Rendering sequences

Gilbert Ritschard

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

Workshop on Sequence Analysis, Lund, May 8-9, 2008

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

Research project

The work presented in this seminar is part of the FNS project

- Mining event histories: Towards new insights on personal Swiss life courses

- Project FN 100012-113998
- Start: February 1, 2007
  End: January 31, 2009

- It is work in progress

Research Team

- Gilbert Ritschard, professor of Statistics for social sciences
  U. of Geneva
  main applicant
- Eric Widmer, professor of Sociology,
  U. of Geneva (since September 2007, formerly at U. Lausanne)
  co-applicant

- Alexis Gabadinho (scientific collaborator, 40%)
  Demography
- Nicolas S. Müller (assistant 50%)
  Sociology, Computer Science
- Matthias Studer (assistant 50%)
  Economics, Sociology

Objectives of the project I

- Explore the feasibility and interest of heuristical data mining approaches for discovering interesting knowledge from event history data.
- Develop data mining techniques by adapting existing ones to the specificity of event history data
- Making newly adapted/developed techniques and validation criteria easily accessible to the social scientist end user through a general toolbox for life course data.
- Demonstrate the complementarity of data mining approaches with classical ones through convincing illustrations of the original new insights that they provide.

Objectives of the project II

- Discover the most characteristic personal Swiss life course patterns, especially broken life patterns, through an overall in-depth analysis of the retrospective biographical database collected by the Swiss Household Panel.

Objective of this presentation

- **Colorize your life courses**
- Preliminary results from the analysis of the retrospective Swiss Household Panel (SHP) survey.
- Focus on visualization of life course data.
- Divorce and de-standardization of life Swiss life courses.

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.

Two broad approaches

- Survival analysis (Event history analysis): Focus on one event (divorce)
  - Which factors influence the hazard rate of experimenting the event?
  - What is the importance of these factors?
- Sequence analysis: sequence describing whole life course.
  - Similarity between pairs of state sequences ($\Rightarrow$ cluster analysis).
  - Typical event pattern.
  - Turbulence and other instability measures of a sequence.

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.	• 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

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

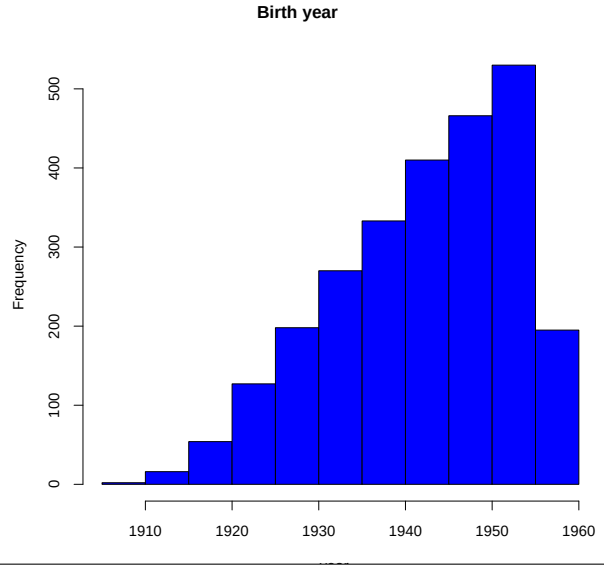
Table: State sequence view, Sandra

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

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 → 2601 remaining individuals, born between 1909 et 1957.

Distribution by birth cohort



Creating 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 :

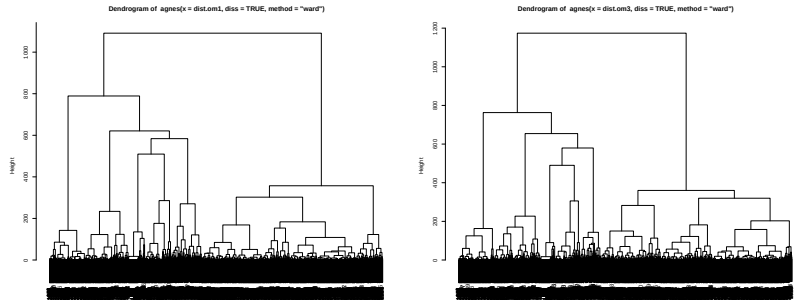
- events:

individual	LHome	marriage	childbirth	divorce
1	1989	1990	1992	NA

- states:

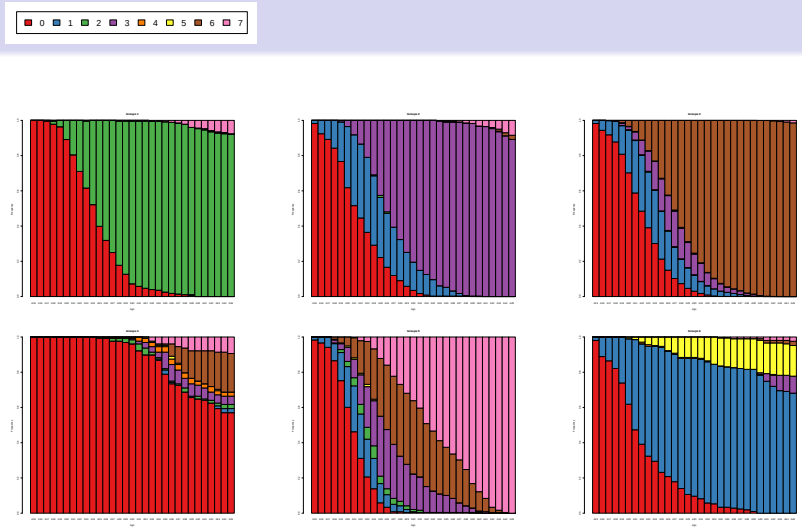
individual	...	1988	1989	1990	1991	1992	1993	...
1	...	0	0	1	3	3	6	...

Dendrogram, OM1 versus OM3



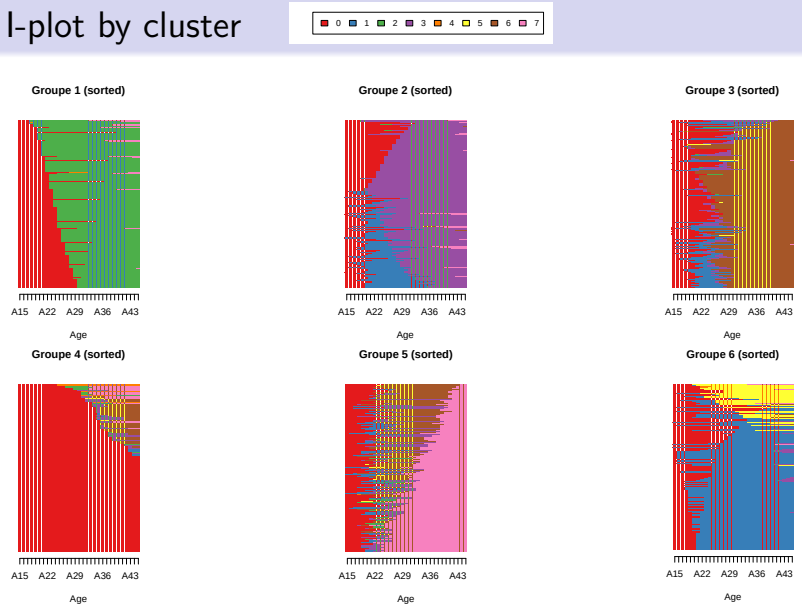
OM1 OM3

State distribution by age, within cluster

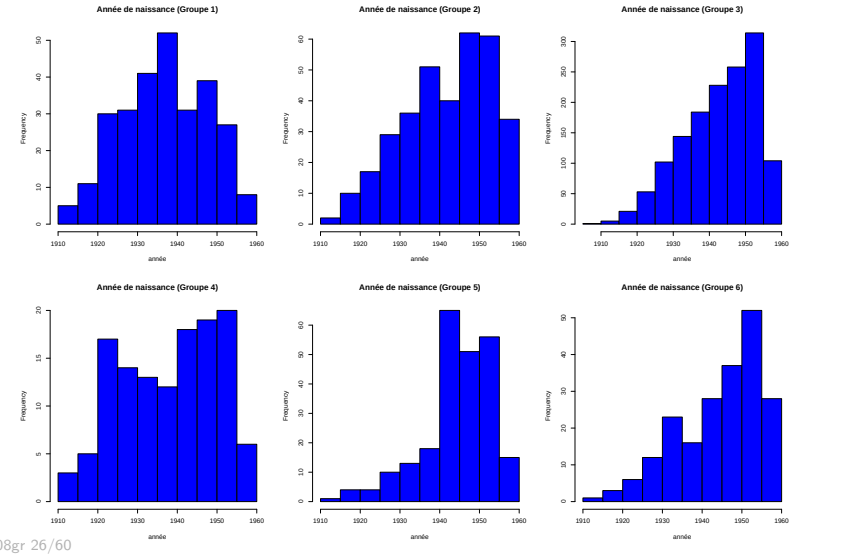


Most frequent sequences by cluster

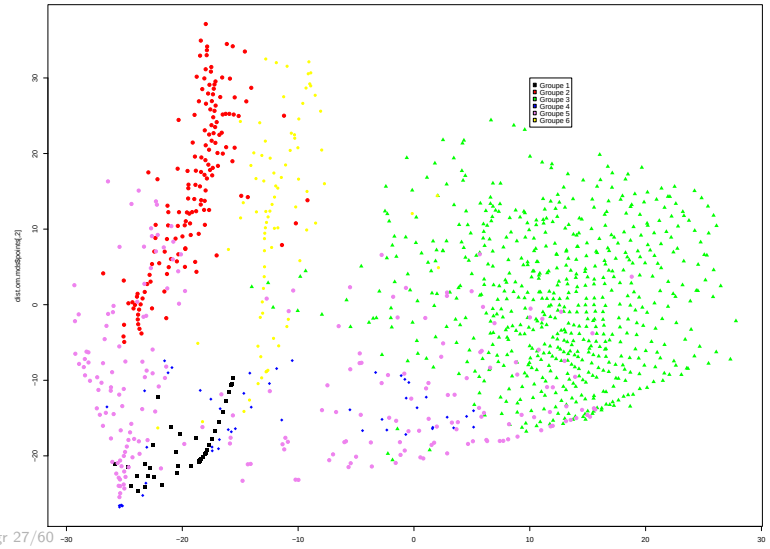
I-plot by cluster



Distribution by birth cohort within each cluster



Multidimensional Scaling



Definitions of event sequences

According to Agrawal and Srikant (1995)

- $I = \{i_1, i_2, ...i_m\}$ set of m distinct events defining an alphabet.
- A transition is a list of non null, not ordered events $(i_1 i_2 ... i_k)$.
- A sequence is an ordered list of transitions.
- A sequence is denoted $\alpha = (\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_q)$, where α_i is a transition.

Sub-sequence (1)

- α is a subsequence of β if each event in α belongs also to β and order between events is preserved.
- We denote it by $\alpha \preceq \beta$.
- Ex: $(B \rightarrow AC)$ is a subsequence of $(AB \rightarrow E \rightarrow ACD)$ since $B \subseteq (AB)$ and $(AC) \subseteq (ACD)$.
- A subsequence is said frequent when it is found in a number of sequences that exceeds a minimal threshold support defined by the user.
- A k-sequence is a subsequence composed of k elements.

Sub-sequence (2)

- Order of event occurrences matters.
- Simultaneity is also is also taken into consideration.
- Presence/absence of intermediary events is allowed.
- Length of gaps does not matter.

Considered events

Code	Description
c	First union
d	Leaving Home
e	First childbirth
m	First marriage
s	Divorce of first marriage

Data format

individual	timestamp	events
5	21	cd
5	28	em
8	16	d
8	30	c
8	32	m
8	40	s

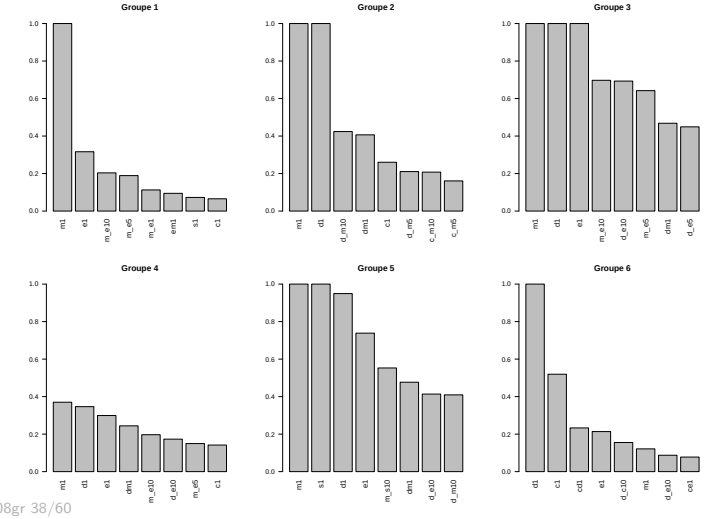
Frequent subsequences

K-sequence	Percentage
m	90.0%
d	85.8%
e	68.1%
m_e10	44.4%
d_e10	43.0%
m_e5	40.1%
dm	36.3%
d_m10	33.1%
d_e5	27.8%
c	22.9%

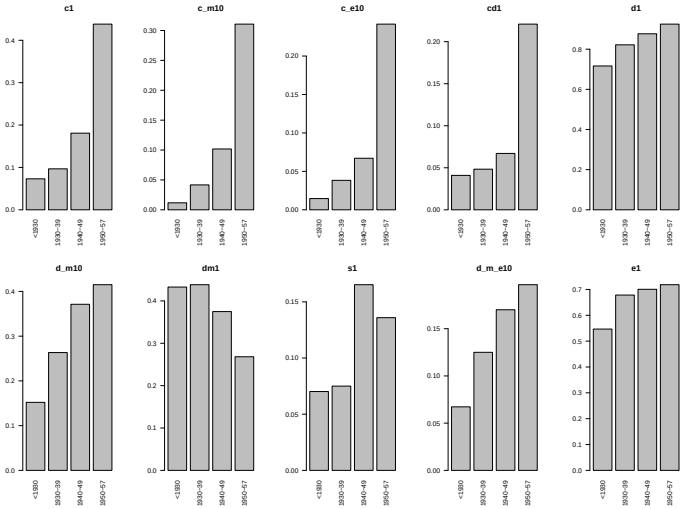
K-sequence	Percentage
dm_e10	19.4%
d_m5	19.2%
m_e1	19.1%
em	18.4%
dm_e5	17.5%
d_m_e10	15.5%
c_m10	14.2%
s	12.3%
de	11.3%
c_m5	11.2%

Clusters of subsequences

Zoom



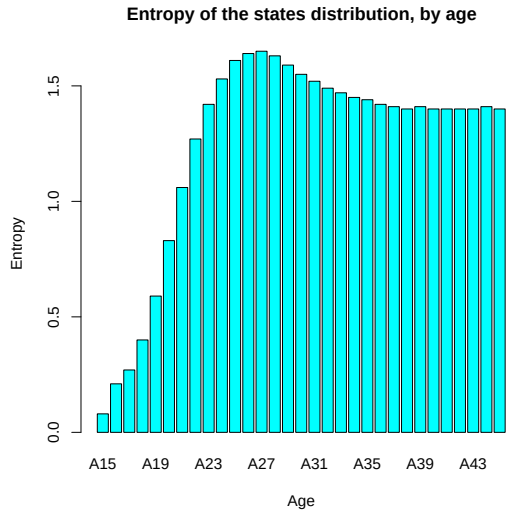
Subsequences and birth cohorts



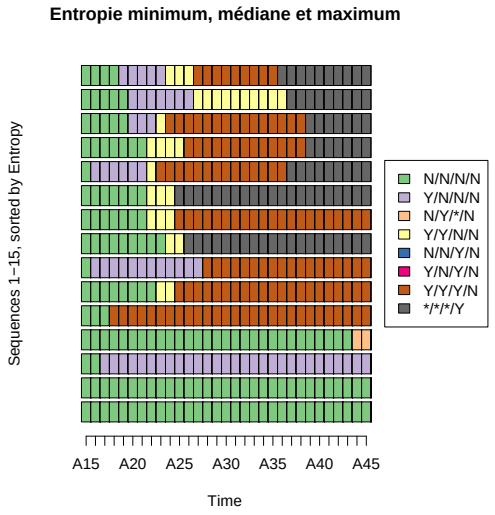
Definition

- Entropy: measure of uncertainty regarding sequence predictability.
- 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.

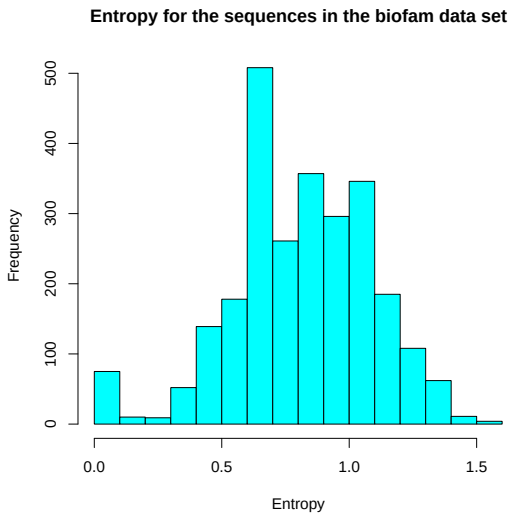
Entropy of the state at each time (age) point



Entropy: Minimum/maximum



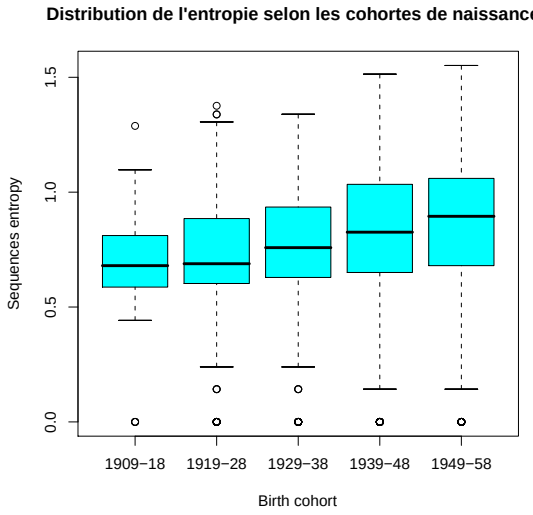
Entropy - histogram



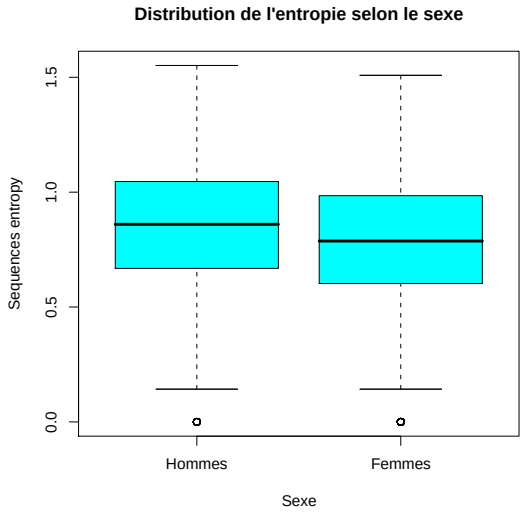
Hypothesis

- Evolutions of familial life trajectories gives rise to an increase in the entropy of individual sequences,
- because they become less predictable and more diversified.

Entropy by birth cohorts



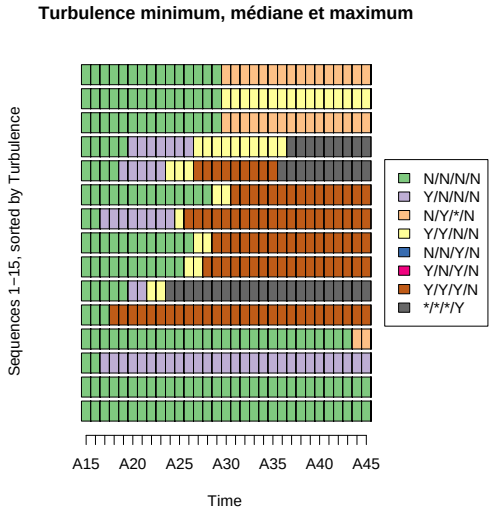
Entropy by sex



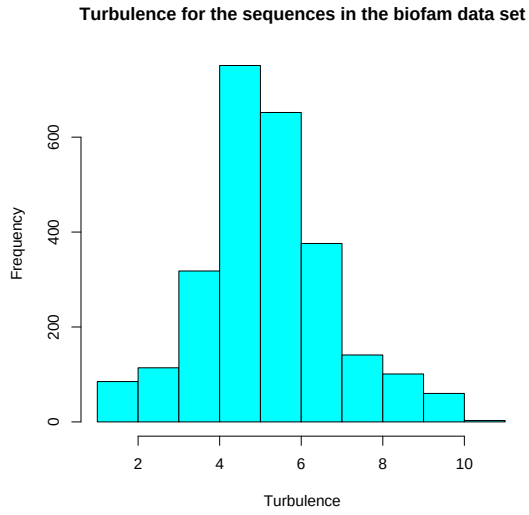
Definition

- 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

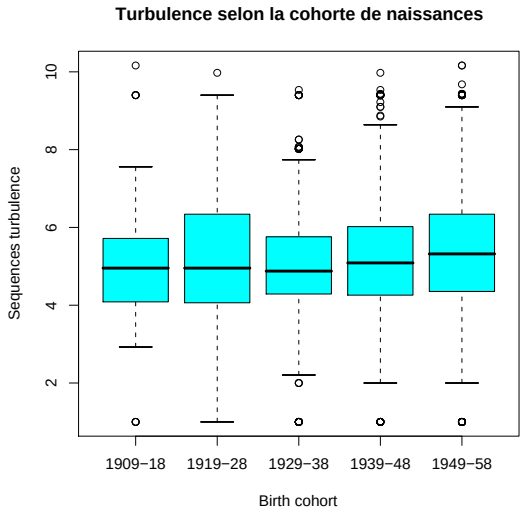
Turbulence - Minimum/maximum



Turbulence - histogram



Turbulence by cohorts



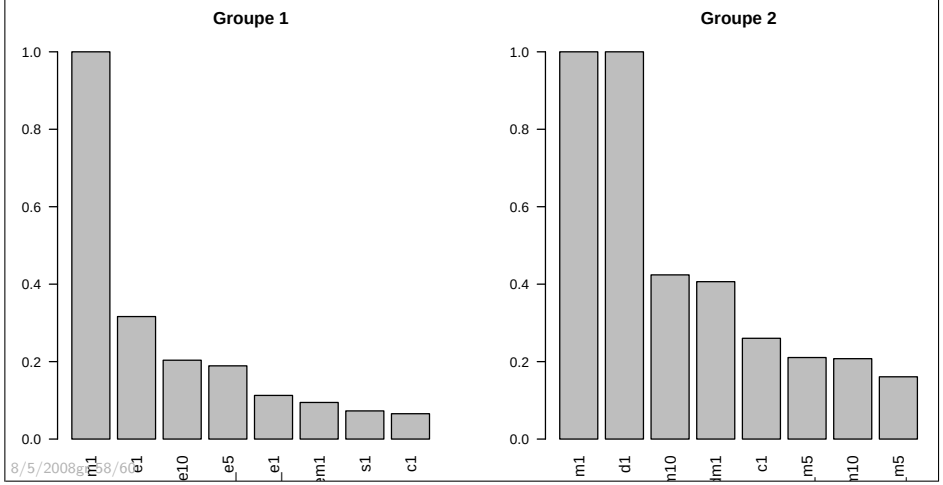
Conclusion

- Work in progress ...
- TraMineR, the toolbox that colorizes your life

THANK YOU!

MERCI !

Clusters et sous-séquences



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

References

Agrawal, R. and R. Srikant (1995). Mining sequential patterns. In P. S. Yu and A. L. P. Chen (Eds.), *Proceedings of the International Conference on Data Engeneering (ICDE), Taipei, Taiwan*, pp. 487–499. IEEE Computer Society.

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.