Mining Event or State Sequences: A Social Science Perspective

Gilbert Ritschard

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

IIS 2008, Zakopane, Poland, June 16-18
My talk is about life courses,

Example of scientific life course
to help you understand what a social scientist does at IIS

<table>
<thead>
<tr>
<th>date</th>
<th>event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970-1979</td>
<td>Studies in econometrics</td>
</tr>
<tr>
<td>1980-1992</td>
<td>Mathematical Economics</td>
</tr>
<tr>
<td>1985-...</td>
<td>Work with Social scientists (Family studies)</td>
</tr>
<tr>
<td></td>
<td>Interest in Statistics for social sciences</td>
</tr>
<tr>
<td>1990-1995</td>
<td>Interest in Neural Networks</td>
</tr>
<tr>
<td>2000-...</td>
<td>KDD and data mining (Clustering, supervised learning)</td>
</tr>
<tr>
<td>2003-...</td>
<td>Work with historians, demographers, psychologists (longitudinal data)</td>
</tr>
<tr>
<td>2005-...</td>
<td>KDD and Data mining approaches for analysing life course data</td>
</tr>
</tbody>
</table>
Outline

1. Sequence Analysis in Social Sciences
2. Survival Trees
3. Visualizing and clustering sequence data
4. Mining Frequent Episodes
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 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 are we dealing with?

Mainly **categorical longitudinal** data describing life courses

An ontology of longitudinal data (Aristotelian tree).
Table: Time stamped events, record for Sandra

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

Table: State sequence view, Sandra

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>civil status</td>
<td>single</td>
<td>single</td>
<td>single</td>
<td>single</td>
<td>married</td>
</tr>
<tr>
<td>education level</td>
<td>primary</td>
<td>secondary</td>
<td>secondary</td>
<td>secondary</td>
<td>secondary</td>
</tr>
<tr>
<td>job</td>
<td>no</td>
<td>no</td>
<td>first</td>
<td>first</td>
<td>first</td>
</tr>
</tbody>
</table>
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 \( S(t, x) \) or hazard
  \[ h(t) = p(T = t \mid T \geq t) \]
  \[ h(t, x) = g\left(t, \beta_0 + \beta_1 x_1 + \beta_2 x_2(t) + \cdots\right) . \]
Survival curves (Switzerland, SHP 2002 biographical survey)

- Leaving home
- Marriage
- 1st Chilbirth
- Parents' death
- Last child left
- Divorce
- Widowing

AGE (years) vs. Survival probability

Women
Analysis of sequences

- Frequencies of given subsequences
  - Essentially event sequences.
  - 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

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

<table>
<thead>
<tr>
<th></th>
<th>men</th>
<th>women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>1414</td>
<td>1656</td>
<td>3070</td>
</tr>
<tr>
<td>1st marriage dissolution</td>
<td>231</td>
<td>308</td>
<td>539</td>
</tr>
<tr>
<td></td>
<td>16.3%</td>
<td>18.6%</td>
<td>17.6%</td>
</tr>
</tbody>
</table>
Distribution by birth cohort

Birth year

Frequency

1910 1920 1930 1940 1950 1960
0 100 200 300 400 500
Marriage duration until divorce
Survival curves
Discrete time model (logistic regression on person-year data)

exp(B) gives the Odds Ratio, i.e. change in the odd \( \frac{h}{1 - h} \) when covariate increased by 1 unit.

<table>
<thead>
<tr>
<th></th>
<th>exp(B)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthyr</td>
<td>1.0088</td>
<td>0.002</td>
</tr>
<tr>
<td>university</td>
<td>1.22</td>
<td>0.043</td>
</tr>
<tr>
<td>child</td>
<td>0.73</td>
<td>0.000</td>
</tr>
<tr>
<td>language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td>1.47</td>
<td>0.000</td>
</tr>
<tr>
<td>French</td>
<td>1.26</td>
<td>0.007</td>
</tr>
<tr>
<td>German</td>
<td>1</td>
<td>ref</td>
</tr>
<tr>
<td>Italian</td>
<td>0.89</td>
<td>0.537</td>
</tr>
<tr>
<td>Constant</td>
<td>0.0000000004</td>
<td>0.000</td>
</tr>
</tbody>
</table>
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

$$TW = \sum_{i} w_i \left( d_{i1} - E(D_i) \right) \left( \frac{w_i^2 \text{var}(D_i)}{w_i^2 \text{var}(D_i)} \right)^{1/2}$$

- **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

**Birth Cohort**
- **≤ 1940**
  - **Language**
    - Non French
      - S < 90% at 26
        - S(30) = 89%
        - n = 667
        - e = 79
    - French
      - S < 90% at 11
        - S(30) = 74%
        - n = 174
        - e = 44
  - University
    - No
      - S < 90% at 29
        - S(30) = 90%
        - n = 616
        - e = 67
    - Yes
      - S < 90% at 10
        - S(30) = 76%
        - n = 51
        - e = 12

- **> 1940**
  - S < 90% at 9
    - S(30) = 73%
    - n = 2778
    - e = 499
  - Language
    - Yes
      - S < 90% at 11
        - S(30) = 75%
        - n = 2175
        - e = 361
    - No, miss.
      - S < 90% at 5
        - S(30) = 64%
        - n = 603
        - e = 138
  - University
    - No
      - S < 90% at 6
        - S(30) = 65%
        - n = 517
        - e = 115
    - Yes
      - S < 90% at 3
        - S(30) = 59%
        - n = 86
        - e = 23
Divorce, Switzerland, Differences in KM Survival Curves II
Divorce, Switzerland, Relative risk

Birth Cohort

\( \lambda = 1 \)
\( n = 3619 \)
\( e = 622 \)

\( \Delta \text{Dev} = 55.9 \)

\( \leq 1940 \)
\( \lambda = 0.6 \)
\( n = 841 \)
\( e = 123 \)

\( \Delta \text{Dev} = 18.4 \)

Language

Non French

\( \lambda = 0.48 \)
\( n = 667 \)
\( e = 79 \)

French

\( \lambda = 1.1 \)
\( n = 174 \)
\( e = 44 \)

\( \geq 1940 \)
\( \lambda = 1.2 \)
\( n = 2778 \)
\( e = 499 \)

Child

Yes

\( \lambda = 1.06 \)
\( n = 2175 \)
\( e = 361 \)

No, miss.

\( \lambda = 1.88 \)
\( n = 603 \)
\( e = 138 \)
Hazard model with interaction

- Adding interaction effects detected with the tree approach
- Improves significantly the fit \( \text{sig } \Delta \chi^2 = 0.004 \)

<table>
<thead>
<tr>
<th></th>
<th>exp(B)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>born after 1940</td>
<td>1.78</td>
<td>0.000</td>
</tr>
<tr>
<td>university</td>
<td>1.22</td>
<td>0.049</td>
</tr>
<tr>
<td>child</td>
<td>0.94</td>
<td>0.619</td>
</tr>
<tr>
<td>language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td>1.50</td>
<td>0.000</td>
</tr>
<tr>
<td>French</td>
<td>1.12</td>
<td>0.282</td>
</tr>
<tr>
<td>German</td>
<td>1</td>
<td>ref</td>
</tr>
<tr>
<td>Italian</td>
<td>0.92</td>
<td>0.677</td>
</tr>
<tr>
<td>b_before_40*French</td>
<td>1.46</td>
<td>0.028</td>
</tr>
<tr>
<td>b_after_40*child</td>
<td>0.68</td>
<td>0.010</td>
</tr>
<tr>
<td>Constant</td>
<td>0.008</td>
<td>0.000</td>
</tr>
</tbody>
</table>
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.
Presentation of 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 $\rightarrow$ 2601 remaining individuals, born between 1909 et 1957.
Distribution by birth cohort

Birth year

Frequency

1910 1920 1930 1940 1950 1960

0 100 200 300 400 500

Distribution by birth cohort

Mining Event or State Sequences
Visualizing and clustering sequence data
Example: the BioFam sequential data set
Creating state sequences

Example of time stamped data:

<table>
<thead>
<tr>
<th>individual</th>
<th>LHome</th>
<th>marriage</th>
<th>childbirth</th>
<th>divorce</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1989</td>
<td>1990</td>
<td>1992</td>
<td>NA</td>
</tr>
</tbody>
</table>
Deriving the states

Need one state for each combination of events:

<table>
<thead>
<tr>
<th></th>
<th>LHome</th>
<th>marriage</th>
<th>childbirth</th>
<th>divorce</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>1</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>yes</td>
<td>yes/no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>yes/no</td>
<td>yes</td>
<td>yes/no</td>
<td>yes</td>
</tr>
</tbody>
</table>
Definition

- **Entropy**: measure of uncertainty regarding sequence predictability.
  - $p_i$, proportion of cases (or time points) in state $i$.
  - Shannon $h(p) = \sum_i -p_i \log_2(p_i)$
  - Other type of entropies: Quadratic (Gini), Daroczy, ...

- Two ways of using entropies.
  - **Entropy of the state at each time (age) point**: Entropy increases with diversity of states observed at each time point (age).
  - **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.
Mining Event or State Sequences
Visualizing and clustering sequence data
Characteristics of sequences

Entropy of the state at each time (age) point

Entropy of bifam state distribution by age

<table>
<thead>
<tr>
<th>Age</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>a15</td>
<td>0.2</td>
</tr>
<tr>
<td>a17</td>
<td>0.4</td>
</tr>
<tr>
<td>a19</td>
<td>0.6</td>
</tr>
<tr>
<td>a21</td>
<td>0.8</td>
</tr>
<tr>
<td>a23</td>
<td></td>
</tr>
<tr>
<td>a25</td>
<td></td>
</tr>
<tr>
<td>a27</td>
<td></td>
</tr>
<tr>
<td>a29</td>
<td></td>
</tr>
</tbody>
</table>
Mining Event or State Sequences
Visualizing and clustering sequence data
Characteristics of sequences

Entropy: Minimum/maximum

Entropie minimum, médiane et maximum

Sequences 1–15, sorted by Entropy
A15 A20 A25 A30 A35 A40 A45
N/N/N/N
Y/N/N/N
N/Y/*/N
Y/Y/N/N
N/N/Y/N
Y/N/Y/N
Y/Y/Y/N
*/*//*/Y

Time
Entropy - histogram

Entropy for the sequences in the biofam data set
Hypothesis

- Evolutions of familial life trajectories gives rise to an increase in the entropy of individual sequences,
- because they become less predictable and more diversified.
Mining Event or State Sequences
Visualizing and clustering sequence data
Characteristics of sequences

Entropy by birth cohorts

Distribution de l'entropie selon les cohortes de naissances
Entropy by sex

Distribution de l'entropie selon le sexe

- Hommes
- Femmes

Sequences entropy

0.0 0.5 1.0 1.5

Sexe
Definition

- **Turbulence** (Elzinga and Liefbroer, 2007): Somewhat similar to entropy.
- Turbulence accounts for state sequencing (which is not the case of the entropy).
- Turbulence accounts of the following two elements:
  - **number of subsequences:**
    - \(x=S,U,M,MC\) - 16 subsequences more turbulent than
      - \(y=S,U,S,C\) - 15 subsequences
  - **variance of duration in each state:**
    - \(S/10\ U/2\ M/132\) is less turbulent than
      - \(S/48\ U/48\ M/48\)
Mining Event or State Sequences
Visualizing and clustering sequence data
Characteristics of sequences

Turbulence - Minimum/maximum

Turbulence minimum, médiane et maximum

Sequences 1−15, sorted by Turbulence

A15 A20 A25 A30 A35 A40 A45

N/N/N/N
Y/N/N/N
N/Y/*/N
Y/Y/N/N
N/N/Y/N
Y/N/Y/N
Y/Y/Y/N
*/*/*/Y

Time
Turbulence - histogram

Turbulence for the sequences in the biofam data set
Turbulence by cohorts

Turbulence selon la cohorte de naissances

Birth cohort

Sequences turbulence
Once you are able to compute 2 by 2 distances between sequences you can among others:

- Cluster sequences
- 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.
Mining Event or State Sequences
Visualizing and clustering sequence data
Distances between sequences: Clustering

Dendrogram, OM1 versus OM3
different indel costs (1 vs 3)
State distribution by age, within cluster
Mining Event or State Sequences

Visualizing and clustering sequence data

Distances between sequences: Clustering

I-plot by cluster

Groupe 1 (sorted)

Groupe 2 (sorted)

Groupe 3 (sorted)

Groupe 4 (sorted)

Groupe 5 (sorted)

Groupe 6 (sorted)
Distribution by birth cohort within each cluster
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 distances $d_{ij}$ best approximate the distances $\delta_{ij}$ between sequences.
  - Plotting the points in the $(f_1, f_2)$ space.
Mining Event or State Sequences
Visualizing and clustering sequence data
Multidimensional Scaling representation of sequences

Multidimensional Scaling
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 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?

Node constraint:
- LH,17
  - \( w = 2 \)

Edge constraints:
- \((0,1,10)\)
- Elastic
- \((0,3)\)
- Parallel

Event constraints:
- \( w = 1 \)
Counting methods (Joshi et al., 2001)

Searching \((U, C)\)

- \(\text{min gap} = 1\), \(\text{max gap} = 2\), \(\text{win size} = 2\)

- \(\text{indiv. with episode} \quad \text{COBJ} = 1\)
- \(\text{windows with episode} \quad \text{CWIN} = 3\)
- \(\text{min win. with episode} \quad \text{CminWIN} = 2\)
- \(\text{distinct occurrences} \quad \text{CDIS}_o = 5\)
- \(\text{dist. occ. without overlap} \quad \text{CDIS} = 3\)
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?
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://mephisto.unige.ch/biominning](http://mephisto.unige.ch/biominning)
- and soon from the CRAN
Thank You!
Divorce, Switzerland, Differences in KM Survival Curves

Language

Non French

French

TW(1) = 22.5, p < .0001

TW(1) = 8.08, p = .0045

TW(1) = 9.77, p = .0018

TW(1) = 37.4, p <= .0001

S < 90% at 21
S(30) = 86%
n = 841
e = 123

S < 90% at 11
S(30) = 74%
n = 174
e = 44

S < 90% at 26
S(30) = 89%
n = 667
e = 79

S < 90% at 11
S(30) = 75%
n = 2175
e = 361

Child

Yes

Language
Clusters and subsequences
Biofam data: Legend

- no event
- left home
- married with/without child
- left home, married
- with child
- left home, with child
- left home, married, child
- divorced

For Further Reading II


For Further Reading III


