Evolution of the Swiss Familial Life Course
Learning from innovative approaches in sequence analysis

Gilbert Ritschard Alexis Gabadinho Nicolas S. Müller
Matthias Studer

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

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

Séminaire du LaboDemo, April 22, 2008

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.
- 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.

Tendances de l'évolution des trajectoires de vie familiale

Les techniques d'analyse de séquences permettent de tester certaines hypothèses concernant les évolutions des trajectoires de vie familiale (Elzinga and Liefbroer, 2007):

- Dé-standardisation: certains états et événnements de la vie familiale caractérisent des proportions moins importantes de la population, se produisent à des âges plus dispersés et leurs durés sont également de plus en plus dispersées.
- Dé-institutionalisation: l'organisation sociale et temporelle des trajectoires de vie devient de moins en moins guidée par des règles normatives, légales ou institutionnelles
- Différentiation: le nombre d'étapes et d'états distincts traversés par les individus augmente.

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.

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 $h(t) = p(T = t \mid T \geq t)$

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

SHP biographical retrospective survey

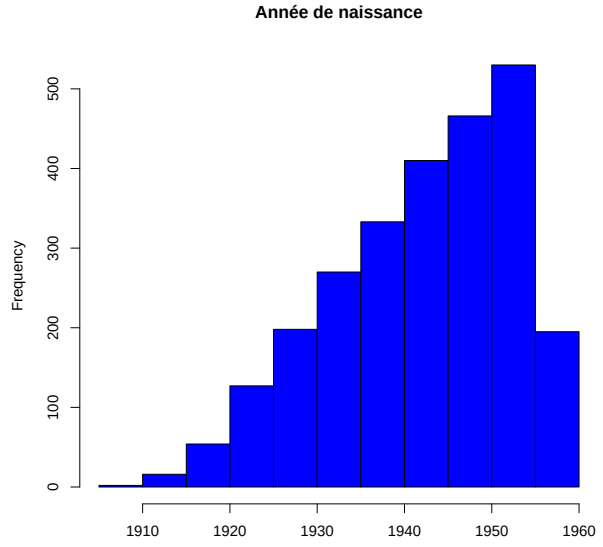
http://www.swisspanel.ch

- 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)

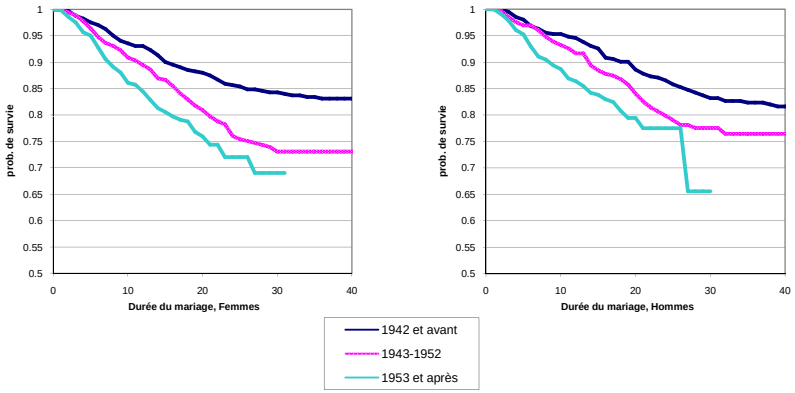
	men	women	Total
Total	1414	1656	3070
1st marriage dissolution	231	308	539
	16.3%	18.6%	17.6%

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

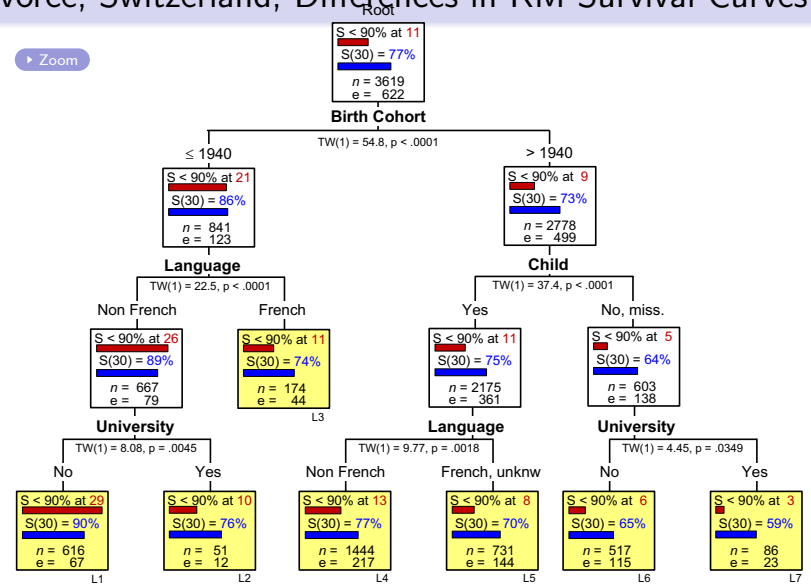
Survival trees: Principle

- Target is survival curve or some other survival characteristic.
- Aim: Partition data set into groups that
- differ as much as possible (max 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

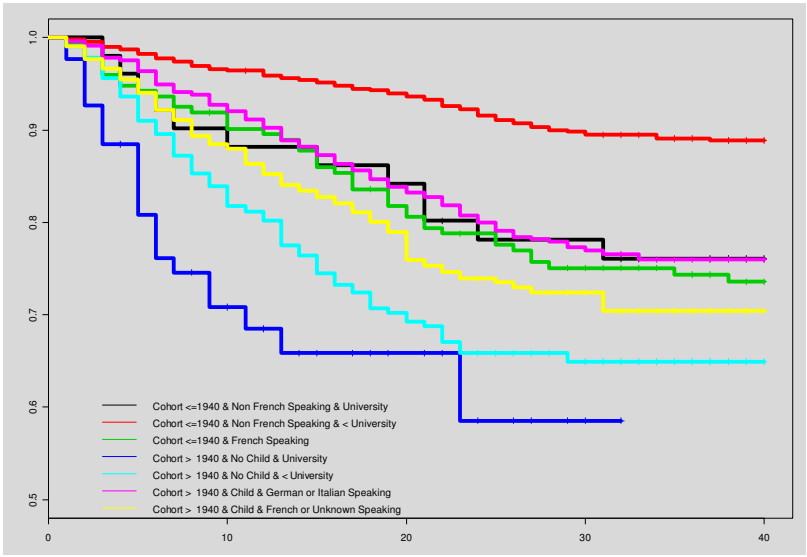
$$TW = \sum_i \frac{w_i\left(d_{i1} - \mathrm{E}(D_i)\right)}{\left(w_i^2 \,\mathrm{var}(D_i)\right)^{1/2}}$$

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

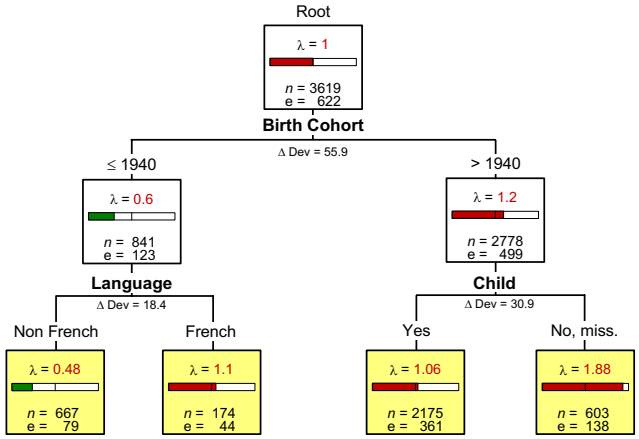
Divorce, Switzerland, Differences in KM Survival Curves I



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

Présentation des données "BioFam"

- Les données utilisées ont été collectées dans le cadre du projet "Vivre en suisse 1990-2000", piloté par le Panel suisse de ménages et supporté par le FNS, l'Office fédéral de la statistique et l'Université de Neuchâtel.
- Enquête biographique rétrospective du Panel suisse des ménages (5560 individus)
- Evénements familiaux retenus : le départ du foyer parental, le premier enfant, le premier mariage et le premier divorce.
- Âge de 15 à 45 ans $\rightarrow$ 2601 individus restants, nés entre 1909 et 1957

Création des séquences

- Exemple de données sous la forme d'événements :

individu	départ	mariage	enfant	divorce
1	1989	1990	1992	NA

Construction des états

Nécessité de construire un état correspondant à chaque combinaison de variables :

	départ	mariage	enfant	divorce
0	non	non	non	non
1	oui	non	non	non
2	non	oui	oui/non	non
3	oui	oui	non	non
4	non	non	oui	non
5	oui	non	oui	non
6	oui	oui	oui	non
7	oui/non	oui	oui/non	oui

Passage des événements aux états

Exemple de transformation :

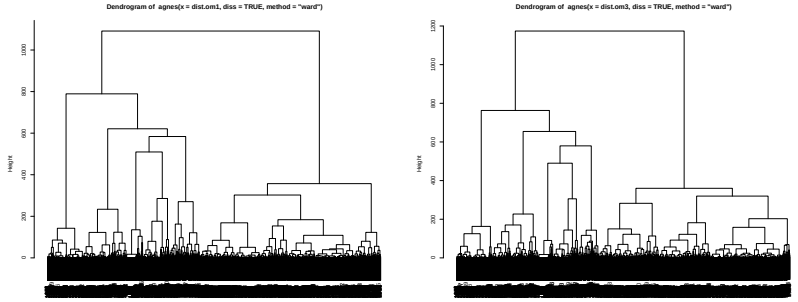
- événements :

individu	départ	mariage	enfant	divorce
1	1989	1990	1992	NA

- états :

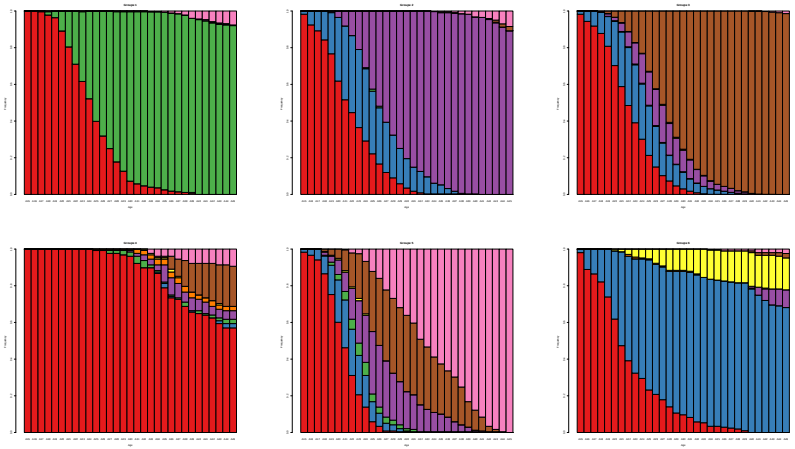
individu	...	1988	1989	1990	1991	1992	1993	...
1	...	0	0	1	3	3	6	...

Dendrogram, OM1 versus OM3

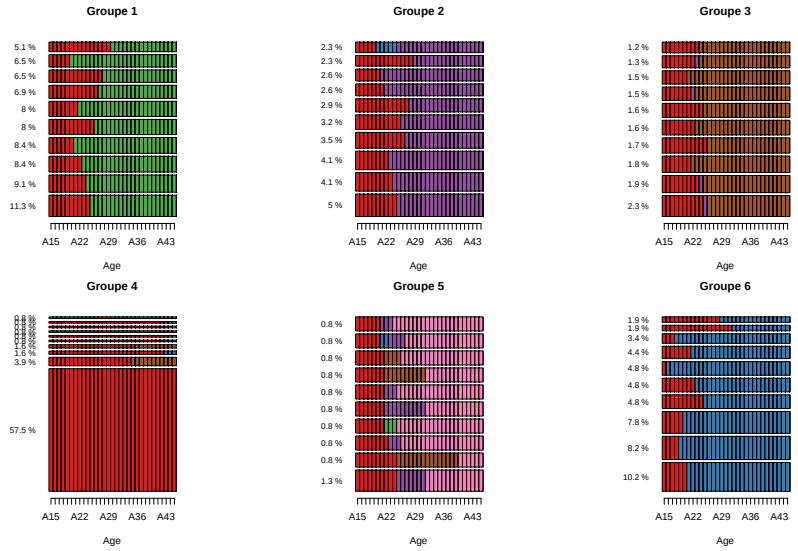


OM1 OM3

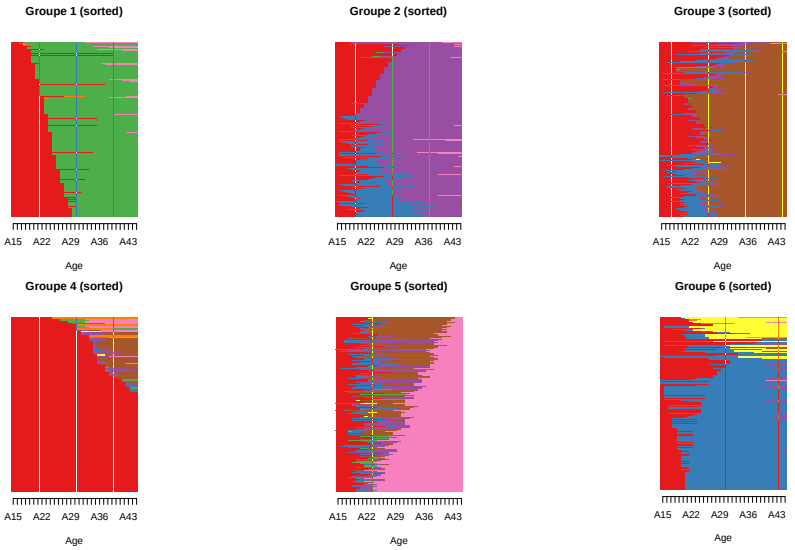
State distribution by age, inside cluster



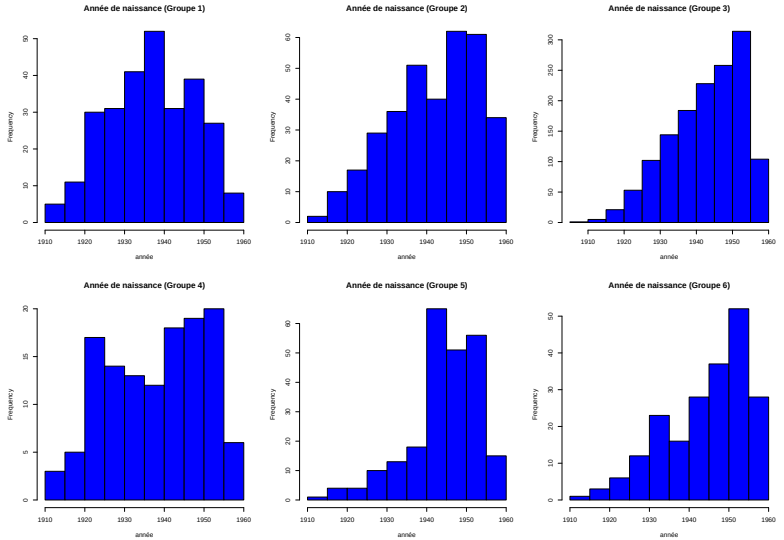
Most frequent sequences by cluster



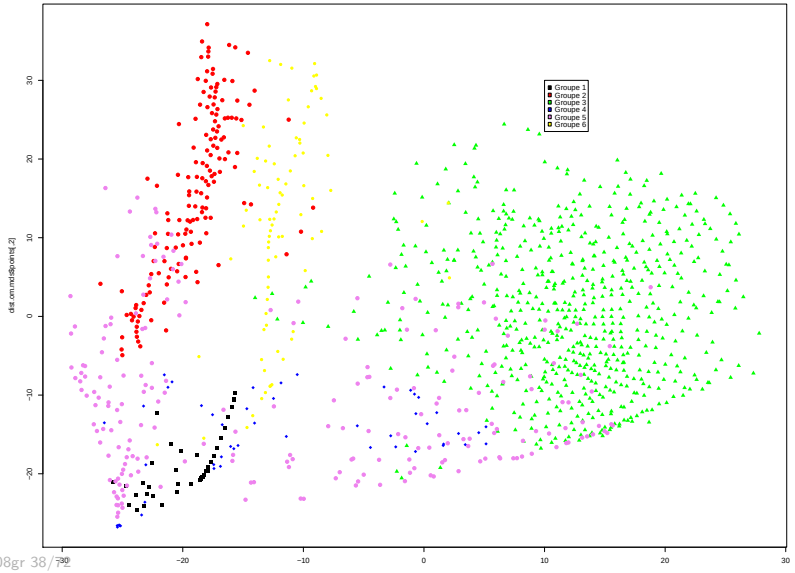
I-plot by cluster



Distribution by birth cohort inside each cluster



Multidimensional Scaling



Définitions de séquence d'événements

Selon Agrawal and Srikant (1995)

- Soit $I = i_1, i_2, ...i_m$ un ensemble de m événements distincts qui définit un alphabet.
- Une transition est une liste non nulle et non ordonnée d'événements notée $(i_1 i_2 ... i_k)$.
- Une séquence est une liste ordonnée de transitions.
- Une séquence α est notée $(\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_q)$, α_i étant une transition.

Sous-séquence

- α est une sous-séquence de β si chaque événement de α est un sous-ensemble de β et que l'ordre des événements est conservé. On le note $\alpha \preceq \beta$.
- Ex : $(B \rightarrow AC)$ est une sous-séquence de $(AB \rightarrow E \rightarrow ACD)$ puisque $B \subseteq (AB)$ et que $(AC) \subseteq (ACD)$.
- Une sous-séquence est dite fréquente quand on l'observe dans un nombre de séquences supérieur à un support minimum défini par l'utilisateur.
- Une k-séquence est une sous-séquence composée de k éléments.

Sous-séquence

- L'ordre d'apparition des événements est pris en compte.
- La simultanéité également.
- La présence/absence d'événements intermédiaire est prise en compte.
- On ne distingue pas la longueur des sauts.

Événements retenus

Code	Description
c	Première mise en couple
d	Première indépendance des parents
e	Naissance du premier enfant
m	Premier mariage
s	Divorce du premier mariage

Format des données

individu	timestamp	événements
5	21	cd
5	28	em
8	16	d
8	30	c
8	32	m
8	40	s

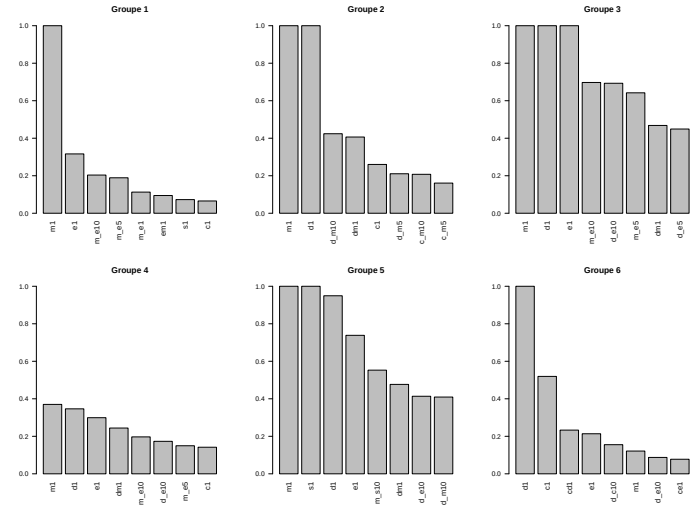
Sous-séquences fréquentes

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%

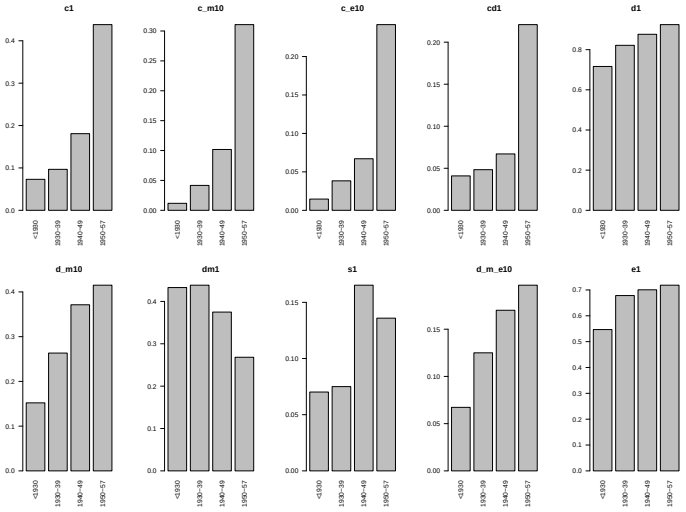
K-sequence	Percentage
dm_e10	19.4%
d_m5	19.2%
m_e1	19.1%
em	18.4%
dm_e5	17.5%
d_m_e10	15.5%
c_m10	14.2%
s	12.3%
de	11.3%
c_m5	11.2%

Clusters et sous-séquences

Zoom



Sous-séquences et cohorte

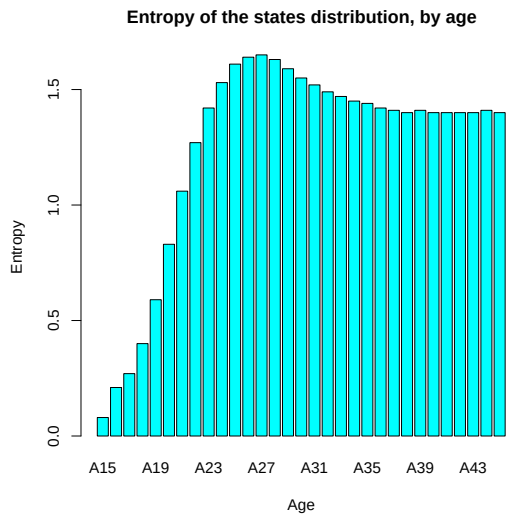


Définition

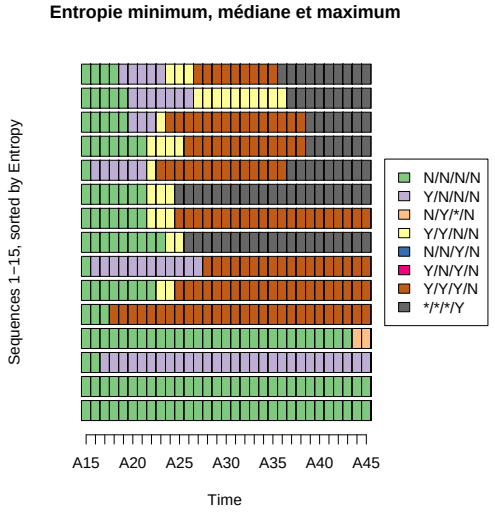
Entropie: mesure de l'incertitude, de la prédictibilité des séquences. Deux mesures peuvent être appliquées aux séquences afin d'en mesurer les caractéristiques.

- entropie des états à chaque âge: l'entropie augmente avec la diversité des états dans lesquels les individus se trouvent.
- entropie des séquences individuelles: l'entropie augmente avec la diversité des états traversés par chaque individu et avec le temps passé dans chacun des états.

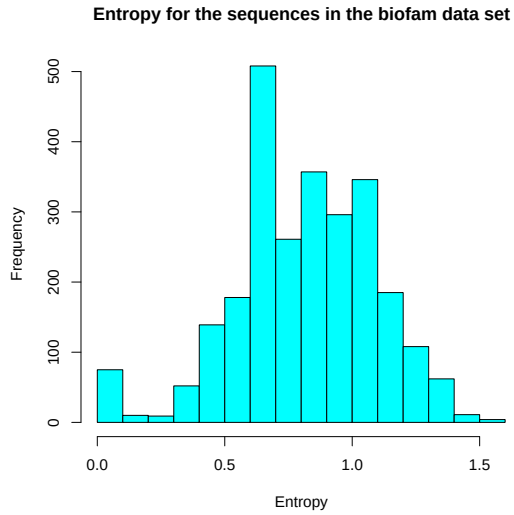
Entropie des états à chaque âge



Entropie - Minimum/maximum



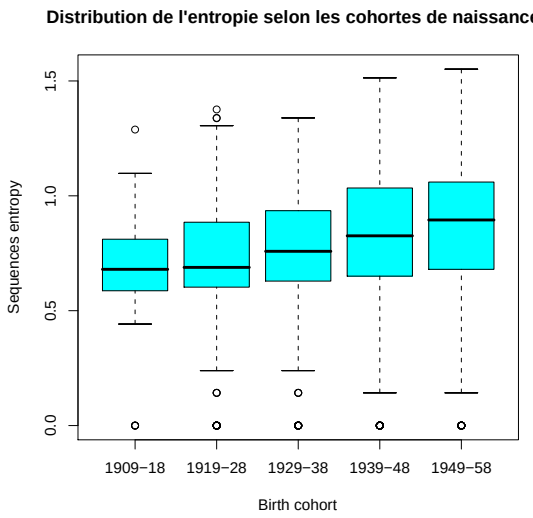
Entropie - histogramme



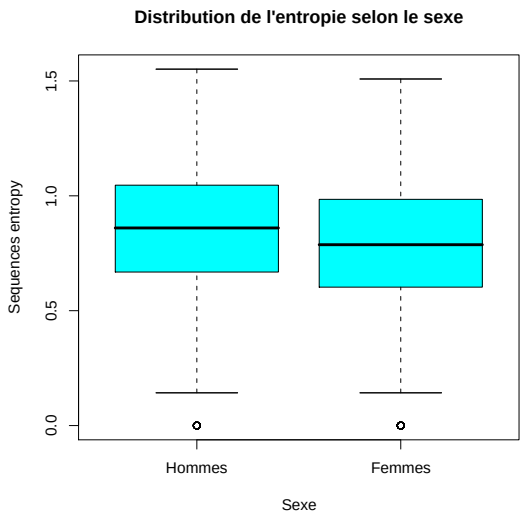
Hypothèse

- Les évolutions des trajectoires de vie familiale entrainent une augmentation de l'entropie des séquences individuelles, car celles-ci sont de moins en moins prédictibles et de plus en plus diversifiées.

Entropie - par cohortes



Entropie - par sexe

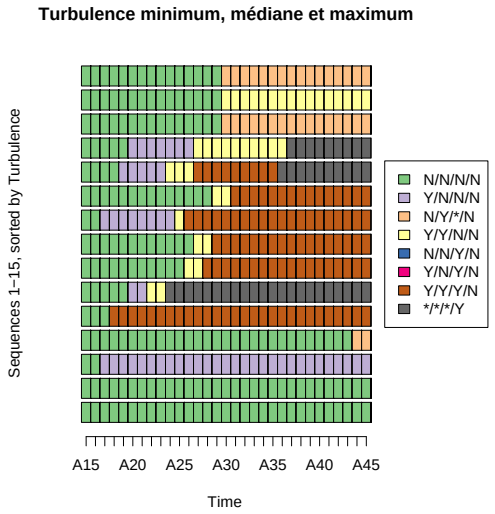


Définition

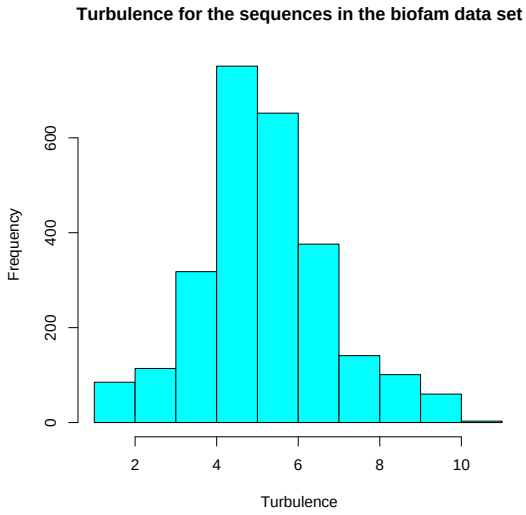
Turbulence (Elzinga and Liefbroer, 2007): mesure qui se rapproche de l'entropie mais qui tient compte en plus de l'ordre dans lequel les états se succèdent. Cette mesure comporte deux éléments:

- le nombre de sous-séquences:
  x=S,U,M,MC - 16 sous-séquences plus turbulente que
  y=S,U,S,C - 15 sous-séquences
- la variance des durées passées dans chaque état:
  S/10 U/2 M/132 trajectoire moins turbulente que
  S/48 U/48 M/48

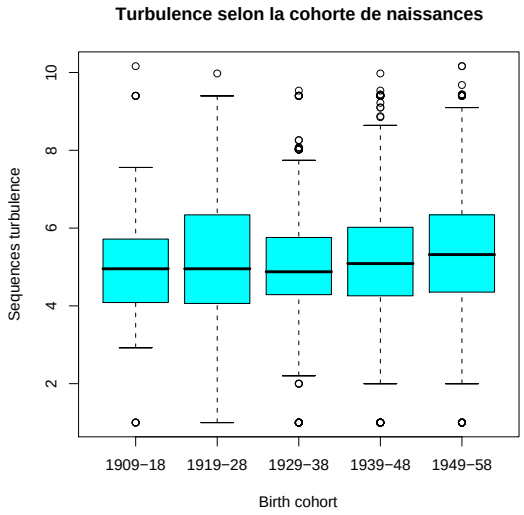
Turbulence - Minimum/maximum



Turbulence - histogramme



Turbulence par cohortes



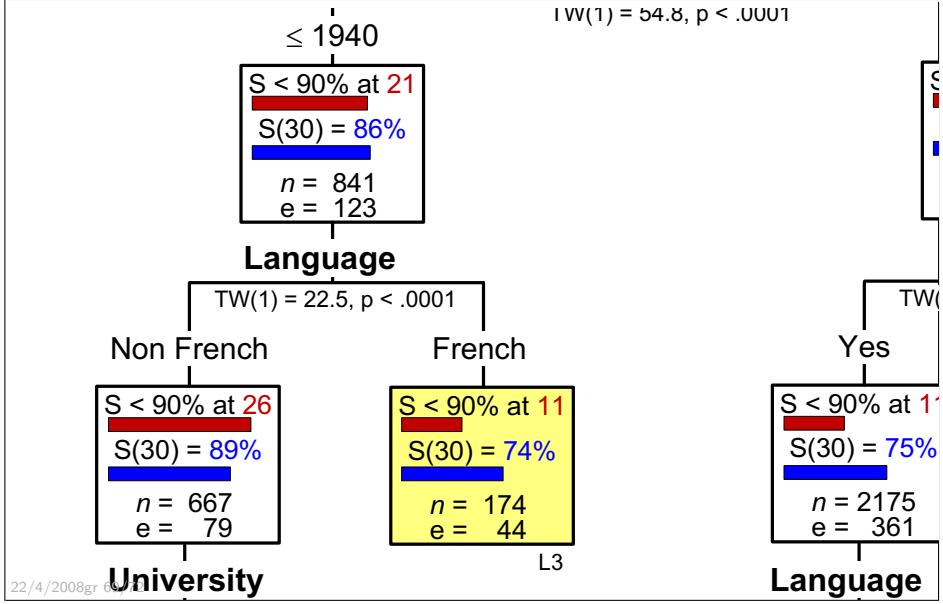
Conclusion

- Le travail se poursuit ...
- TraMineR, la boite à outils qui vous colorie la vie
- sera présentée Jeudi 24 avril, à 12h15, salle M-4220

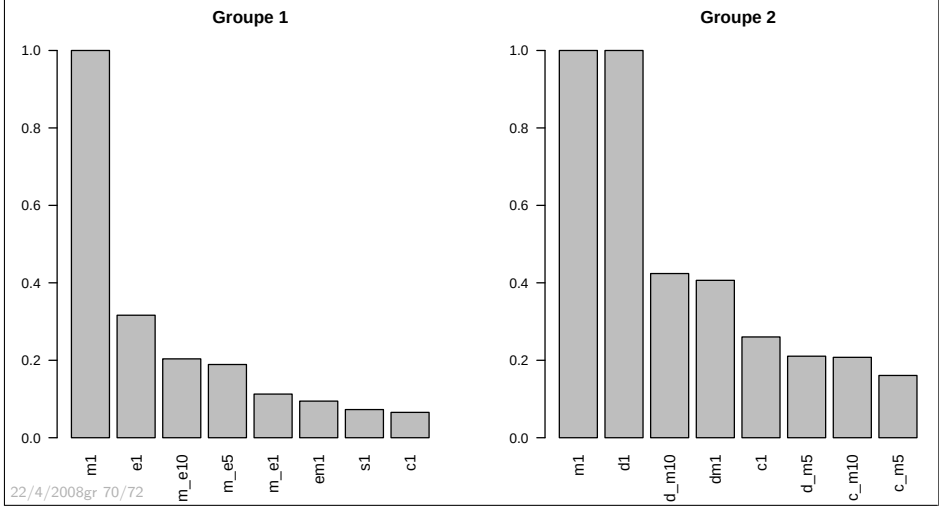
THANK YOU!

MERCI !

Divorce, Switzerland, Differences in KM Survival Curves I



Clusters et sous-séquences



Biofam data: Legend

- aucun événement
- parti
- marié avec ou sans enfants
- parti, marié
- avec enfants
- parti, avec enfants
- parti, marié, avec enfants
- divorcé

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.

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. 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.

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

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