Exploring the sequencing and timing of life events

Gilbert Ritschard

Reto Bürgin and Matthias Studer

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

Lausanne Conference on Sequential Analysis
University of Lausanne, June 6-8, 2012

Objectives

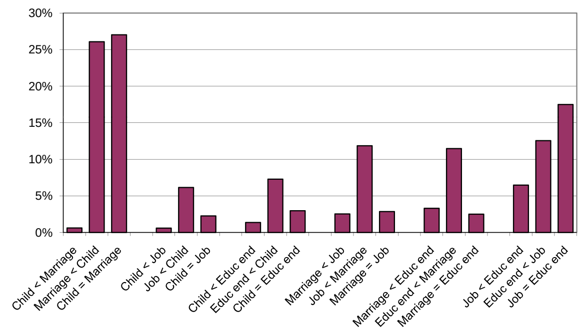
- (Non tree) data-mining-based methods
  - Discovering interesting information from sequences of life events, i.e. on how people sequence important life events
    - What is the most typical succession of family or professional life events?
    - Are there standard ways of sequencing those events?
    - 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

Objectives (continued)

- Demonstrate the kind of results that can be obtained by mining event subsequences
- Search for
  - most frequent subsequences
  - subsequences that best discriminate groups (provided covariate)
- But also, computing dissimilarities between event sequences
- which permits then
  - clustering event sequences
  - principal coordinate analysis (multi-dimensional scaling)
  - find out medoids or density-based representative sequences
  - discrepancy analysis and regression trees ...

What's new

- Previous attempts with event sequences in social sciences (e.g. Billari et al., 2006; Ritschard et al., 2007) mainly consisted in counting predefined subsequences.



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

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

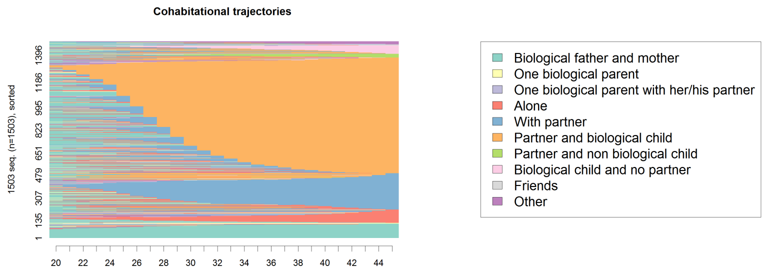
- Event sequence: events occur at a given (time) position
  - Interest in their order, in their sequencing
  - Can be time stamped (TSE)

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

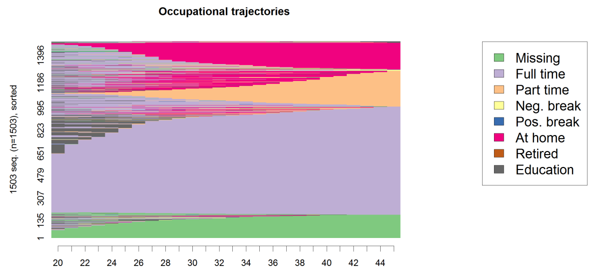
The Biographical SHP Data

- 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
- 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"

Creating the event sequences

- We create the cohabitational event sequence object as follows using the previous matrix (denoted `transition.coh.mat`)

```
R> shpevt.coh <- seqecreate(seqs.coh, tevent=transition.coh.mat)
```

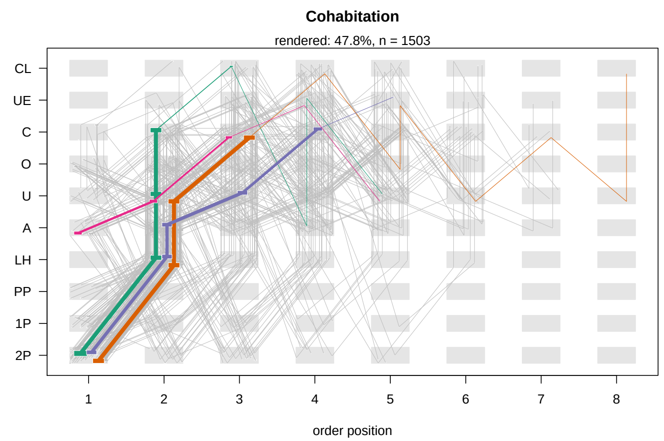
- For occupational trajectories, we define an event for the start of each spell in a different state

```
R> shpevt.occ <- seqecreate(seqs.occ, tevent="state")
```

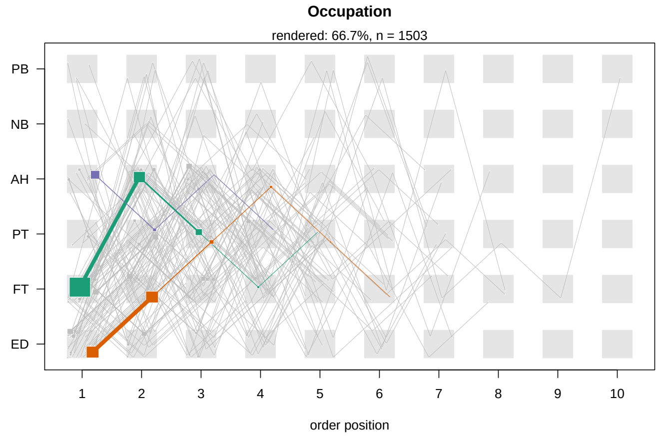
  after having merged the 'At home' `AH` and 'Retired' `R` states.

Rendering cohabitational event sequences

(Bürgin et al., 2012)



Rendering occupational event sequences
(Bürgin et al., 2012)



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).

Events and transitions

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

Example

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

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

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".

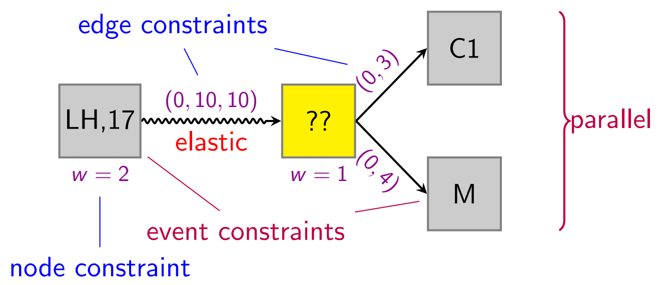
Frequent and discriminant subsequences

- 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.

Episode structure constraints

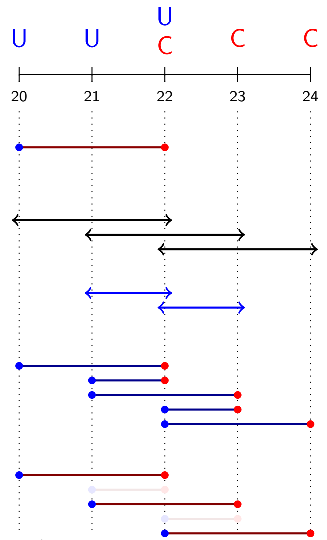
Joshi et al. (2001)

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

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 cohabitational subsequences - 2

10 most frequent subsequences, min support 50

- With at least 2 events and 3-year maximum time span

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

	Subsequence	Support	Count	#Transitions	#Events
1	(LH,U)	0.454	682	1	2
2	(C,U)	0.291	437	1	2
3	(2P) → (LH)	0.275	414	2	2
4	(U) → (C)	0.274	412	2	2
5	(A,LH)	0.244	367	1	2
6	(C,LH)	0.180	270	1	2
7	(C,LH,U)	0.175	263	1	3
8	(LH) → (C)	0.166	250	2	2
9	(A) → (U)	0.158	237	2	2
10	(2P) → (A)	0.148	223	2	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 occupational subsequences - 2

Most frequent subsequences, min support = 50

- With at least 2 events and 3-year maximum time span

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

	Subsequence	Support	Count	#Transitions	#Events
1	(ED) → (FT)	0.185	288	2	2
2	(FT) → (AH)	0.067	100	2	2
3	(ED) → (AH)	0.042	73	2	2
4	(PT) → (FT)	0.036	56	2	2
5	(PT) → (AH)	0.034	53	2	2
6	(ED) → (PT)	0.031	52	2	2

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

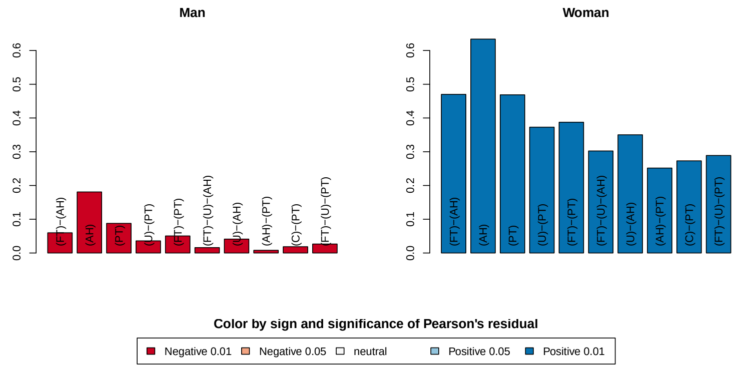
Cohabitational subsequences that best discriminate sex

Remember that we observe only since age 20!

	Subsequence	Chi-2	Support	Freq. Men	Freq. Women	Diff
1	(LH)	38.3	0.72	0.795	0.651	0.144
2	(2P) → (U)	22.4	0.58	0.642	0.521	0.122
3	(LH) → (U)	19.0	0.27	0.316	0.216	0.101
4	(LH) → (C)	18.3	0.38	0.436	0.328	0.109
5	(2P) → (LH)	18.3	0.62	0.676	0.567	0.108
6	(2P) → (A) → (U)	17.5	0.21	0.253	0.164	0.089

Cohabitational subsequences that discriminate sex

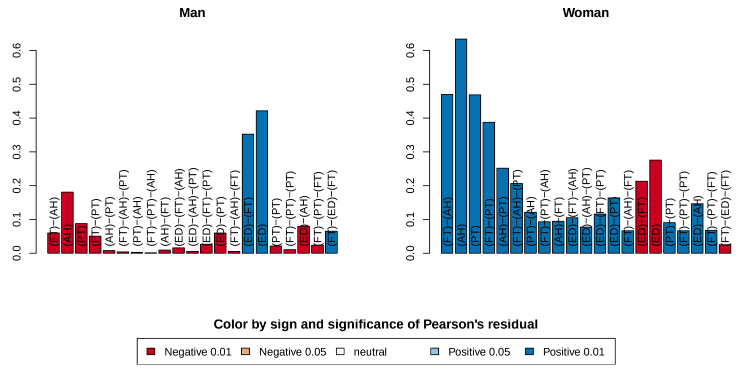
at the 1% level



Occupational 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	(FT) → (PT)	247.5	0.22	0.051	0.387	-0.337
5	(AH) → (PT)	195.5	0.13	0.008	0.252	-0.244
6	(FT) → (AH) → (PT)	161.5	0.11	0.004	0.206	-0.202

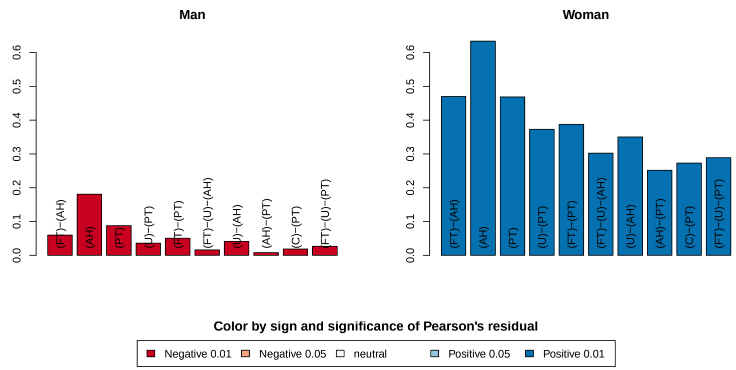
Occupational subsequences that discriminate sex
at the 0.1% level



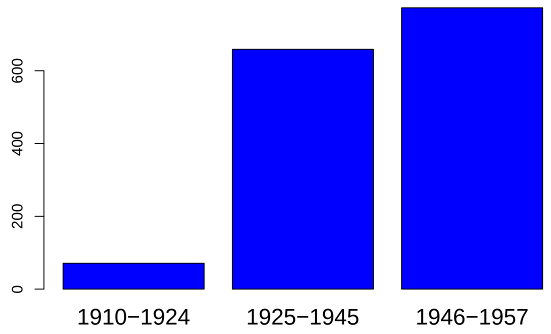
Mixed events: Subsequences that best discriminate sex

	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

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



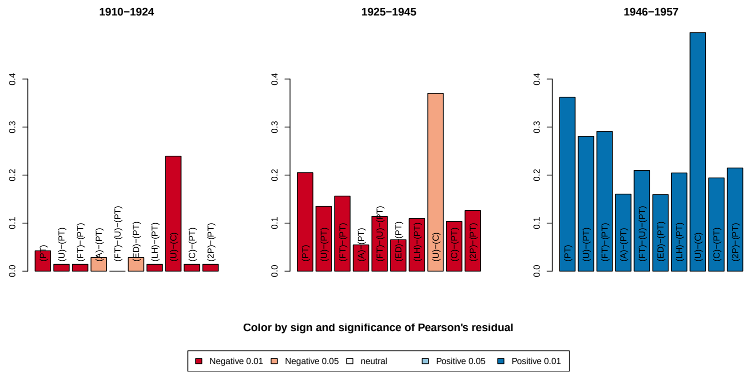
Birth cohort distribution



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

Mixed events: Subsequences that best discriminate birth cohorts



Pairwise dissimilarities

- Optimal matching distance for event sequences (Studer et al., 2010; Moen, 2000)
  - the insertion/deletion of an event;
  - a change in the time stamp of a given event;
- Costs: indel = 1 and unit time displacement = 0.1
- Normalized distance

$$d_{N,ome}(x,y) = \frac{2d_{ome}(x,y)}{\Omega(x)+\Omega(y)+d_{ome}(x,y)}$$

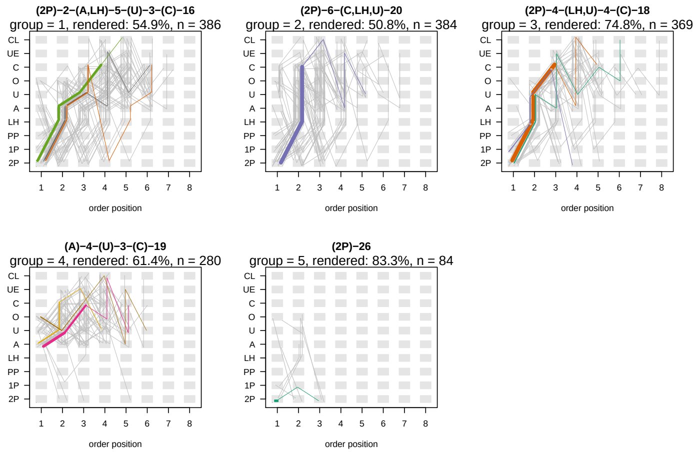
where $d_{ome}(x,y)$ is the OME dissimilarity between the time-stamped event sequences x and y, and $\Omega(x)$ the total cost for inserting all the events of x.

Four cohabitational types (PAM solution)

	Man	Woman	Overall
(2P) $\xrightarrow{2}$ (A,LH) $\xrightarrow{5}$ (U) $\xrightarrow{3}$ (C) $\xrightarrow{16}$	0.298	0.216	0.257
(2P) $\xrightarrow{6}$ (C,LH,U) $\xrightarrow{20}$	0.266	0.245	0.255
(2P) $\xrightarrow{4}$ (LH,U) $\xrightarrow{4}$ (C) $\xrightarrow{18}$	0.249	0.242	0.246
(A) $\xrightarrow{4}$ (U) $\xrightarrow{3}$ (C) $\xrightarrow{19}$	0.138	0.234	0.186
(2P) $\xrightarrow{26}$	0.049	0.063	0.056

	1910-1924	1925-1945	1946-1957	Overall
(2P) $\xrightarrow{2}$ (A,LH) $\xrightarrow{5}$ (U) $\xrightarrow{3}$ (C) $\xrightarrow{16}$	0.183	0.235	0.282	0.257
(2P) $\xrightarrow{6}$ (C,LH,U) $\xrightarrow{20}$	0.380	0.310	0.198	0.255
(2P) $\xrightarrow{4}$ (LH,U) $\xrightarrow{4}$ (C) $\xrightarrow{18}$	0.211	0.211	0.278	0.246
(A) $\xrightarrow{4}$ (U) $\xrightarrow{3}$ (C) $\xrightarrow{19}$	0.113	0.164	0.212	0.186
(2P) $\xrightarrow{26}$	0.113	0.080	0.030	0.056

Cluster of cohabitational trajectories

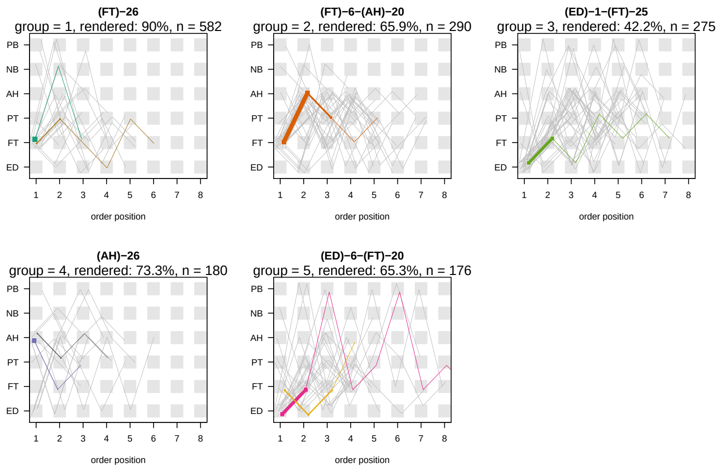


Occupational trajectory types (PAM solution)

	Man	Woman	Overall
(FT) $\xrightarrow{26}$	0.488	0.286	0.387
(FT) $\xrightarrow{6}$ (AH) $\xrightarrow{20}$	0.041	0.345	0.193
(ED) $\xrightarrow{1}$ (FT) $\xrightarrow{25}$	0.185	0.181	0.183
(AH) $\xrightarrow{26}$	0.100	0.140	0.120
(ED) $\xrightarrow{6}$ (FT) $\xrightarrow{20}$	0.186	0.048	0.117

	1910-1924	1925-1945	1946-1957	Overall
(FT) $\xrightarrow{26}$	0.338	0.404	0.378	0.387
(FT) $\xrightarrow{6}$ (AH) $\xrightarrow{20}$	0.141	0.209	0.184	0.193
(ED) $\xrightarrow{1}$ (FT) $\xrightarrow{25}$	0.127	0.155	0.212	0.183
(AH) $\xrightarrow{26}$	0.239	0.135	0.096	0.120
(ED) $\xrightarrow{6}$ (FT) $\xrightarrow{20}$	0.155	0.097	0.131	0.117

Clusters of occupational trajectories



Conclusion

- Three approaches for event sequences
  - frequent episodes
  - discriminant episodes
  - cluster analysis
- Complementary insights
  - most common characteristics
  - salient distinctions between groups
  - identify types of trajectories
- Easy to extend to other types of analyses (representative sequences, discrepancy analyses, ...)

Conclusion 2

- Work continues ...
- There are often too many frequent subsequences!
- How can we structure those subsequences?
  - Eliminate redundant subsequences, i.e., when you experience one subsequence you also experiment all its subsequences.
    - Count only maximal frequent subsequences
    - For (FT) → (AH) → (PT) we would not count the occurrence of (FT) → (AH), (FT) → (PT) or (AH) → (PT)
  - Group together sequences shared by same individuals.
    - Clustering frequent subsequences

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

Billari, F. C., J. Fürnkranz, and A. Prskawetz (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population 22*(1), 37–65.

Bürgin, R., G. Ritschard, et E. Rousseaux (2012). Exploration graphique de données séquentielles. In *Atelier Fouille Visuelle de Données : méthologie et évaluation, EGC 2012, Bordeaux*, pp. 39–50. Association EGC.

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.

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*.

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.

References III

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.

Moen, P. (2000). *Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining*. PhD thesis, University of Helsinki.

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.

Ritschard, G., M. Studer, N. Muller, and A. Gabadinho (2007). Comparing and classifying personal life courses: From time to event methods to sequence analysis. In *2nd Symposium of COST Action C34 (Gender and Well-Being). The Transmission of Well-Being: Marriage Strategies and Inheritance Systems in Europe from 17th-20th Centuries. University of Minho, Guimaraes, Portugal, April 25-28, 2007*.

References IV

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.

Studer, M., N. S. Müller, G. Ritschard, et A. Gabadinho (2010). Classer, discriminer et visualiser des séquences d'événements. *Revue des nouvelles technologies de l'information RNTI E-19*, 37–48.

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.