Mining Event Histories:
Some New Insights on Personal Swiss Life Courses

Gilbert Ritschard

Dept of Econometrics and Laboratory of Demography, University of Geneva
http://mephisto.unige.ch

PaVie Seminar, Lausanne, October 22, 2008

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

- My talk is about life courses,
- So, let me start with an example of scientific life course

date	event
1970-1979	Studies in econometrics
1980-1992	Mathematical Economics
1985-...	Work with Social scientists (Family studies)
	Interest in Statistics for social sciences
1990-1995	Interest in Neural Networks
2000-...	KDD and data mining (Clustering, supervised learning)
2003-...	Work with historians, demographers, psychologists (longitudinal data)
2005-...	KDD and Data mining approaches for analysing life course data
2007-...	Start a SNF project on "Mining Event Histories"

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

- 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?

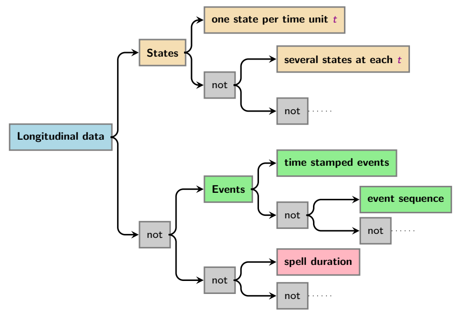
Motivation: KD in Social sciences

- In KDD (Knowledge discovery in databases) and data mining, focus on prediction and classification.
- Improve prediction and classification errors.
- In Social science, aim is understanding/explaining (social) behaviors.
- Hence focus is on process rather than output.

What kind of data?

- What kind of data are we dealing with?
- Mainly categorical longitudinal data describing life courses
- Data can be in different forms ...

ontology of longitudinal data (Aristotelean tree)



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

Transforming time stamped events into state sequences

Example: the "BioFam" data

- Data from the retrospective survey conducted in 2002 by the Swiss Household Panel (SHP)
- (with support of Federal Statistical Office, Swiss National Fund for Scientific Research, University of Neuchatel.)
- Retrospective survey: 5560 individuals
- Retained familial life events: Leaving Home, First childbirth, First marriage and First divorce.
- Age 15 to 45 → 2601 remaining individuals, born between 1909 et 1957.

Creating state sequences

- Example of time stamped data:

individual	LHome	marriage	childbirth	divorce
1	1989	1990	1992	NA

Deriving the states

Need one state for each combination of events:

	LHome	marriage	childbirth	divorce
0	no	no	no	no
1	yes	no	no	no
2	no	yes	yes/no	no
3	yes	yes	no	no
4	no	no	yes	no
5	yes	no	yes	no
6	yes	yes	yes	no
7	yes/no	yes	yes/no	yes

From events to states

Example of transformation :

- events:

individual	LHome	marriage	childbirth	divorce
1	1989	1990	1992	NA

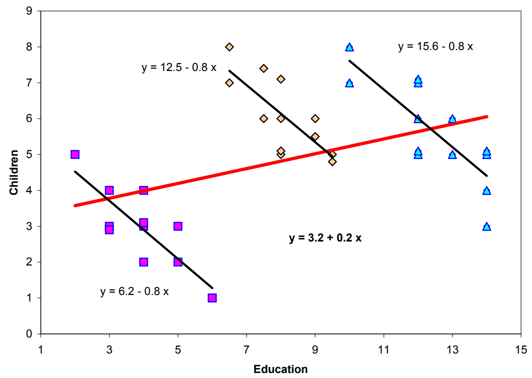
- states:

individual	...	1988	1989	1990	1991	1992	1993	...
1	...	0	0	1	3	3	6	...

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

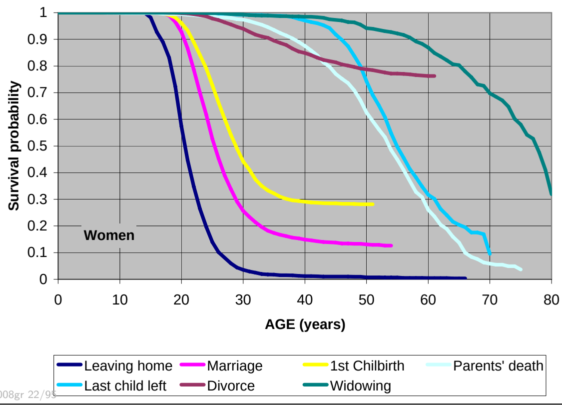
- Survival or Event history analysis (Blossfeld and Rohwer, 2002)
  - Focuses on one event.
  - Concerned with duration until event occurs
    or with hazard of experiencing event.
- Survival curves: Distribution of duration until event occurs

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

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

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

Survival curves (Switzerland, SHP 2002 biographical survey)



Analysis of sequences

- Frequencies of given subsequences
  - Essentially event sequences, e.g. (First job → Marriage).
  - Subsequences considered as categories ⇒ 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.

Typology of methods for life course data

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

SHP biographical retrospective survey

http://www.swisspanel.ch

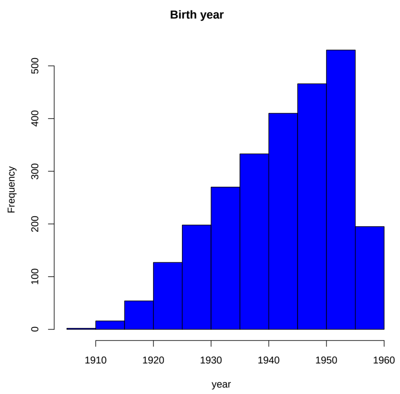
- SHP retrospective survey: 2001 (860) and 2002 (4700 cases).
- We consider only data collected in 2002.
- Data completed with variables from 2002 wave (language).

Characteristics of retained data for divorce

(individuals who get married at least once)

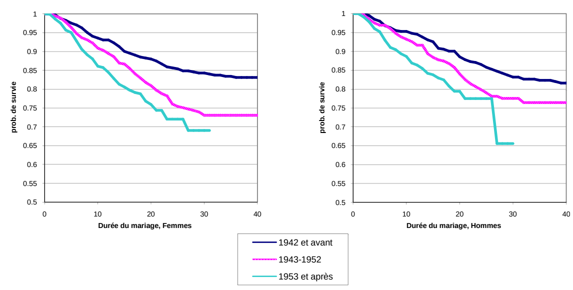
	men	women	Total
Total	1414	1656	3070
1st marriage dissolution	231	308	539
	16.3%	18.6%	17.6%

Distribution by birth cohort



Marriage duration until divorce

Survival curves



Marriage duration until divorce

Hazard model

- Discrete time model (logistic regression on person-year data)
- exp(B) gives the Odds Ratio, i.e. change in the odd $h/(1-h)$ when covariate increased by 1 unit.

		exp(B)	Sig.
birthyr		1.0088	0.002
university		1.22	0.043
child		0.73	0.000
language	unknwn	1.47	0.000
	French	1.26	0.007
	German	1	ref
	Italian	0.89	0.537
Constant		0.0000000004	0.000

Survival trees: Principle

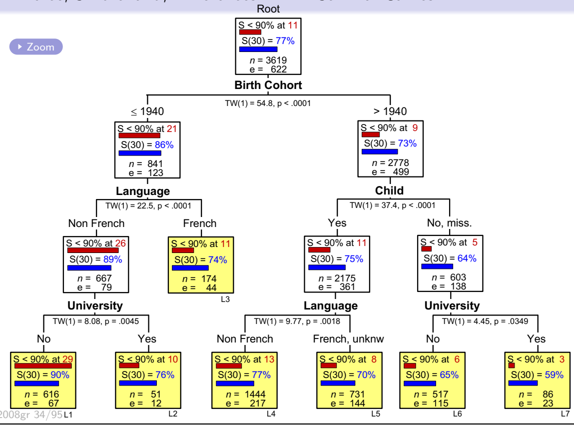
- Target is survival curve or some other survival characteristic.
- Aim: Partition data set into groups that
- differ as much as possible (max between class variability)
  - Example: Segal (1988) maximizes difference in KM survival curves by selecting split with smallest p-value of Tarone-Ware Chi-square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

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

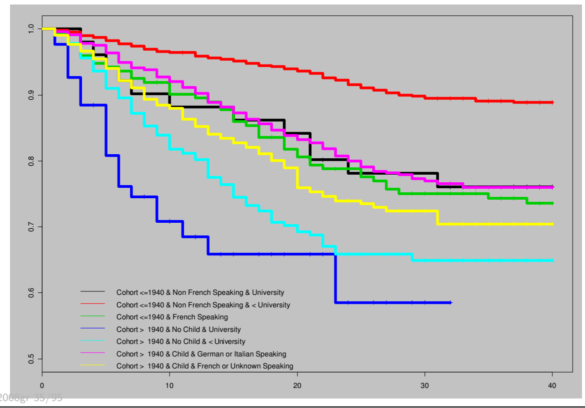
- are as homogeneous as possible (min within class variability)
  - Example: Leblanc and Crowley (1992) maximize gain in deviance (-log-likelihood) of relative risk estimates.

Divorce, Switzerland, Differences in KM Survival Curves I

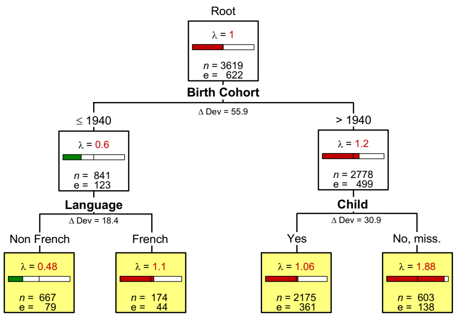
Zoom



Divorce, Switzerland, Differences in KM Survival Curves II



Divorce, Switzerland, Relative risk



Hazard model with interaction

- Adding interaction effects detected with the tree approach
- improves significantly the fit (sig $\Delta\chi^2 = 0.004$)

		exp(B)	Sig.
born after 1940		1.78	0.000
university		1.22	0.049
child		0.94	0.619
language	unknwn	1.50	0.000
	French	1.12	0.282
	German	1	ref
	Italian	0.92	0.677
b_before_40*French		1.46	0.028
b_after_40*child		0.68	0.010
Constant		0.008	0.000

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.

Sequence analysis

- 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).

Rendering sequences

- **Colorize your life courses**

- Results from the analysis of the retrospective Swiss Household Panel (SHP) survey.
- Focus on visualization of life course data.

Evolution tendencies in familial life course trajectories

Sequence analysis techniques permit to test hypotheses about evolution in these familial life trajectories. (Elzinga and Liefbroer, 2007):

- De-standardization: Some states and events of familial life are shared by decreasing proportions of the population, occur at more dispersed ages and their duration is also more scattered.
- De-institutionalization: Social and temporal organization of life courses becomes less driven by normative, legal or institutional rules.
- Differentiation: Number of distinct steps lived by individual increases.

Characterizing sets of sequences

- Sequence of transversal measures (between entropy, ...)

id	t_1	t_2	t_3	···
1	B	B	D	···
2	A	B	C	···
3	B	B	A	···

- Summary of longitudinal measures (sequence entropy, ...)

id	t_1	t_2	t_3	···
1	B	B	D	···
2	A	B	C	···
3	B	B	A	···

- Other global characteristics: Central sequence, Sequence diversity, ...

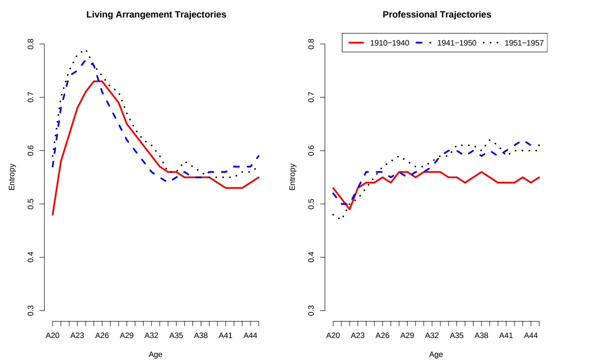
Entropy

- Entropy: Measure of uncertainty regarding state predictability.
  - p_i, proportion of cases (or time points) in state i.
  - Shannon $h(p) = \sum_i -p_i \log_2(p_i)$
  - Other types of entropies: Quadratic (Gini), Daroczy, ...
- Two ways of using entropies.
  - (Transversal) entropy of the state at each time (age) point: Entropy increases with diversity of states observed at each time point (age).
  - (Longitudinal) entropy of each individual sequences: Entropy increases with diversity of states during the observed life course and varies with the time spend in each state.

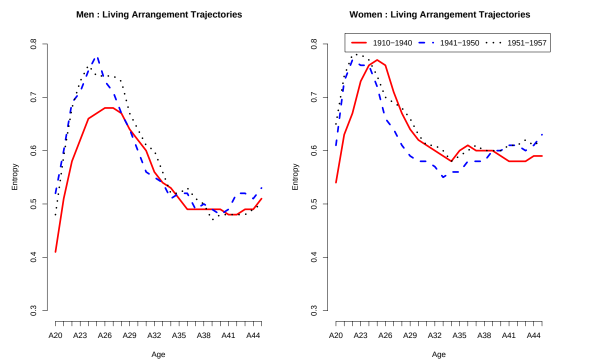
Illustrative data

- Data from the 2002 SHP biographical survey
- Interested in relationship between
  - Cohabitational trajectories (10 states)
  - Professional trajectories (8 states)
- We use the coding retained by Gauthier (2007)
- Focus on ages 20 to 45 (sequence length = 26 years)
- 1503 cases (751 women, 752 men)

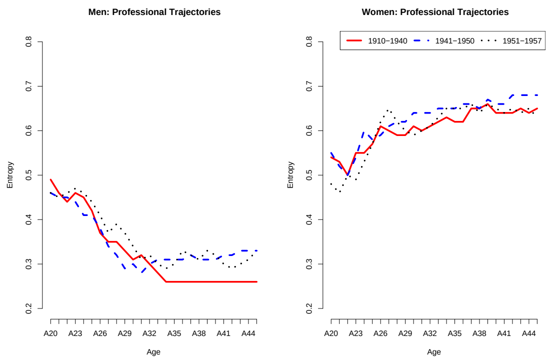
Transversal entropy at each time (age) point



Transversal entropy at each time (age) point



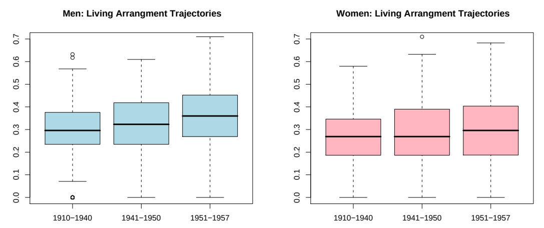
Transversal entropy at each time (age) point



Hypothesis about longitudinal entropies

- Cohabitational and professional life trajectories
  - become less stable
  - more diversified
- Their entropy tends to increase for younger generations.
- Are increases in professional trajectories related to increases in cohabitational trajectories?

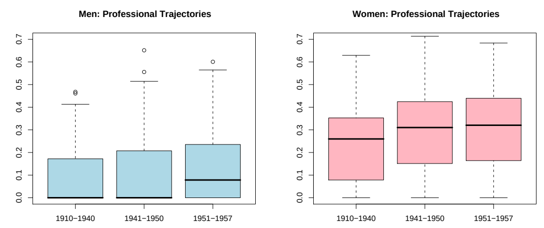
Entropy of cohabitational trajectories



$p(F > f) = .000^{***}$ — all 2 by 2 differences significant

$p(F > f) = .073^{*}$ — coh3 significantly (.02) different from coh1

Entropy of professional trajectories



$p(F > f) = .002^{***}$

$p(F > f) = .001^{***}$

coh3 not significantly different from coh2

Correlation between cohabitational and professional entropies

	Overall	Men	Women
1910-1940	0.08 *	0.11 *	0.19 ***
1941-1950	0.12 **	0.14 **	0.30 ***
1951-1957	0.15 **	0.25 ***	0.31 ***

Clustering, Multidimensional scaling and more

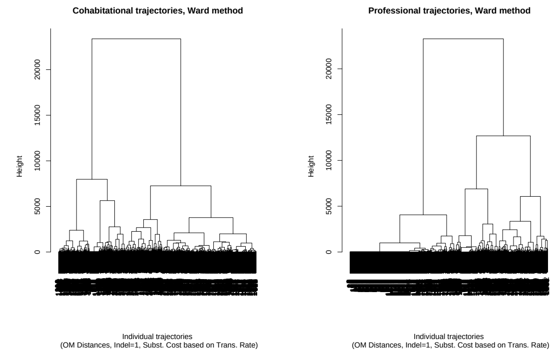
- Once you are able to compute 2 by 2 distances between sequences you can among others:
- Cluster sequences
- Analyse the trajectory heterogeneity (Generalized ANOVA)
- Make scatter plot representation of sets of sequences using multidimensional scaling.

Distances between sequences

- Edit distance (known as Optimal matching in Social sciences) (Levenshtein, 1966; Needleman and Wunsch, 1970; Abbott and Forrest, 1986)
  - $d(x, y)$ Total cost of insert, deletion and substitution changes required to transform sequence x into y.
  - Different solutions depending on indel and substitution costs.
- Other metrics proposed by (Elzinga, 2008)
  - LCP: Longest common prefix (also longest common postfix)
  - LCS: Longest common subsequence (same as OM with indel cost = 1, and substitution cost = 2).
  - NMS: Number of matching subsequences
  - ...

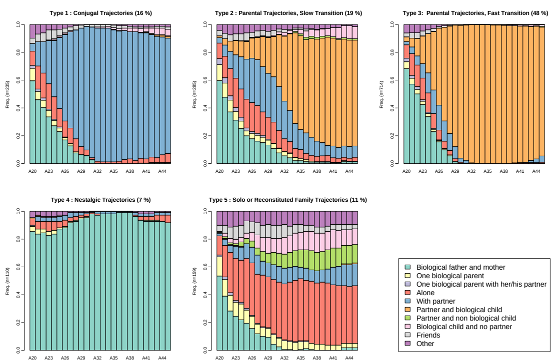
Elzinga (2008) proposes a nice formalization of these metrics.

Clustering with OM distances: Dendrograms

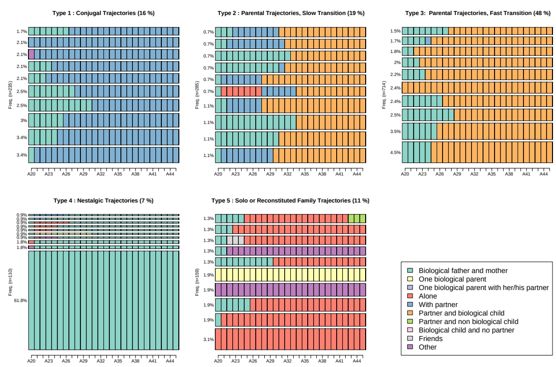


LA Prof

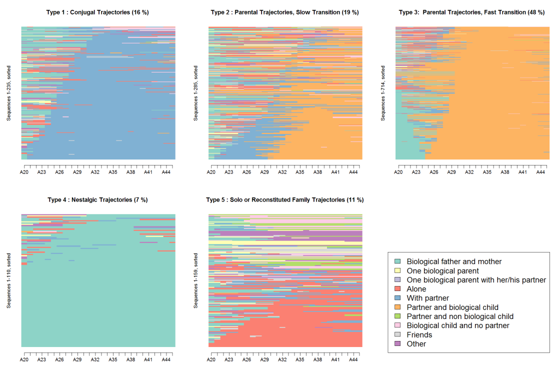
LA, State distribution by age, within cluster



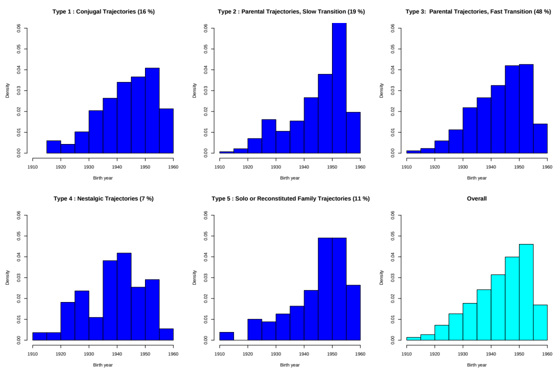
LA, Most frequent sequences by cluster



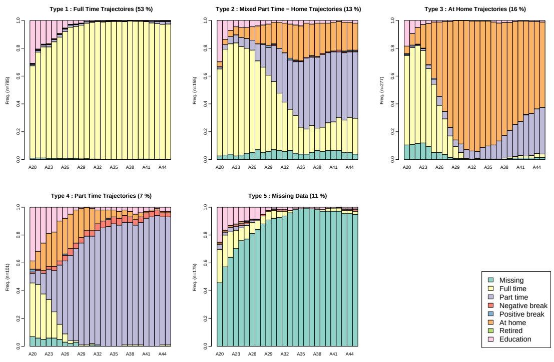
LA, Sequence diversity within cluster



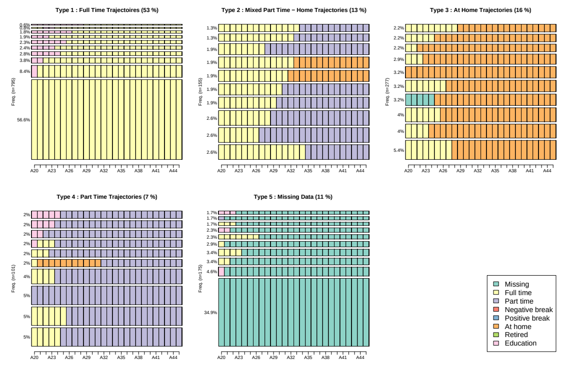
LA, Birth year distribution by cluster



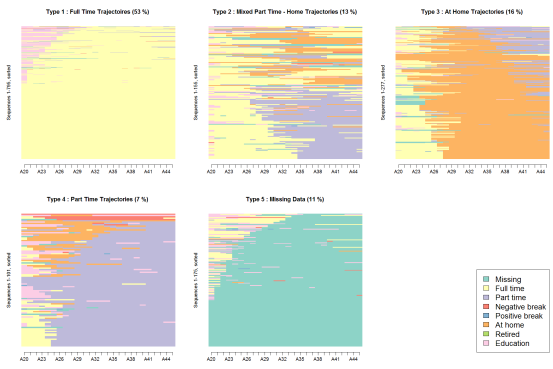
Prof, State distribution by age, within cluster



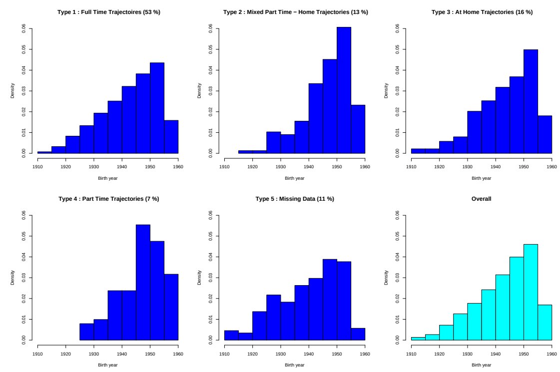
Prof, Most frequent sequences by cluster



Prof, Sequence diversity within cluster



Prof, Birth year distribution by cluster



Heterogeneity of set of sequences

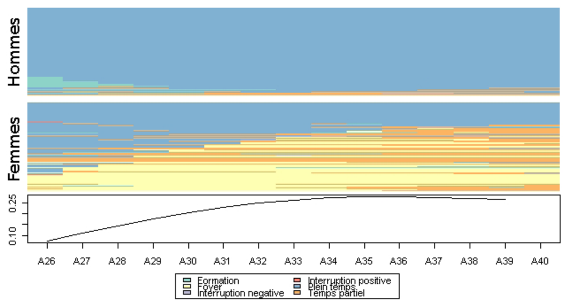
- Sum of squares can be expressed in terms of the distances between each pair of points

$$SS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}$$

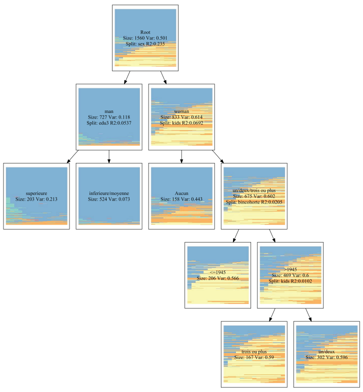
- Setting d_{ij} to the OM, LCP, LCS, ... distance, we get a measure of diversity or heterogeneity of sequences.
- Can apply ANOVA principle to sequences.

Heterogeneity analysis

Professional trajectories by sex



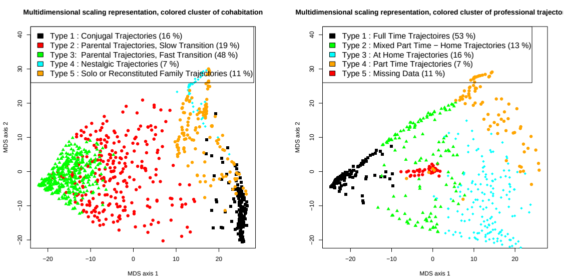
Sequence Tree



Multidimensional Scaling: Principle

- Let D be a distance matrix between sequences.
- D computed using OM, LPS, LCS, ... metrics.
- Multidimensional Scaling consists in
  - Finding a set of real valued variables (f_1, f_2) such that the $\delta_{ij} = \sqrt{(f_{i1} - f_{j1})^2 + (f_{i2} - f_{j2})^2}$ best approximate the distances between sequences.
  - Plotting the points in the (f_1, f_2) space.

Multidimensional Scaling



Mining Frequent Episodes

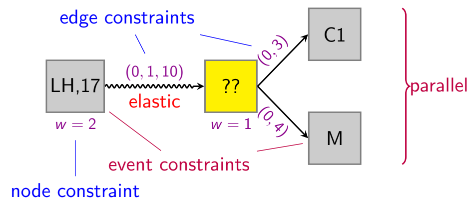
- (Time stamped) event sequences
- What can we expect from frequent episodes mining?
  - GSP (Srikant and Agrawal, 1996)
  - MINEPI, WINEPI (Mannila et al., 1997)
  - TCG, TAG (Bettini et al., 1996)
  - SPADE (Zaki, 2001)
- Are there specific issues when applying these methods in social sciences?

Frequent episodes. What is it?

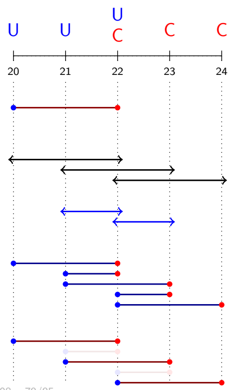
- Episode: Collection of events occurring frequently together.
- Mining typical (frequent) episodes:
  - Specialized case of mining frequent itemsets.
  - Time dimension ⇒ Partially ordered events.
- More complex than unordered itemsets: User must
  - specify time constraints (and episode structure constraints).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



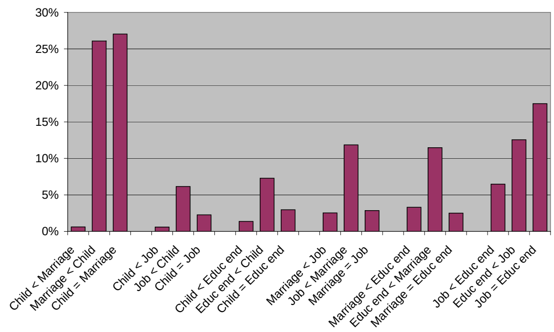
Counting methods (Joshi et al., 2001)



Searching (U,C)
min gap= 1, max gap= 2, win size= 2

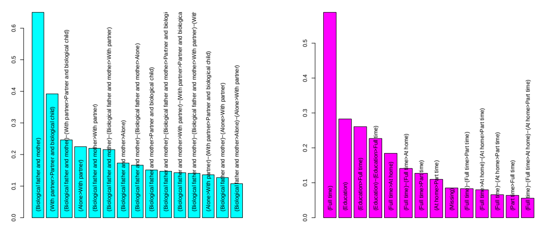
indiv. with episode	COBJ = 1
windows with episode	CWIN = 3
min win. with episode	CminWIN = 2
distinct occurrences	CDIS_o = 5
dist. occ. without overlap	CDIS = 3

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

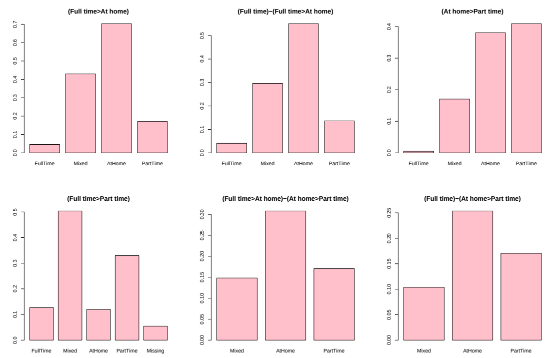


Switzerland, SHP 2002 biographical survey ($n = 5560$).

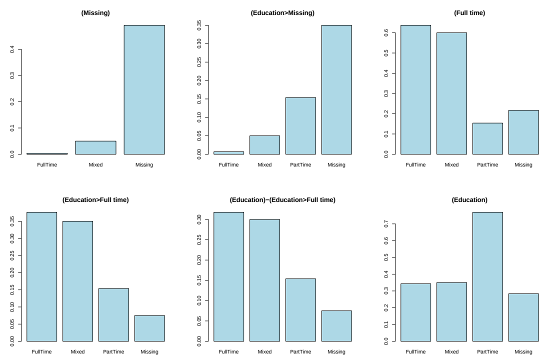
Frequent episodes, cohabitational and professional trajectories



Discriminant episodes, professional trajectories, women



Discriminant episodes, professional trajectories, men



Summary

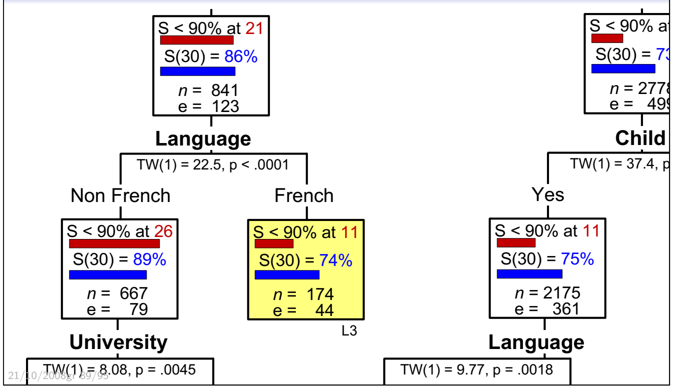
- Data mining approaches (survival trees, clustering sequences, frequent episodes) have promising future in life course analysis.
  - Complement classical statistical outcomes with new insights.
- Their use within social sciences raises specific issues:
  - Accounting for multi-level effects when growing 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.
  - ...

Our TraMineR R-package

- Let me finish with an Add ...
- TraMineR, a free life trajectory mining tool
- for the free open source R statistical environment.
- downloadable from `http://cran.r-project.org` (CRAN)
- see also `http://mephisto.unige.ch/biomining`

Thank You!

Divorce, Switzerland, Differences in KM Survival Curves I



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.

For Further Reading II

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*, Advances in Life Course Research, Vol. 10, pp. 267–288. Amsterdam: Elsevier.

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

Elzinga, C. H. (2008). Sequence analysis: Metric representations of categorical time series. *Sociological Methods and Research*. forthcoming.

For Further Reading III

Elzinga, C. H. and A. C. Liefbroer (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population 23*, 225–250.

Gauthier, J.-A. (2007). *Empirical categorizations of social trajectories: A sequential view on the life course*. Thèse, Université de Lausanne, Faculté des sciences sociales et politique (SSP), Lausanne.

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

For Further Reading IV

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*.

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.

For Further Reading V

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.

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.

For Further Reading VI

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