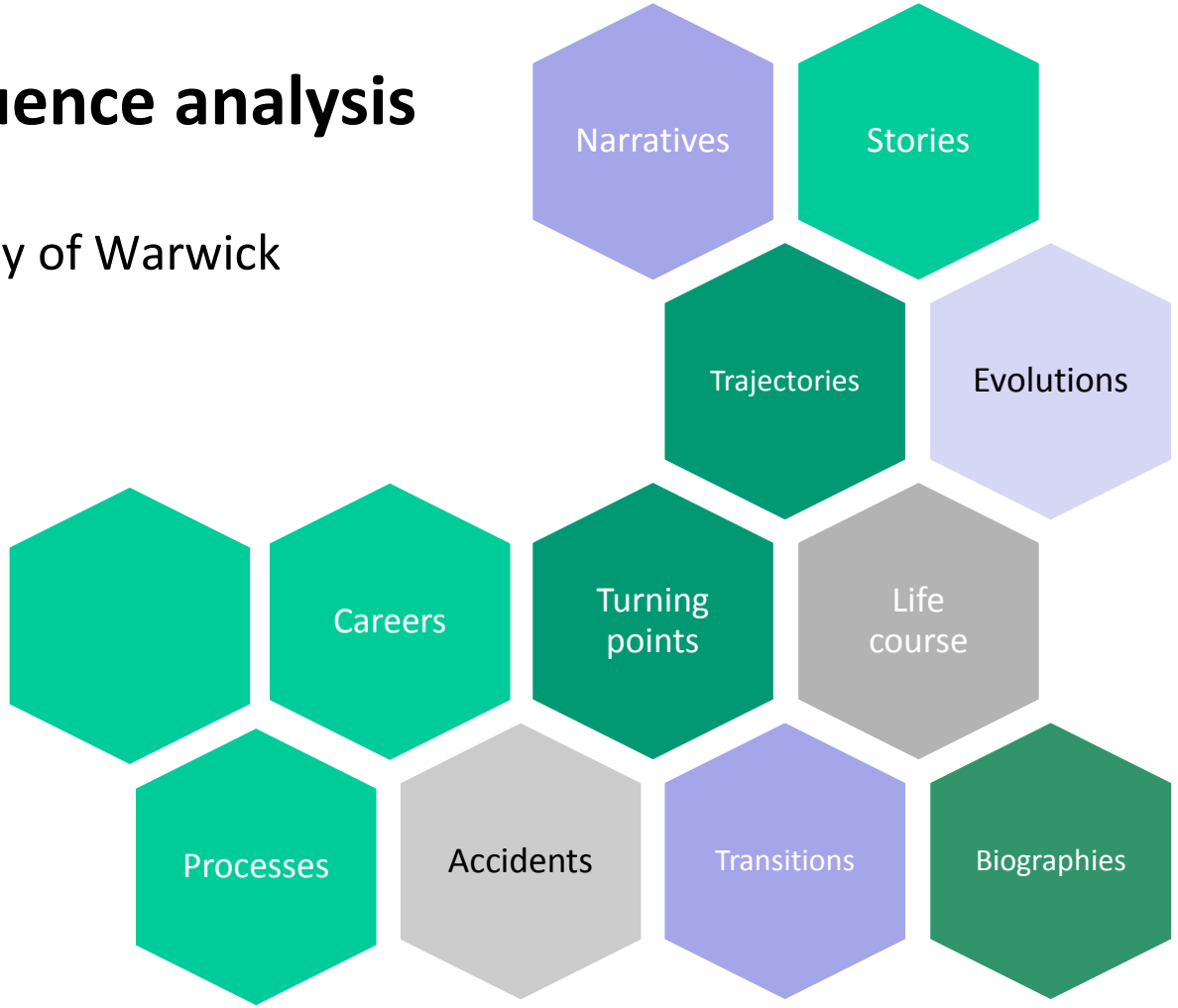
Introduction to sequence analysis

Philippe Blanchard, University of Warwick

IL027
12.02.2020



What is a sequence?

Statistical formulation

Individual	Life domain	Alphabet of states	t1	t2	t3	t4	t5	t6	
	A	a1, a2, a3	a2	a1	a2	a2	a2	a2	
i1	B	b1, b2	b1	b1	na	na	b2	b2	**One full sequence**
	C	c1, c2, c3, c4	c1	c1	c2	c3	c3	c3	
	A	a1, a2, a3	a1	a1	a1	a1	a1	a1	
i2	B	b1, b2	b1	b2	b2	b2	b2	b2	**One biography**
	C	c1, c2, c3, c4	c1	c1	c3	c3	c3	c3	

A **transition** between states c1 and c3

Cases

ID	Life domain	Alphabet	2000	2001	2002	2003	2004	2005	
	Work	Unemployed, at Work, Retired	W	U	W	W	W	W	**One trajectory**
1	Parenting	No child, One child or more	N	N	na	na	O	O	
	Coupling	Single, in Couple, Married, Divorced/separated	S	S	C	C	M	M	**One constant subsequence**
	Work	Unemployed, at Work, Retired	U	U	U	U	U	U	
2	Parenting	No child, One child or more	N	O	O	O	O	O	
	Coupling	Single, in Couple, Married, Divorced/separated	S	S	M	M	D	D	

The transition happens to be a wedding

Notations

- Alphabet A° of p_a elements a_i inside life domain A: $A^\circ = \{a_1, a_2 \dots a_{p_a}\}$

- Sequence S of length s inside life domain A:

  $S = (a_{i_1}, a_{i_2} \dots a_{i_s})$ *with* $\{a_{i_1}, a_{i_2} \dots a_{i_s}\} \subset A^\circ$

- Subsequence S' inside sequence S:

  $S' = (a_{i_m}, a_{i_{m+1}} \dots a_{i_{m+p}})$ *with* $\{m, m+p\} \subset \{1 \dots s\}$

- Non successive subsequence S'' of sequence S:

  $S'' = (a_{i_q} \dots a_{i_r})$ *with* $\{q \dots r\} \subset \{1, 2 \dots s\}$ *and* $q < \dots < r$

- Episode S''', made of one unique state a_{s_0} :

  $S''' = (a_{i_0}, a_{i_0} \dots a_{i_0})$ *with* $i_0 \in \{1, 2 \dots s\}$

- Biography B made of sequences

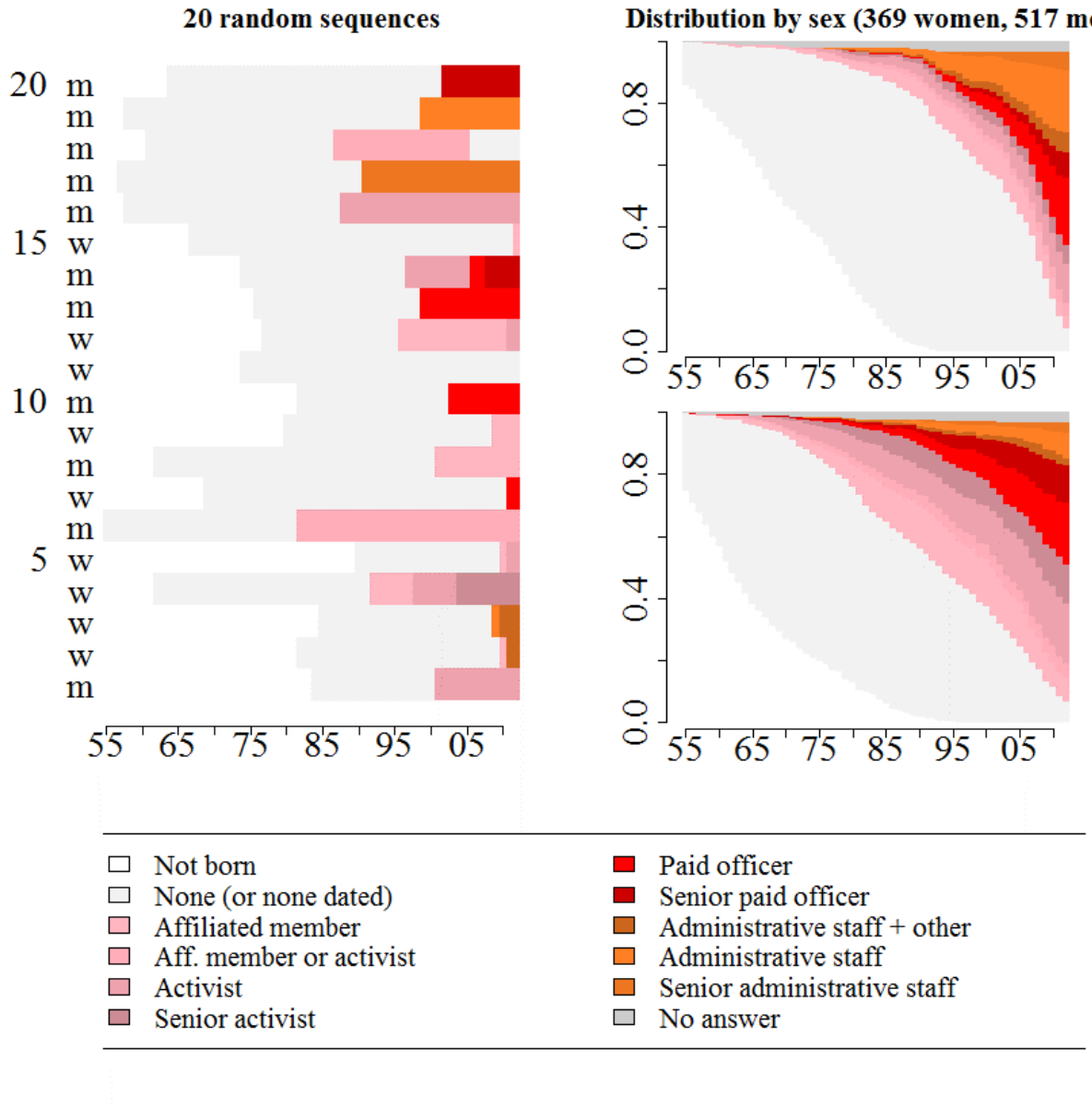
  $S_A, S_B \dots S_M$ *composed from alphabets* A°, B° … M° :

  $B = (S_A, S_B \dots S_M)$

  $B = \left[(a_{i_1}, a_{i_2} \dots a_{i_s}), (b_{j_1}, b_{j_2} \dots b_{j_s}) \dots (m_{k_1}, m_{k_2} \dots m_{k_s})\right]$

  with $\{a_{i_1}, a_{i_2} \dots a_{i_s}\} \subset A^\circ$, $\{b_{j_1}, b_{j_2} \dots b_{j_s}\} \subset B^\circ$ … $\{m_{k_1}, m_{k_2} \dots m_{k_s}\} \subset M^\circ$

What is a sequence?



Blanchard et al. 2013. "Gendered Trade Unions Careers: From Transitions to Turning Points"

What does sequence analysis do?

- Describes and represents sequences with statistics and plots
- Compares and classifies sequences
- Mines sequences
- Extracts/creates prototypical/discriminating sequences
- Links sequence profiles with external variables
- And more connections with other, non sequential methods.

Applications

- Sociology of work
  - Historical evolution of female trajectories of finance executives? (Blair-Loy 1999)
- Life course analysis
  - How do work, relationship and housing trajectories relate to quality of life? (Wiggins et a.. 2007)
- History
  - Do multi-positionning along time follow some universal *cursus honorum*? (Lemercier 2005)
- Ethnography
  - Can evolutions in English folk dance steps be derived from a unique historical pattern? (Abbot & Forrest 1986)
- Geography
  - How do patterns of travelling behaviours relate to sociodemographics? (Schlich 2003)
- Linguistics
  - Can a writing style be recognized through the analysis of typical phrases? (Barzilay & Lee 2002)

Other applications (implemented or potential)

- Work, relationship and housing trajectories related to quality of life (Wiggins et al. 2007)
- Multi-positioned careers in XIXth century economic committees (Lemercier 2005)
- Shopping decisions inside a market or along an online session (Joh et al. 2003)
- Worldly dynamic of diffusion of social legislation between countries? (Abbott & Deviney 1992)
- Engaging and disengaging processes in social movements (Fillieule and Blanchard 2013)
- Places visited in a trip (Saarlos et al. 2010)
- Historical evolution of the structure of scientific articles (Abbott and Barman 1997)
- Topics, arguments or keywords in a face-to-face or online conversation
- Stages of development of a story (novel, speech, fairy tale…)
- Mobilising events along a protest cycle
- Sounds, gestures or skills acquired along a child's development
- Etc.

Shortages of usual statistical methods

- **What methods?**
  - Basics
    - Summary statistics on individuals and groups
    - Handmade mining and counting of sequences
    - Excel sequence graphs
  - Multivariate
    - Clustering
    - Correspondence analysis
    - Synchronic regression models
  - Longitudinal
    - Time series
    - Event history
    - Regression models with time included

- **Their limits**
  - Focus on continuous variables
  - Focus on 1 or 2 events to explain
  - Ignore length of subsequences
  - Ignore order of states
  - Strict dependence on one given time axis
  - Censorship
  - No global view on sequences
  - Graphs not adpated and hard to make

What is new with sequence analysis?

1. Holistic approach
2. Strong descriptive ambition
3. Accounts for the nature of states or events (categorical data)
4. ... for the length of each episode
5. ... and for their order
6. Does not make any hypothesis about the shape of distribution of states or about the underlying cause of sequences
7. Various individuals, various Ns, various lengths.

A standard treatment sequence

1. Collecting data: questionnaire, archives, web-dataset...
2. Conceptualizing trajectories to be studied: nature of stages, time limits, time unit...
3. Building a sequential dataset: alphabet, data cleaning and formatting.
4. Exploring: summary measures, plots, mining.
5. Pair comparison of sequences: choice of a metric, calculation of NxN distance matrix ("DM")
6. Exploring/synthesizing DM: cross-tabs, clustering, ANOVA...
7. Describing and interpring the outputs (tables, clusters, tests, regressions...)
8. Exploring and mining again sequences according to the results.

6 : Êtes-vous ou avez-vous été :
(plusieurs réponses possibles)

		De quand à quand ?
- volontaire	☐	de 19 ___ à 19 ___
- permanent(e)	☐	de 19 ___ à 19 ___

7 : Quand avez-vous fini votre formation initiale de volontaire ? Mois ___ 19 ___

11 : Combien de temps par semaine estimez-vous, en moyenne, consacrer ou avoir consacré aux activités de l'association ? *(en nombre d'heures par semaine et par périodes)*

Nombre d'heures par semaine	**De quand à quand ?**
_____	de 19 ___ à 19 ___
_____	de 19 ___ à 19 ___
_____	de 19 ___ à 19 ___

12 : Parmi les phrases suivantes, quelles sont celles qui se sont appliquées à votre situation personnelle ?
(plusieurs réponses possibles)

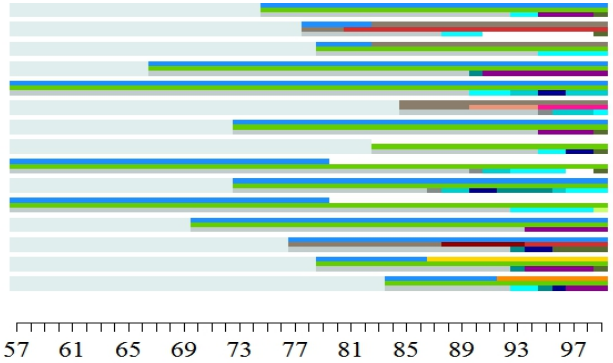
- J'étais utilisateur de l'association et je suis devenu volontaire	☐	
- Je suis / j'ai été volontaire et utilisateur de l'association		☐
- Je suis / j'ai été volontaire mais je n'arrive pas à faire appel à l'association	☐	
- Autre, précisez :		☐

16 : A quelle date avez-vous quitté l'association ? **Mois ___ 19 ___**

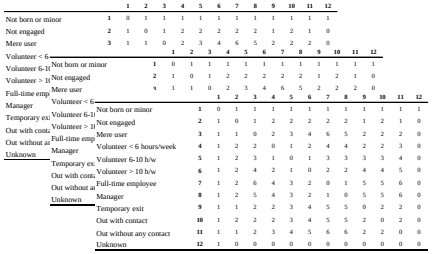
18 : Après votre engagement à AIDES, avez-vous gardé des contacts avec d'autres membres de AIDES ?

- Oui	☐
- Non	☐

Questionnaire



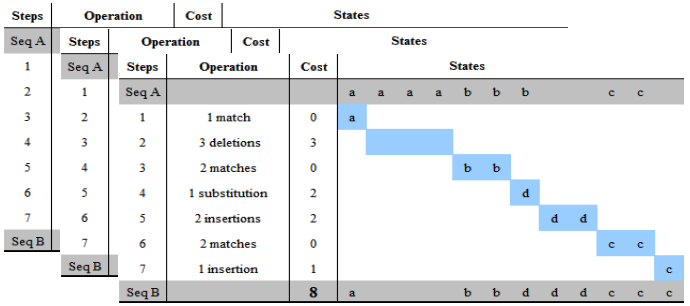
Longitudinal database



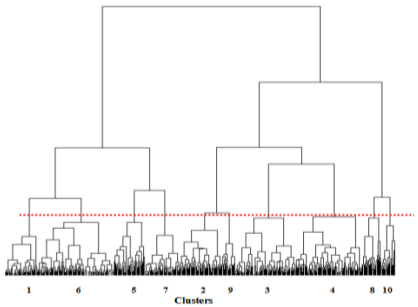
3 cost matrices

id	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
1	0	13	2	6	22	16	6	3	17	6
2	13	0	15	17	29	26	17	16	25	17
3	2	15	0	6	20	15	6	3	16	6
4	6	17	6	0	22	15	0	3	14	0
5	22	29	20	22	0	17	22	21	18	22
6	16	26	15	15	17	0	15	14	10	15
7	6	17	6	0	22	15	0	3	14	0
8	3	16	3	3	21	14	3	0	15	3
9	17	25	16	14	18	10	14	15	0	14

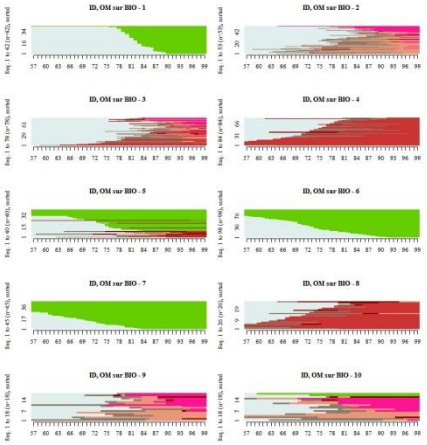
NxN distance matrix



3D optimal matching



Clustering



Typology of biographies

Description of clusters (cross-tabs, regressions)

Plots (hist. / biogr. / biol. times)

Subsequence mining

Methods for sequence comparison (1)

- **"Common chunks" of sequences** (Elzinga 1986; *CHESA* 2007)
  - Relative proportion of common states (Elzinga 1986)
  - LCP: Longest common prefix/suffix (Elzinga 1986; Gabadhino et al. 2011)
  - LCS: Longest common subsequence (Elzinga 1986; Gabadhino et al. 2011)
  - LCS with discarded duration (Dijkstra & Taris 1995)
  - Number of distinct common subsequences (Elzinga 1986)
  - Number of matching subsequences (Elzinga 1986)

- **Advantages**
  - Intuitive procedures
  - Provides intuitive measures of (dis)similarity
  - Potentially very powerful

- **Disadvantages**
  - High calculation cost
  - Not all implemented in current packages
  - Little collective empirical feedback so far.

Methods for sequence comparison (2)

- **"Optimal matching"** (Sankoff and Kruskal 1983; Abbott and Forrest 1986; Elzinga 2003)

1. Define elementary operations
   - Substitution, Insertion, Deletion, Match
2. Attribute “cost” to each elementary operation
   - Substitution costs: anti-identity, objective and transition costs, empirically trained, mixed
   - Insertion and deletion: usually fixed, close to half of average scost
   - Match: usually 0
3. Compare sequences by pairs, incrementally from first to last spell
4. Select cheapest sequence of operations (Needleman & Wunsch 1970)

Optimal matching: the straight way

Steps	Operation	Cost	States								
Seq A			a	a	a	a	b	b	b	c	c
1	1 match	0	a								
2	2 substitutions	4		b	b						
3	1 substitution	2				d					
4	2 substitutions	4					d	d			
5	1 substitution	2							c		
6	2 matches	0								c	c
Seq B		**12**	a	b	b	d	d	d	c	c	c

a *Not involved in Aides yet*
b *Involved*
c *Not involved anymore*

Optimal matching: Optimising

Steps	Operation	Cost	States
Seq A			a a a a b b b c c
1	1 match	0	a
2	3 deletions	3	
3	2 matches	0	b b
4	1 substitution	2	d
5	2 insertions	2	d d
6	2 matches	0	c c
7	1 insertion	1	c
Seq B		**8**	a b b d d d c c c

a *Not involved in Aides yet*
b *Involved*
c *Not involved anymore*

Optimal matching: Optimising

Sequences:

s=CAMBRIDGE
g=CAMPING

Distance Sequence alignment :

1) substitute s_4(B:P), s_5(R:I), s_6(I:N), s_7(D:G) delete s_8(G), s_9(E) => d=10

2) substitute s_4(B:P), delete s_5(R), substitute s_6(D:N), delete s_8 (E) => d=6

Source: Schlich (2001)