Experiences with some longitudinal exploratory data mining problems

Gilbert Ritschard
Matthias Studer and Emmanuel Rousseaux

NCCR LIVES and Institute for demographic and life course studies
University of Geneva
http://mephisto.unige.ch

COMPSTAT, Geneva, August 19-22, 2014

Objectives (continued)

- Recall kind of results that can be obtained by mining event subsequences
  - most frequent subsequences
  - association rules between subsequences
    (cf. Emmanuel Rousseaux, Session CS75, Friday 22, 9 am)
  - subsequences that best discriminate groups (provided covariate)
- Problem How to deal with nested subsequences?
  - If (LHome) → (Marriage) → (Childbirth) is frequent, shall we also consider people following that path when counting the frequency of subsequence (LHome) → (Marriage)?
  - Could be more interesting to know how many people with (LHome) → (Marriage), did not have child birth afterwards.

Objectives

- Data-mining-based methods (pattern mining)
  - Discovering interesting information from sequences of life events, i.e., on how people sequence important life events
    - What is the most typical succession of family or professional life events?
    - Are there standard ways of sequencing those events?
    - What are the most typical events that occur after a given subsequence such as after leaving home and ending education?
    - How is the sequencing of events related to covariates?
    - Which event sequencings do best discriminate groups such as men and women?
  - Mining of frequent (Agrawal and Srikant, 1995; Mannila et al., 1995; Bettini et al., 1996; Mannila et al., 1997; Zaki, 2001) and discriminant event subsequences (Ritschard et al., 2013)

Event sequences versus state sequences

- State sequence: states last a whole interval period

age	20	21	22	23	24	25	26
state	2P	2P	A	A	UC	UC	UC

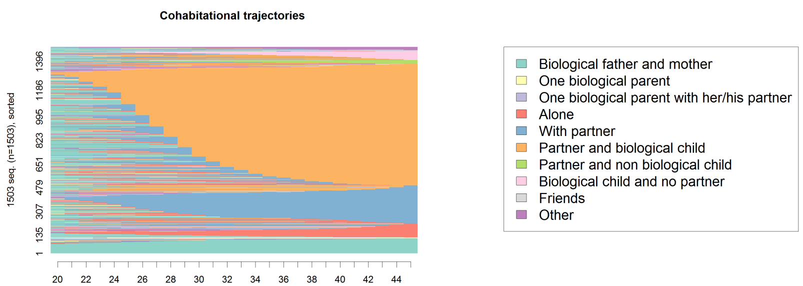
- Event sequence: events occur at a given (time) position
  - Interest in their order, in their sequencing
  - Can be time stamped (TSE)

id	Timestamp	Event
101	22	Leaving Home
101	24	Start living with partner
101	24	Childbirth

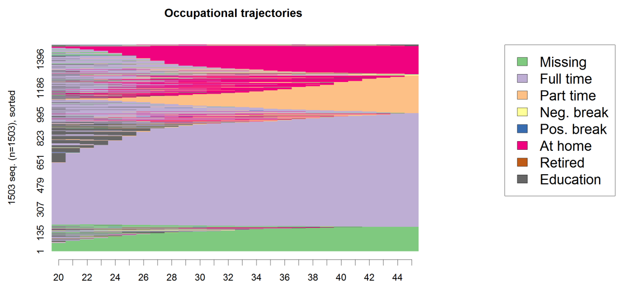
The Biographical SHP Data

- Sequences derived from the biographical survey conducted in 2002 by the Swiss Household Panel www.swisspanel.ch
- Retain the 1503 cases studied in Widmer and Ritschard (2009) with techniques for state sequences
- Two channels: Cohabitational and occupational
- Only individuals aged 45 or more at survey time
- Focus on life trajectory between 20 and 45 years
- Granularity is yearly level

The Cohabitational State Sequences



The Occupational State Sequences



Short and long state labels

Cohabitational		Occupational	
2P	Biological father and mother	Mi	Missing
1P	One biological parent	FT	Full time
PP	One biological parent with her/his partner	PT	Part time
A	Alone	NB	Neg. break
U	With partner	PB	Pos. break
UC	Partner and biological child	AH	At home
UN	Partner and non biological child	RE	Retired
C	Biological child and no partner	ED	Education
F	Friends		
O	Other		

Events associated to cohabitational state transitions

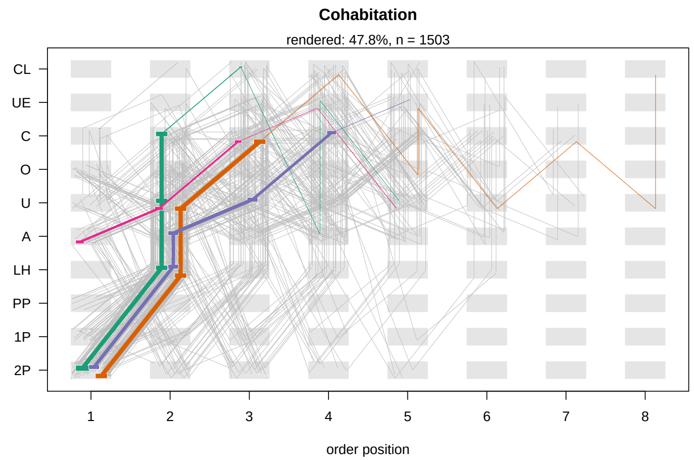
- For cohabitational trajectories, we convert states to events by defining the events associated to the state transitions

	2P	1P	PP	A	U	UC	UN	C	F	O
2P	"2P"	"1P"	"PP"	"LH,A"	"LH,U"	"LH,U,C"	"LH,U,C"	"LH,C"	"LH,A"	"LH,O"
1P	"2P"	"1P"	"PP"	"LH,A"	"LH,U"	"LH,U,C"	"LH,U,C"	"LH,C"	"LH,A"	"LH,O"
PP	"2P"	"1P"	"PP"	"LH,A"	"LH,U"	"LH,U,C"	"LH,U,C"	"LH,C"	"LH,A"	"LH,O"
A	"2P"	"1P"	"PP"	"A"	"U"	"U,C"	"U,C"	"C"	""	"O"
U	"2P"	"1P"	"PP"	"UE,A"	"U"	"C"	"C"	"C"	"UE,A"	"UE,O"
UC	"2P"	"1P"	"PP"	"UE,CL,A"	"CL"	"U,C"	"CL,C"	"UE"	"UE,CL,A"	"UE,CL,O"
UN	"2P"	"1P"	"PP"	"UE,CL,A"	"CL"	"C"	"U,C"	"UE,C"	"UE,CL,A"	"UE,CL,O"
C	"2P"	"1P"	"PP"	"CL,A"	"CL,U"	"U"	"CL,C"	"C"	"CL,A"	"CL,O"
F	"2P"	"1P"	"PP"	""	"U"	"U,C"	"U,C"	"C"	"A"	"O"
O	"2P"	"1P"	"PP"	"A"	"U"	"U,C"	"U,C"	"C"	"A"	"O"

- For occupational trajectories, we assign an event to the start of each spell in a state.

Rendering occupational event sequences
(Bürgin and Ritschard, 2014)

Rendering cohabitational event sequences
(Bürgin and Ritschard, 2014)



Frequent subsequences versus Frequent itemsets - 1

- Mining of frequent itemsets and association rules has been popularized in the 90's with the work of Agrawal and Srikant (1994); Agrawal et al. (1995) and their Apriori algorithm.
  - Find out items that customers often buy together
  - Symptoms that often occur together before a failure

Frequent subsequences versus Frequent itemsets - 2

- Interest on sequences for accounting for the time order of the buys or symptoms
- Mining typical event sequences is a specialized case of the mining of frequent itemsets
  - More complicated however
  - Must specify a counting method: How should we count multiple occurrences of a subsequence in a same sequence?
  - Which time span should be covered? Maximal gap between two events? ...
- Best known algorithms by Bettini et al. (1996), Srikant and Agrawal (1996), Mannila et al. (1997) and Zaki (2001).
- Algorithm in TraMineR is adaptation of the tree search described in Masseglia (2002).

Subsequence

- A subsequence B of a sequence A is an event sequence such that
  - each event of B is an event of A,
  - events of B are in same order as in A.

Example

A (LHome, Union) → (Marriage) → (Childbirth).

B (LHome, Marriage) → (Childbirth).

C (LHome) → (Childbirth).

- C is a subsequence of A and B, since order of events is respected.
- B is not a subsequence of A, since we don't know in B whether "LHome" occurs before "Marriage".

Events and transitions

- Event sequence: ordered list of transitions.
- Transition (transaction): a set of non ordered events.

Example

(LHome, Union) → (Marriage) → (Childbirth)

- (LHome, Union) and (Marriage) are transitions.
- "LHome", "Union" et "Marriage" are events.

Frequent and discriminant subsequences

- Support of a subsequence: number of sequences that contain the subsequence.
  - Frequent subsequence: sequence with support greater than a minimal support.
  - A subsequence is discriminant between groups when its support varies significantly across groups.

Frequent cohabitational subsequences

10 most frequent subsequences, min support = 50

- With at least 2 events

  Remember that we assigned the state at age 20 as start event

	Subsequence	Support	Count	#Transitions	#Events
1	(2P) → (LH)	0.621	934	2	2
2	(2P) → (U)	0.582	874	2	2
3	(2P) → (C)	0.477	717	2	2
4	(LH,U)	0.454	682	1	2
5	(U) → (C)	0.429	645	2	2
6	(2P) → (LH,U)	0.392	589	2	3
7	(LH) → (C)	0.382	574	2	2
8	(A) → (U)	0.376	565	2	2
9	(2P) → (LH) → (C)	0.325	489	3	3
10	(C,U)	0.291	437	1	2

Frequent occupational subsequences

Most frequent subsequences, min support = 50

- With at least 2 events

  Remember that we assigned the state at age 20 as start event

	Subsequence	Support	Count	#Transitions	#Events
1	(ED) → (FT)	0.283	425	2	2
2	(FT) → (AH)	0.265	398	2	2
3	(FT) → (PT)	0.219	329	2	2
4	(AH) → (PT)	0.130	195	2	2
5	(ED) → (AH)	0.113	170	2	2
6	(ED) → (PT)	0.112	168	2	2
7	(FT) → (FT)	0.112	168	2	2
8	(FT) → (AH) → (PT)	0.105	158	3	3
9	(FT) → (ED)	0.073	109	2	2
10	(ED) → (FT) → (PT)	0.071	107	3	3

Frequent subsequences easily extends to multichannel

- Here we have cohabitational and occupational trajectories
- Merging the two series of time stamped events
  - we get mixed cohabitational/occupational event sequences

Merged cohabitational and occupational sequences

12 most frequent subsequences, min support 150

	Subsequence	Support	Count	#Transitions	#Events
1	(FT) → (U)	0.695	1045	2	2
2	(2P) → (LH)	0.621	934	2	2
3	(FT) → (C)	0.583	876	2	2
4	(2P) → (U)	0.582	874	2	2
5	(FT) → (LH)	0.555	834	2	2
6	(2P) → (C)	0.477	717	2	2
7	(LH,U)	0.454	682	1	2
8	(U) → (C)	0.429	645	2	2
9	(2P) → (LH,U)	0.392	589	2	3
10	(LH) → (C)	0.382	574	2	2
11	(2P,FT)	0.378	568	1	2
12	(A) → (U)	0.376	565	2	2

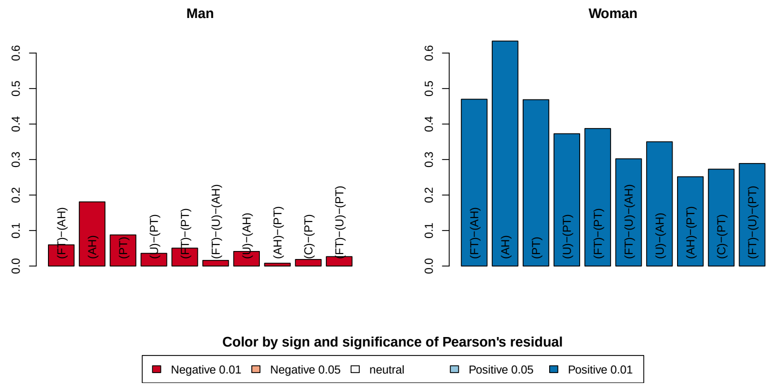
Interesting frequent subsequences

- To get interesting knowledge we need to compare
  - most frequent subsequences
  - with longer less frequent subsequences in which they are included.
- For example,

	Subsequence	Support	Count	#Transitions	#Events
2	(2P) → (LH)	0.621	934	2	2
4	(2P) → (U)	0.582	874	2	2
9	(2P) → (LH,U)	0.392	589	2	3

- Here, we know that
  - among the 62.1% who left home (LH) after living with both parents (2P) when 20 years old
  - $39.2/62.1 = 63\%$ left home to start a union the same year

Mixed events: Subsequences that best discriminate sex at the 0.1% level

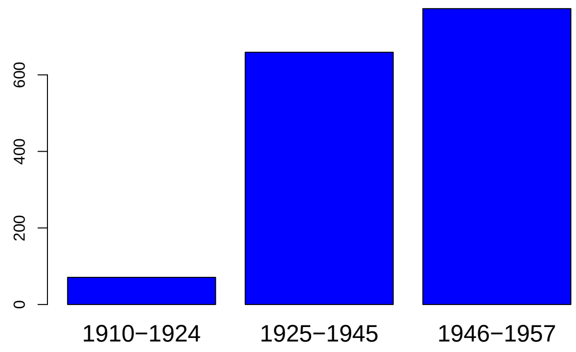


Mixed events: Subsequences that best discriminate sex

	Subsequence	Chi-2	Support	Freq. Men	Freq. Women	Diff
1	(FT) → (AH)	322.7	0.26	0.060	0.470	-0.410
2	(AH)	317.5	0.41	0.181	0.634	-0.453
3	(PT)	269.7	0.28	0.088	0.469	-0.381
4	(U) → (PT)	260.4	0.20	0.036	0.373	-0.337
5	(FT) → (PT)	247.5	0.22	0.051	0.387	-0.337
6	(FT) → (U) → (AH)	228.2	0.16	0.016	0.302	-0.286
7	(U) → (AH)	226.0	0.20	0.041	0.350	-0.309
8	(AH) → (PT)	195.5	0.13	0.008	0.252	-0.244
9	(C) → (PT)	193.3	0.15	0.019	0.273	-0.254
10	(FT) → (U) → (PT)	192.7	0.16	0.027	0.289	-0.262

- Mainly occupational events (FT, PT and AH)
- In conjunction with a few cohabitational ones (U and C)

Birth cohort distribution

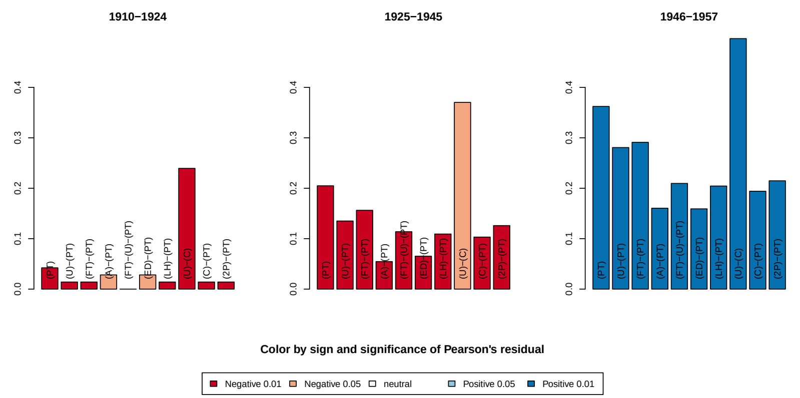


Mixed events: Subsequences that best discriminate birth cohorts

	Subsequence	Chi-2	Support	1910-25	1926-45	1946-57
1	(PT)	64.5	0.28	0.042	0.205	0.362
2	(U) → (PT)	63.0	0.20	0.014	0.135	0.281
3	(FT) → (PT)	56.1	0.22	0.014	0.156	0.291
4	(A) → (PT)	46.3	0.11	0.028	0.055	0.160
5	(FT) → (U) → (PT)	38.5	0.16	0.000	0.114	0.210
6	(ED) → (PT)	36.8	0.11	0.028	0.065	0.159
7	(LH) → (PT)	35.9	0.15	0.014	0.109	0.204
8	(U) → (C)	34.2	0.43	0.239	0.370	0.497
9	(C) → (PT)	34.0	0.15	0.014	0.103	0.194
10	(2P) → (PT)	32.7	0.17	0.014	0.126	0.215

- Mainly emergence of Part-time (PT)

Mixed events: Subsequences that best discriminate birth cohorts



Too many frequent subsequences

- There are often too many frequent subsequences!
- How can we structure those subsequences?
  - Eliminate redundant subsequences: When you experience one subsequence you also experiment all its subsequences.
    - Count only maximal subsequences
    - If subsequence (FT) → (AH) → (PT) is observed,
    - we would not count the occurrence of (FT) → (AH), (FT) → (PT) or (AH) → (PT)

Frequent maximal subsequence: Definition

Frequent maximal subsequence

- A subsequence is frequent maximal if frequent when in each sequence we count only those subsequences that are not themselves a subsequence of another frequent subsequence present in the same sequence.

- Example: The subsequence (2P) → (LH) will be considered a maximal subsequence of sequences which do not also have a frequent supersequence such as (2P) → (LH,U).

Maximal frequent sequence in pattern mining

- Our definition of a frequent maximal subsequence differs from the notion of maximal frequent sequence used in pattern mining, where a frequent sequence is said maximal if none of its supersequence is frequent.
- In pattern mining, if *s* is a maximal frequent sequence, then none of its subsequences is a maximal frequent subsequence, even if it occurs frequently in sequences which do not include *s*.
  - e.g., if (U) → (C) is frequent, then (U) would not be considered.
- This is not very useful for life trajectories where we may be interested to know that
  - It is frequent to start a union (U) without having a child afterwards (U) → (C)

Frequent maximal subsequences: algorithm

1. Find frequent subsequences for the selected support
2. Starting from the longest obtained frequent subsequence
   - Adjust the count of each of its subsequence (by reducing their counts by the number of occurrences of the considered frequent sequence).
   - Delete from the list subsequences with counts falling below the support threshold.
3. Iterate on frequent subsequences ordered in decreasing order of length (using their already adjusted counts)

Max subsequences, cohabitational-occupational events

12 most frequent maximal subsequences, min support 150

	Subsequence	Support	Count	#Transitions	#Events
1	(2P) → (C,LH,U)	0.160	241	2	4
2	(FT) → (U) → (AH)	0.159	239	3	3
3	(FT) → (U) → (PT)	0.158	237	3	3
4	(FT) → (A,LH) → (U)	0.152	228	3	4
5	(2P,ED) → (FT) → (U)	0.140	210	3	4
6	(FT) → (C,LH,U)	0.140	210	2	4
7	(AH) → (C)	0.137	206	2	2
8	(2P) → (LH) → (AH)	0.133	200	3	3
9	(AH) → (U)	0.130	195	2	2
10	(2P,FT) → (LH,U)	0.129	194	2	4
11	(2P) → (LH) → (PT)	0.128	193	3	3
12	(2P,FT) → (AH)	0.126	190	2	3

Max subsequences, cohabitational-occupational events

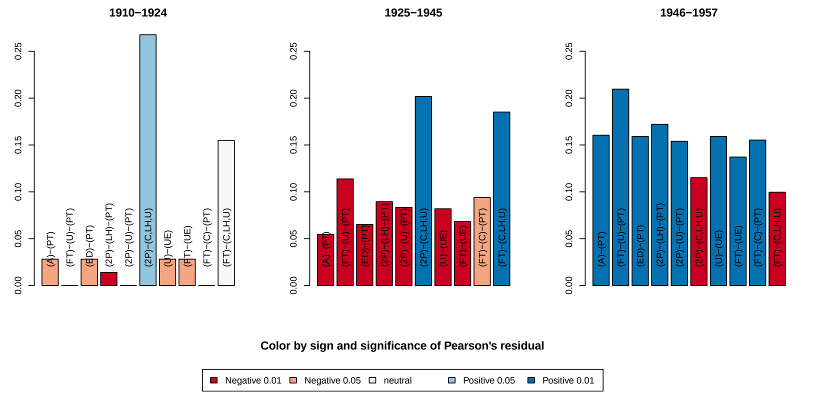
12 most frequent maximal subsequences, min support 200

	Subsequence	Support	Count	#Transitions	#Events
1	(2P,FT) → (LH,U)	0.229	344	2	4
2	(A) → (U) → (C)	0.194	291	3	3
3	(2P,ED) → (LH)	0.189	284	2	3
4	(ED) → (FT) → (C)	0.189	284	3	3
5	(2P) → (A,LH) → (U)	0.181	272	3	4
6	(2P,FT) → (LH) → (C)	0.178	268	3	4
7	(2P) → (LH,U) → (C)	0.168	253	3	4
8	(2P) → (PT)	0.166	250	2	2
9	(FT) → (LH,U) → (C)	0.166	250	3	4
10	(2P) → (C,LH,U)	0.160	241	2	4
11	(FT) → (U) → (AH)	0.159	239	3	3
12	(FT) → (U) → (PT)	0.158	237	3	3

Solutions change with chosen support

- As seen, solutions vary with chosen minsupport
- For minsupport $= 0$, we get the set of complete event sequences.
- We are working on criteria to select an optimal minsupport
  - to minimize the number of subsequences with no representative
  - maximize the average number of representatives
  - ...

Frequent max-subsequences discriminating between cohorts



Frequent max-subsequences discriminating birth cohorts

Minsupport=150

	Subsequence	Chi-2	Support	1910-25	1926-45	1946-57
1	(A) → (PT)	46.3	0.11	0.028	0.055	0.160
2	(FT) → (U) → (PT)	38.5	0.16	0.000	0.114	0.210
3	(ED) → (PT)	36.8	0.11	0.028	0.065	0.159
4	(2P) → (LH) → (PT)	30.4	0.13	0.014	0.090	0.172
5	(2P) → (U) → (PT)	27.0	0.12	0.000	0.083	0.154
6	(2P) → (C,LH,U)	26.2	0.16	0.268	0.202	0.115
7	(U) → (UE)	26.1	0.12	0.028	0.082	0.159
8	(FT) → (UE)	22.9	0.10	0.028	0.068	0.137
9	(FT) → (C) → (PT)	22.8	0.12	0.000	0.094	0.155
10	(FT) → (C,LH,U)	21.8	0.14	0.155	0.185	0.100

Sequential association rules

Sequential association rule

A rule $\text{subseq}_1 \rightarrow \text{subseq}_2$ such that

1. Has a minimal support
2. When subseq_1 occurs, it is most often followed by subseq_2

- Extracted from frequent sequences.
- Extraction criteria:
  - Confidence: $p(\text{subseq}_2 \mid \text{subseq}_1)$
  - Lift: $\dfrac{p(\text{subseq}_2 \mid \text{subseq}_1)}{p(\text{subseq}_2)}$
  - ...

Extracting association rules

- From the mined frequent subsequences, we can extract association rules :

```
##                      Rules Support   Conf  Lift
## 153   (2P,ED) => (LH)-(C)     167 0.5719 1.498
## 171     (FT)-(AH) => (PT)     158 0.3970 1.427
## 55        (2P,ED) => (LH)     284 0.9726 1.345
## 74       (2P) => (C,LH,U)     241 0.2349 1.342
## 35     (2P,FT) => (LH,U)      344 0.6056 1.335
## 72         (2P) => (C,LH)     246 0.2398 1.335
## 175          (1P) => (LH)     151 0.9557 1.321
## 177 (2P,FT) => (LH,U)-(C)     150 0.2641 1.306
## 99     (2P) => (A,LH)-(C)     212 0.2066 1.278
## 12        (2P,FT) => (LH)     523 0.9208 1.273
```


Issues with association rules

- Classical definition assume the left hand and the right hand subsequences are frequent.
- Which implication rule should be used?
  - There are over 50 interestingness criteria (Gras' intensity of implication, ...)
- How can we get rules for rare events (or subsequences)?
  (This will be the topic of Rousseaux's presentation)

Conclusion

- Type of outcomes for event sequences
  - frequent episodes
  - discriminant episodes
  - association rules
  - cluster analysis (not addressed in this presentation)
- Complementary insights
  - most common characteristics
  - salient distinctions between groups
  - implication rules between common charcateristics
  - identify types of trajectories
- Easy to extend to other types of analyses (representative sequences, discrepancy analyses, ...)

Conclusion

- Looking at frequent max-subsequences produces more directly interpretable results
- Issue: Solutions vary with the minsupport threshold

Thank You!

References I

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo (1995). Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 307–328. Menlo Park, CA: AAAI Press.

Agrawal, R. and R. Srikant (1994). Fast algorithm for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo (Eds.), *Proceedings 1994 International Conference on Very Large Data Base (VLDB'94), Santiago de Chile*, San-Mateo, pp. 487–499. Morgan-Kaufman.

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.

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.

References II

Bürgin, R. and G. Ritschard (2014). A decorated parallel coordinate plot for categorical longitudinal data. *The American Statistician 68*(2), 98–103.

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2009). Mining sequence data in R with the TraMineR package: A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva.

Mannila, H., H. Toivonen, and A. I. Verkamo (1995). Discovering frequent episodes in sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, 1995*, pp. 210–215. AAAI Press.

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.

Masseglia, F. (2002). *Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel*. Ph. D. thesis, Université de Versailles Saint-Quentin en Yvelines.

References III

Ritschard, G., R. Bürgin, and M. Studer (2013). Exploratory mining of life event histories. In J. J. McArdle and G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Quantitative Methodology, pp. 221–253. New York: Routledge.

Ritschard, G., A. Gabadinho, N. S. Müller, and M. Studer (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

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.

Widmer, E. and G. Ritschard (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life Course Research 14*(1-2), 28–39.

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