
- Their use within social sciences raises specific issues:
  - Accounting for multi-level effects when growing survival tree or mining association rules.
  - Handling time varying predictors in survival trees.
  - Selecting relevant counting methods (event dependent)?
  - Suitable criteria for measuring association strength between frequent episodes.
  - ...

Thank You!

Divorce, Switzerland, Differences in KM Survival Curves I

Return




TW(1) = 54.8, p < .0001
≤ 1940
S < 90% at 21
S(30) = 86%
n = 841
e = 123
Language
TW(1) = 22.5, p < .0001
Non French
French
S < 90% at 26
S(30) = 89%
n = 667
e = 79
S < 90% at 11
S(30) = 74%
n = 174
e = 44
University

For Further Reading I

Abbott, A. and J. Forrest (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

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 P. Ghisletta, J.-M. Le Goff, R. Levy, D. Spini, and E. Widmer (Eds.), *Towards an Interdisciplinary Perspective on the Life Course*, Advancements in Life Course Research, Vol. 10, pp. 267–288. Amsterdam: Elsevier.

For Further Reading II

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

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

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*.

For Further Reading III

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

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

Mannila, H., H. Toivonen, and 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.

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

For Further Reading IV

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.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning 42*(1/2), 31–60.