Sequential data analysis with TraMineR, Part 1

Gilbert Ritschard

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

APA-ATI Workshop on Exploratory Data Mining
University of Southern California, Los Angeles, CA, July 2009

Outline

1. Introduction
2. Concepts and definitions
3. Rendering and summarizing state sequences

Outline

1. Introduction
2. Concepts and definitions
3. Rendering and summarizing state sequences

Section outline

1. Introduction
   - Objectives
   - Overview of what you will learn

Objectives

- Concepts and questioning about sequential categorical data
- Types of sequences: with or without time content, states, transitions, events.
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

Objectives

- Concepts and questioning about sequential categorical data
- Types of sequences: with or without time content, states, transitions, events.
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

Objectives

- Concepts and questioning about sequential categorical data
- Types of sequences: with or without time content, states, transitions, events.
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

Objectives

- Concepts and questioning about sequential categorical data
- Types of sequences: with or without time content, states, transitions, events.
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

The research project

Course mainly based on results of NSF project

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

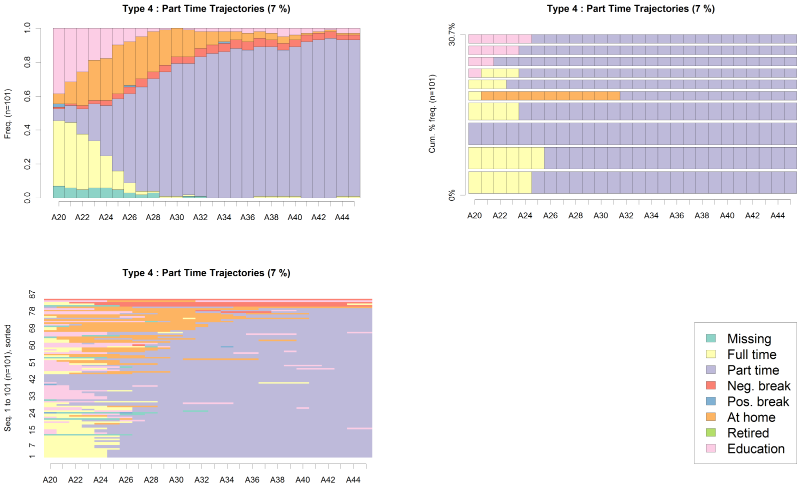
- Project FN 100012-113998 and FN-100015-122230
- Start: February 1, 2007 End: January 31, 2011

- Gilbert Ritschard, main applicant
- Eric Widmer, professor of Sociology, co-applicant
- Alexis Gabadinho, Demography
- Nicolas S. Müller, Sociology, Computer science
- Matthias Studer, Economics, Sociology

Section outline

1. Introduction
    - Objectives
    - Overview of what you will learn

Rendering sequences



Characterizing set of sequences

- Sequence of transversal measures (modal state, between entropy, ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Summary of longitudinal measures (within entropy, transition rates, mean duration ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Other global characteristics: Centro-type sequence, diversity of sequences, ...

Characterizing set of sequences

- Sequence of transversal measures (modal state, between entropy, ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Summary of longitudinal measures (within entropy, transition rates, mean duration ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Other global characteristics: Centro-type sequence, diversity of sequences, ...

Characterizing set of sequences

- Sequence of transversal measures (modal state, between entropy, ...)

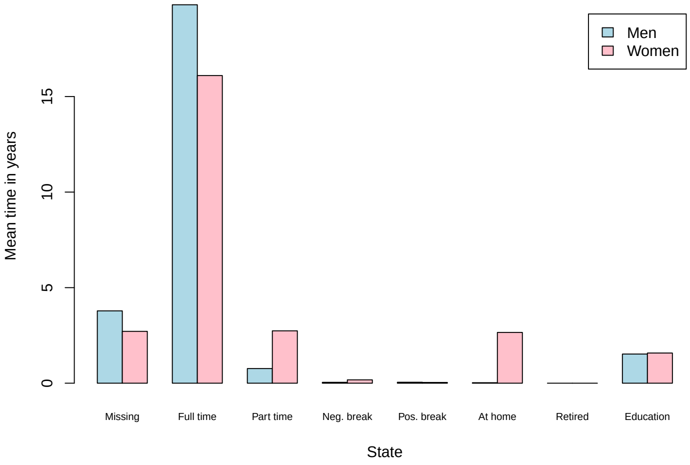
id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Summary of longitudinal measures (within entropy, transition rates, mean duration ...)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Other global characteristics: Centro-type sequence, diversity of sequences, ...

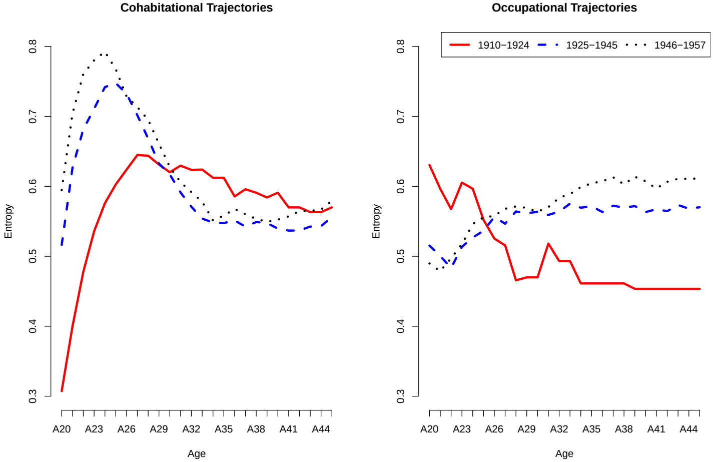
Mean time in each state



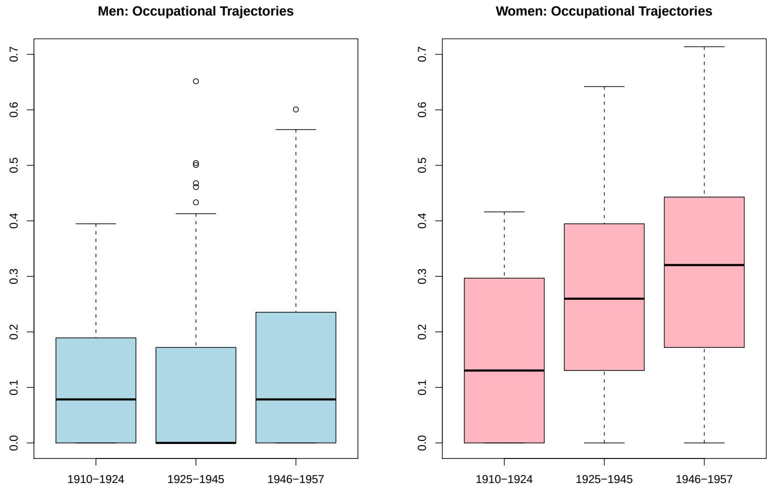
Transition rates

	[-> 0]	[-> 1]	[-> 2]	[-> 3]	[-> 4]	[-> 5]	[-> 6]	[-> 7]
Missing	0.969	0.005	0.004	0.001	0.001	0.011	0.000	0.008
Full time	0.003	0.971	0.009	0.001	0.001	0.013	0.000	0.003
Part time	0.005	0.026	0.939	0.001	0.001	0.018	0.000	0.010
Neg. break	0.040	0.047	0.027	0.880	0.000	0.007	0.000	0.000
Pos. break	0.105	0.316	0.105	0.000	0.404	0.018	0.000	0.053
At home	0.003	0.007	0.032	0.000	0.000	0.956	0.000	0.002
Retired	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
Education	0.044	0.236	0.045	0.001	0.002	0.006	0.000	0.664

Heterogeneity: Sequence of transversal entropies



Longitudinal entropy



Dissimilarities between pairs of sequences

- Distance between sequences
    - Different metrics metrics (LCP, LCS, OM)
- Once we have 2 by 2 dissimilarities, we can
    - Determine a central sequence (centro-type)
    - Measure the discrepancy between sequences
    - Clustering a set of sequences
    - MDS scatterplot representation of sequences
    - Heterogeneity analysis of a set of sequences (ANOH)
    - Dissimilarity analysis (Induction trees)

Dissimilarities between pairs of sequences

- Distance between sequences
  - Different metrics metrics (LCP, LCS, OM)
- Once we have 2 by 2 dissimilarities, we can
  - Determine a central sequence (centro-type)
  - Measure the discrepancy between sequences
  - Clustering a set of sequences
  - MDS scatterplot representation of sequences
  - Heterogeneity analysis of a set of sequences (ANOH)
  - Dissimilarity analysis (Induction trees)

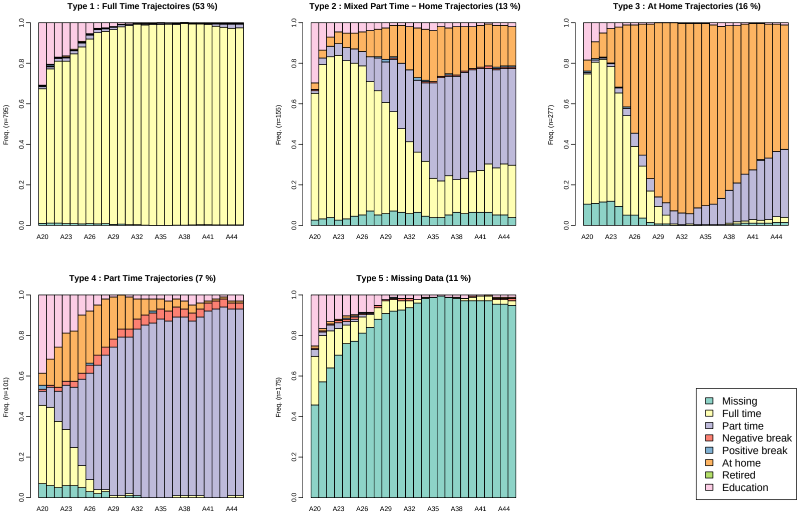
Dissimilarities between pairs of sequences

- Distance between sequences
  - Different metrics metrics (LCP, LCS, OM)
- Once we have 2 by 2 dissimilarities, we can
  - Determine a central sequence (centro-type)
  - Measure the discrepancy between sequences
  - Clustering a set of sequences
  - MDS scatterplot representation of sequences
  - Heterogeneity analysis of a set of sequences (ANOH)
  - Dissimilarity analysis (Induction trees)

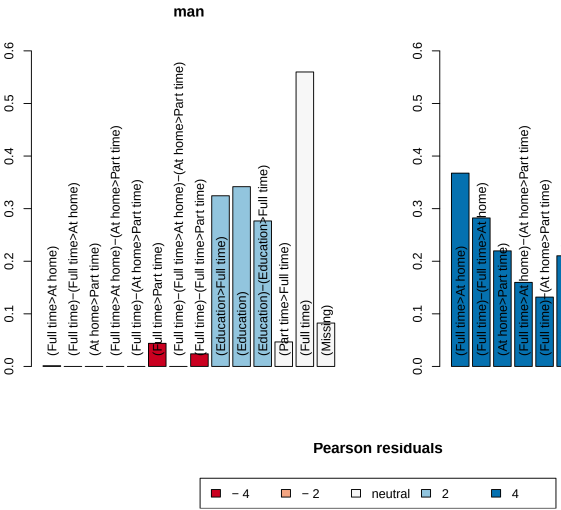
Dissimilarities between pairs of sequences

- Distance between sequences
  - Different metrics metrics (LCP, LCS, OM)
- Once we have 2 by 2 dissimilarities, we can
  - Determine a central sequence (centro-type)
  - Measure the discrepancy between sequences
  - Clustering a set of sequences
  - MDS scatterplot representation of sequences
  - Heterogeneity analysis of a set of sequences (ANOH)
  - Dissimilarity analysis (Induction trees)

Cluster analysis: determining typologies



Event sequences: discriminating sub-sequences



What you will not find in this course ...

- Transition analysis by means of Markovian and other statistical models.
- for Markovian models, see for instance Berchtold and Raftery (2002)

- Survival analysis
- e.g. Hosmer and Lemeshow (1999), Hothorn et al. (2006)

- Determination of association rules between sub-sequences
- Rarely considered in the literature! (NSM is woking hard on it!!)

What you will not find in this course ...

- Transition analysis by means of Markovian and other statistical models.
- for Markovian models, see for instance Berchtold and Raftery (2002)

- Survival analysis
- e.g. Hosmer and Lemeshow (1999), Hothorn et al. (2006)

- Determination of association rules between sub-sequences
- Rarely considered in the literature! (NSM is woking hard on it!!)

What you will not find in this course ...

- Transition analysis by means of Markovian and other statistical models.
- for Markovian models, see for instance Berchtold and Raftery (2002)

- Survival analysis
- e.g. Hosmer and Lemeshow (1999), Hothorn et al. (2006)

- Determination of association rules between sub-sequences
- Rarely considered in the literature! (NSM is woking hard on it!!)

Outline

1 Introduction

2 Concepts and definitions

3 Rendering and summarizing state sequences

Section outline

2 Concepts and definitions

- Definitions and types of sequences
- Some examples
- Alternative sequence data organizations

Sequence

Definition:

- Alphabet A: finite set
- Sequence of length k: ordered list of k successively chosen elements of A

- Examples:
  - Text: $A =$ set of letters, but can also be set of words, of n-grams, ...
  - Biology: $A =$ set of nucleotides, of proteins, ...
  - On-off signals: $A = \{0, 1\}$
  - Buying behaviors: $A =$ set of items.
  - Life course: $A =$ set of considered cohabitation states, types of occupation, ...

Sequences: notations

- Sequence x of length k
  - $x = (x_1, x_2, \ldots, x_k)$
  - If no ambiguity: $x = x_1 x_2 \cdots x_k$
  - separator necessary when A includes a composite symbol (ex: S single, M married, MC married with child S-S-M-M-MS-MS-MS)

Types of sequences

Nature of sequences

Depends on

- Information conveyed by position j in the sequence
    - Temporal dimension?
- Nature of the elements of the alphabet
    - objects or changes
    - states, transitions or events

	Temporal dimension	
Alphabet	No	Yes
Objects/States	Object sequence	State sequence
Transitions/Events	(sequence of object changes)	Event sequence

Types of sequences

Nature of sequences

Depends on

- Information conveyed by position j in the sequence
  - Temporal dimension?
- Nature of the elements of the alphabet
  - objects or changes
  - states, transitions or events

	Temporal dimension	
Alphabet	No	Yes
Objects/States	Object sequence	State sequence
Transitions/Events	(sequence of object changes)	Event sequence

Types of sequences

Nature of sequences

Depends on

- Information conveyed by position j in the sequence
  - Temporal dimension?
- Nature of the elements of the alphabet
  - objects or changes
  - states, transitions or events

	Temporal dimension	
Alphabet	No	Yes
Objects/States	Object sequence	State sequence
Transitions/Events	(sequence of object changes)	Event sequence

Types of sequences

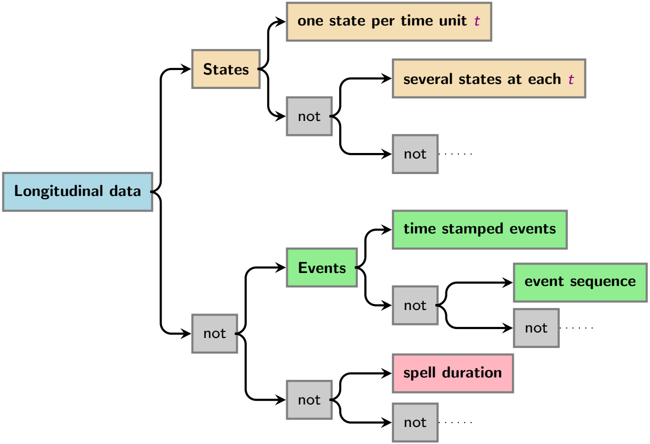
Nature of sequences

Depends on

- Information conveyed by position j in the sequence
  - Temporal dimension?
- Nature of the elements of the alphabet
  - objects or changes
  - states, transitions or events

	Temporal dimension	
Alphabet	No	Yes
Objects/States	Object sequence	State sequence
Transitions/Events	(sequence of object changes)	Event sequence

Ontology of chronological data
(Aristotelian tree)



Section outline

2 Concepts and definitions

- Definitions and types of sequences
- Some examples
- Alternative sequence data organizations

Alternative views of chronological sequences

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
marital status	single	single	single	single	married
education level	primary	secondary	secondary	secondary	secondary
job	no	no	first	first	first

Alternative views of chronological sequences

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
marital status	single	single	single	single	married
education level	primary	secondary	secondary	secondary	secondary
job	no	no	first	first	first

Transforming time stamped events into state sequences

Example: 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.

Transforming time stamped events into state sequences

Example: 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.

Transforming time stamped events into state sequences

Example: 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.

Transforming time stamped events into state sequences

Example: 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.

Deriving the states

Associate one state to 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	...

- Can we automatize the transformation of
  - events into states?
  - states into events?

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	...

- Can we automatize the transformation of
  - events into states?
  - states into events?

Section outline

2 Concepts and definitions

- Definitions and types of sequences
- Some examples
- Alternative sequence data organizations

State sequences

Formats supported by TraMineR

Code	Data type	Several rows for same case	Usage examples
STS	State-sequence	No	Markov modeling, OM
SPS	State-permanence	No	Markov modeling, OM
SSS*	State-start	No	Markov modeling, OM
SRS	Shifted-replicated-sequence	Yes	Mobility tree
DSS	Distinct-state-sequence	No	OM without time reference
SPELL	Spell	Yes	Survival analysis
PPER*	Person-period	Yes	Discrete survival analysis

Formats of state sequences: examples - I

Code	Example										
STS	*Id*	18	19	20	21	22	23	24	25	26	27
	101	S	S	S	M	M	MC	MC	MC	MC	D
	102	S	S	S	MC	MC	MC	MC	MC	MC	MC
SPS	*Id*	1	2	3	4						
	101	(S,3)	(M,2)	(MC,4)	(D,1)						
	102	(S,3)	(MC,7)								
SSS*	*Id*	1	2	3	4						
	101	(S,18)	(M,21)	(MC,23)	(D,27)						
	102	(S,18)	(MC,21)								
SRS	*Id*	$t-9$	$t-8$	$t-7$	$t-6$	$t-5$	$t-4$	$t-3$	$t-2$	$t-1$	t
	101	S	S	S	M	M	MC	MC	MC	MC	D
	101	.	S	S	S	M	M	MC	MC	MC	MC
	101	.	.	S	S	S	M	M	MC	MC	MC
	⋮										
	101	.	.	.	.	.	.	.	.	S	S
	102	S	S	S	MC	MC	MC	MC	MC	MC	MC
	102	.	S	S	S	MC	MC	MC	MC	MC	MC
	⋮										
DSS	*Id*	1	2	3	4						
	101	S	M	MC	D						
	102	S	MC								

Formats of state sequences: examples - II

Code	Example				
SPELL	*Id*	*Index*	*From*	*To*	*State*
	101	1	18	20	Single (S)
	101	2	21	22	Married (M)
	101	3	23	26	Married w Children (MC)
	101	4	27	27	Divorced (D)
	102	1	18	20	Single (S)
	102	2	21	27	Married w Children (MC)
PPER*	*Id*	*Index*	*Age*	*State*	
	101	1	18	Single (S)	
	101	2	19	Single (S)	
	101	3	20	Single (S)	
	101	4	21	Married (M)	
	⋮	⋮	⋮		
	101	10	27	Divorced (D)	
	102	1	18	Single (S)	
	⋮	⋮	⋮		

Event sequences

Formats supported by TraMineR

Code	Data type	Several rows for same case	Usage examples
FCE*	Fixed-column-event	No	Survival analysis
HTSE*	Horizontal time-stamped-event	No	Event sequence mining
TSE	Vertical time-stamped-event	Yes	Event sequence mining

Event sequences: examples

Code	Example								
FCE*	*Id*	*#marr.*	*1st marr.*	*2nd marr.*	· · ·	*#child.*	*1st child*	*2nd child*	· · ·
	101	1	21	.	.	2	23	26	.
	102	1	21	.	.	1	21	.	.
HTSE*	*Id*	1	2	3	· · ·				
	101	(marriage, 21)	(childbirth, 23)	(childbirth, 26)	(divorce, 27)				
	102	(marriage, 21)	(childbirth, 21)						
TSE	*Id*	*Time*	*Event*						
	101	21	Marriage						
	101	23	Childbirth						
	101	26	Childbirth						
	101	27	Divorce						
	102	21	Marriage						
	102	21	Childbirth						

Outline

1. Introduction
2. Concepts and definitions
3. Rendering and summarizing state sequences

The ‘mvad’ data set

- For illustration, we use the mvad data set (McVicar and Anyadike-Danes, 2002)
- Data about transition from school to employment in North Ireland
- 712 cases
- 72 monthly activity statuses (July 1993-June 1999)
- 14 additional variables
- The follow-up starts when respondents finished compulsory school.

mvad variables

id	unique individual identifier
weight	sample weights
male	binary dummy for gender, 1=male
catholic	binary dummy for community, 1=Catholic
Belfast	binary dummies for location of school, one of five Education and Library Board areas in Northern Ireland
N.Eastern	”
Southern	”
S.Eastern	”
Western	”
Grammar	binary dummy indicating type of secondary education, 1=grammar school
funemp	binary dummy indicating father's employment status at time of survey, 1=father unemployed
gcse5eq	binary dummy indicating qualifications gained by the end of compulsory education, 1=5+ GCSEs at grades A-C, or equivalent
fmpr	binary dummy indicating SOC code of father?s current or most recent job,1=SOC1 (professional, managerial or related)
livboth	binary dummy indicating living arrangements at time of first sweep of survey (June 1995), 1=living with both parents
jul93	Monthly Activity Variables are coded 1-6, 1=school, 2=FE, 3=employment, 4=training, 5=joblessness, 6=HE
⋮	”
jun99	”

Creating the sequence object

- Loading the data set and creating the ‘state sequence’ object

```
R> data(mvad)
R> mvad.lab <- seqstatl(mvad[, 17:86])
R> mvad.shortlab <- c("EM", "FE", "HE", "JL", "SC",
+        "TR")
R> mvad.seq <- seqdef(mvad, 17:86, states = mvad.shortlab,
+        labels = mvad.lab)
```

Section outline

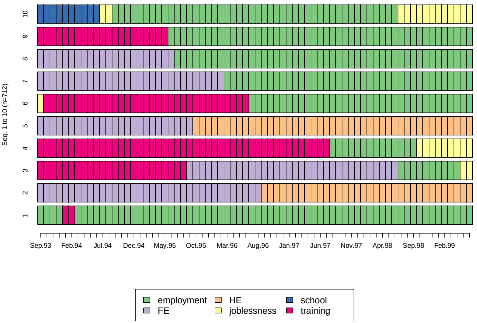
3 Rendering and summarizing state sequences

- Three basic plots
- Sequences of transversal summaries
- Other aggregated summaries
- Longitudinal characteristics of individual sequences

i-plot: Plot of individual sequences (A)

- The plot of individual sequences (i-plot) visualizes each sequence with a horizontal bar. (Scherer, 2001; Brzinsky-Fay et al., 2006)
- i-plot of 10 first sequences (mvad data)

```
R> seqiplot(mvad.seq, cex.legend = 1.3)
```



i-plot: Plot of individual sequences (B)

- The i-plot of the whole set of sequences exhibits the diversity among sequences.
- It may be useful to sort the sequences according to some factor.
- Here is how to i-plot data grouped according to grade obtained at end of compulsory school (gcse5eq) and sorted by religion

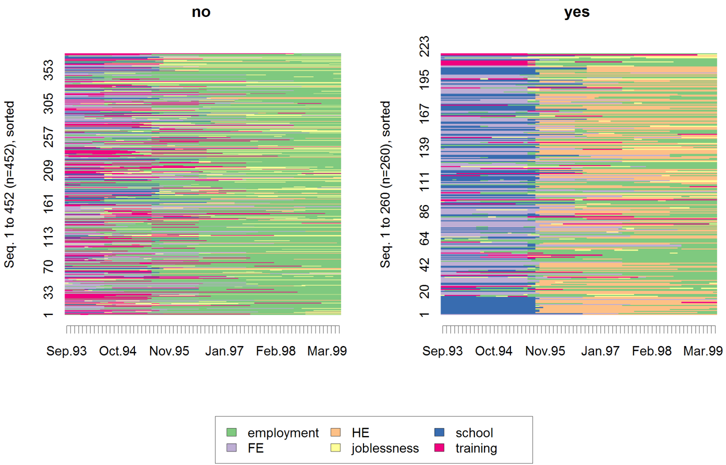
```
R> seqiplot(mvad.seq, tlim = 0, space = 0, group = mvad$gcse5eq,
+        sortv = mvad$catholic, border = NA)
```

i-plot: Plot of individual sequences (B)

- The i-plot of the whole set of sequences exhibits the diversity among sequences.
- It may be useful to sort the sequences according to some factor.
- Here is how to i-plot data grouped according to grade obtained at end of compulsory school (gcse5eq) and sorted by religion

```
R> seqiplot(mvad.seq, tlim = 0, space = 0, group = mvad$gcse5eq,
+        sortv = mvad$catholic, border = NA)
```

i-plots by CS-grade and sorted by religion



Sequence frequencies

- What are the most frequent sequences?
- seqtab() computes the frequencies and displays sequences in decreasing frequency order (here the 10 most frequent)

```
R> seqtab(mvad.seq, tlim = 10)

                    Freq Percent
(EM,70)               50    7.02
(TR,22)-(EM,48)       18    2.53
(FE,22)-(EM,48)       17    2.39
(SC,24)-(HE,46)       16    2.25
(SC,25)-(HE,45)       13    1.83
(FE,25)-(HE,45)        8    1.12
(FE,34)-(EM,36)        7    0.98
(FE,46)-(EM,24)        7    0.98
(FE,10)-(EM,60)        6    0.84
(FE,24)-(HE,46)        6    0.84
```

Sequence frequencies

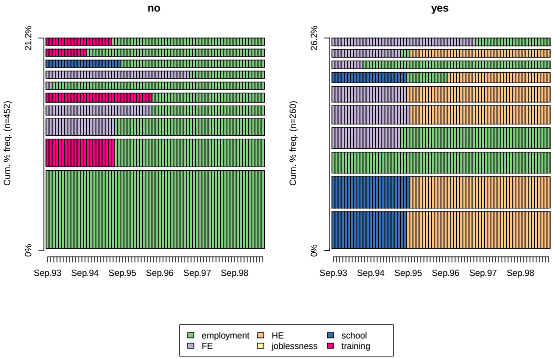
- What are the most frequent sequences?
- seqtab() computes the frequencies and displays sequences in decreasing frequency order (here the 10 most frequent)

```
R> seqtab(mvad.seq, tlim = 10)

                    Freq Percent
(EM,70)               50    7.02
(TR,22)-(EM,48)       18    2.53
(FE,22)-(EM,48)       17    2.39
(SC,24)-(HE,46)       16    2.25
(SC,25)-(HE,45)       13    1.83
(FE,25)-(HE,45)        8    1.12
(FE,34)-(EM,36)        7    0.98
(FE,46)-(EM,24)        7    0.98
(FE,10)-(EM,60)        6    0.84
(FE,24)-(HE,46)        6    0.84
```

f-plot: most frequent sequences

- seqfplot() visualizes the most frequent sequences (here according to gcse5eq).

```
R> seqfplot(mvad.seq, group = mvad$gcse5eq, pbarw = TRUE)
```



Sequence of transversal state distributions

- Distributions at each (age, calendar, ...) position.
- seqstatd() computes the distribution for each position (here just for the 8 first positions).

```
R> seqstatd(mvad.seq[, 1:8])
```

```
         Sep.93  Oct.93  Nov.93  Dec.93  Jan.94  Feb.94  Mar.94  Apr.94
EM        0.117   0.124   0.133   0.138   0.140   0.140   0.149   0.157
FE        0.386   0.388   0.382   0.381   0.369   0.364   0.361   0.353
HE        0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
JL        0.024   0.021   0.020   0.021   0.028   0.038   0.034   0.035
SC        0.251   0.246   0.244   0.242   0.240   0.242   0.240   0.240
TR        0.222   0.222   0.221   0.219   0.222   0.216   0.216   0.215
N       712.000 712.000 712.000 712.000 712.000 712.000 712.000 712.000
Entropy   0.775   0.774   0.777   0.780   0.793   0.805   0.803   0.809
```

Sequence of transversal state distributions

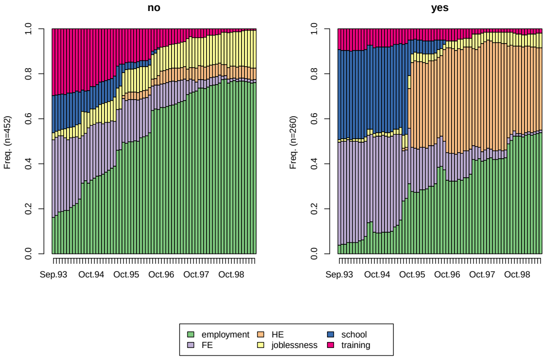
- Distributions at each (age, calendar, ...) position.
- seqstatd() computes the distribution for each position (here just for the 8 first positions).

```
R> seqstatd(mvad.seq[, 1:8])

         Sep.93  Oct.93  Nov.93  Dec.93  Jan.94  Feb.94  Mar.94  Apr.94
EM        0.117   0.124   0.133   0.138   0.140   0.140   0.149   0.157
FE        0.386   0.388   0.382   0.381   0.369   0.364   0.361   0.353
HE        0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
JL        0.024   0.021   0.020   0.021   0.028   0.038   0.034   0.035
SC        0.251   0.246   0.244   0.242   0.240   0.242   0.240   0.240
TR        0.222   0.222   0.221   0.219   0.222   0.216   0.216   0.215
N       712.000 712.000 712.000 712.000 712.000 712.000 712.000 712.000
Entropy   0.775   0.774   0.777   0.780   0.793   0.805   0.803   0.809
```

d-plot: Sequences of transversal distributions

- `seqdplot()` renders the sequence of transversal distributions (here according to `gcse5eq`).

```
R> seqdplot(mvad.seq, group = mvad$gcse5eq)
```



Section outline

3 Rendering and summarizing state sequences
- Three basic plots
- Sequences of transversal summaries
- Other aggregated summaries
- Longitudinal characteristics of individual sequences

Transversal Entropies

- Entropy of each transversal distribution $(p_1, \dots, p_a)$, with $a = |A|$ size of alphabet
- Shannon's Entropy

$$h(p_1, \dots, p_a) = -\sum_{i=1}^{a} p_i \log_2(p_i)$$

- h is 0 when all cases are in same state (good prediction of state at that position)
- h is maximal when states are equi-frequent (worth case for predicting state at that position)

Transversal Entropies

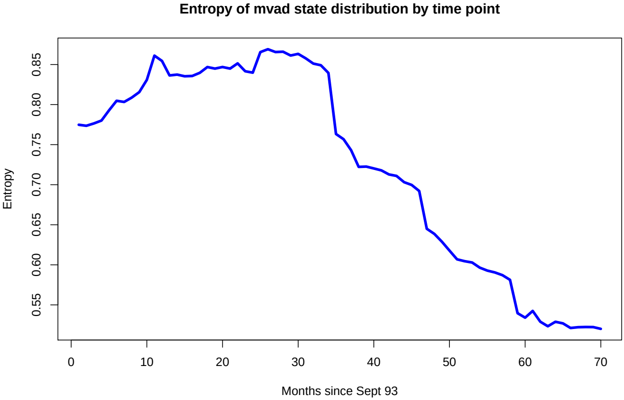
- Entropy of each transversal distribution $(p_1, \ldots, p_a)$, with $a = |A|$ size of alphabet
- Shannon's Entropy

$$h(p_1, \ldots, p_a) = -\sum_{i=1}^{a} p_i \log_2(p_i)$$

- h is 0 when all cases are in same state (good prediction of state at that position)
- h is maximal when states are equi-frequent (worth case for predicting state at that position)

Plotting the series of entropies

```
R> sd <- seqstatd(mvad.seq)
R> plot(sd$Entropy, main = "Entropy of mvad state distribution by time point",
+        xlab = "Months since Sept 93", ylab = "Entropy", type = "l",
+        lwd = 3.5, col = "blue")
```



Section outline

3. Rendering and summarizing state sequences
   - Three basic plots
   - Sequences of transversal summaries
   - Other aggregated summaries
   - Longitudinal characteristics of individual sequences

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

Time spent in each state (A)

- Time spent in each state by individual sequence

```
R> mvad.statd <- seqistatd(mvad.seq)
R> mvad.statd[1:5, ]
```

```
  EM FE HE JL SC TR
1 68  0  0  0  0  2
2  0 36 34  0  0  0
3 10 34  0  2  0 24
4 14  0  0  9  0 47
5  0 25 45  0  0  0
```

- Computing the mean time by column

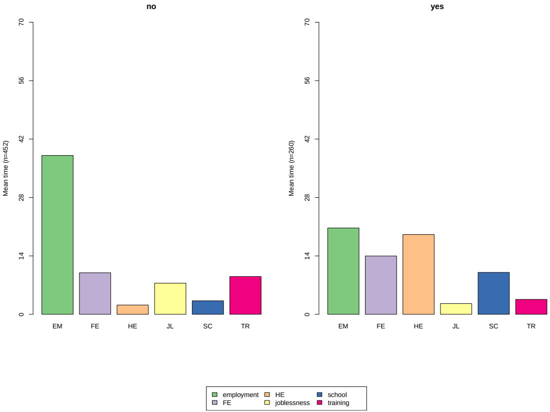
```
R> mt <- apply(mvad.statd, 2, mean)
R> mt
```

```
       EM        FE        HE        JL        SC        TR
31.721910 11.426966  8.398876  5.674157  5.723315  7.054775
```

Plot of mean times

- Plot of mean time by gcse5eq

```
R> seqmtplot(mvad.seq, group = mvad$gcse5eq)
```



Section outline

3 Rendering and summarizing state sequences

- Three basic plots
- Sequences of transversal summaries
- Other aggregated summaries
- Longitudinal characteristics of individual sequences

Transition rates

- Transition rate: estimation of probability to be in state i at t when we are in state j at previous position $t-1$

$$p(x_{it} \mid x_{j(t-1)})$$

```
R> round(seqtrate(mvad.seq), digits = 4)

        [-> EM] [-> FE] [-> HE] [-> JL] [-> SC] [-> TR]
[EM ->]  0.9864  0.0020  0.0025  0.0065  0.0004  0.0022
[FE ->]  0.0279  0.9514  0.0066  0.0090  0.0010  0.0041
[HE ->]  0.0102  0.0002  0.9872  0.0019  0.0000  0.0005
[JL ->]  0.0418  0.0084  0.0023  0.9387  0.0005  0.0084
[SC ->]  0.0142  0.0081  0.0182  0.0056  0.9509  0.0029
[TR ->]  0.0383  0.0036  0.0000  0.0136  0.0004  0.9442
```

Longitudinal entropy

- Entropy computed within each sequence
- is 0 when the sequence contains only a single state (when the person stays in same state during the observed period, for example `A-A-A-A-A-A-A-A`)
- maximum when sequence has a same number of each state in the alphabet (person spent same time in each possible state, for example `A-A-B-B-C-C-D-D`)
- By default, TraMineR normalizes the longitudinal entropy by the entropy of the alphabet

$$h_{std}(p_1, \dots, p_a) = \frac{-\sum_{i=1}^{a} p_i \log_2(p_i)}{h(A)}$$

with p_i proportion of positions in same state i.

Longitudinal entropy

- Entropy computed within each sequence
- is 0 when the sequence contains only a single state (when the person stays in same state during the observed period, for example `A-A-A-A-A-A-A-A`)
- maximum when sequence has a same number of each state in the alphabet (person spent same time in each possible state, for example `A-A-B-B-C-C-D-D`)
- By default, TraMineR normalizes the longitudinal entropy by the entropy of the alphabet

$$h_{std}(p_1, \ldots, p_a) = \frac{-\sum_{i=1}^{a} p_i \log_2(p_i)}{h(A)}$$

with p_i proportion of positions in same state i.

Computing the longitudinal entropies (B)

- `seqient()` computes the longitudinal entropies (here for *mvad* sequences)

```
R> mvad.ient <- seqient(mvad.seq)
R> mvad.ient[1:6, ]

         1          2          3          4          5          6
0.07240966 0.38662498 0.61243051 0.47611545 0.36375226 0.42259527
```

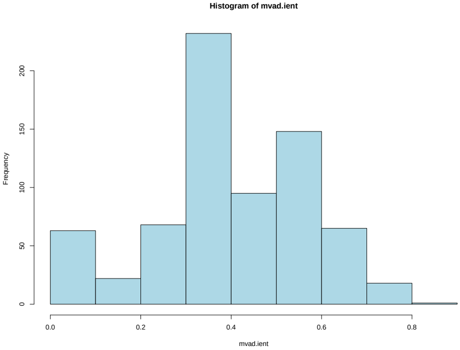
- We check that values are comprised between 0 and 1 (by default the entropy is normalized)

```
R> min(mvad.ient)

[1] 0

R> max(mvad.ient)

[1] 0.854786
```

Longitudinal entropies - Histogram

- Distribution of entropies for *mvad* data

```
R> hist(mvad.ient, col = "LightBlue")
```



Turbulence

- Entropy does not account for the state sequencing
- Turbulence: alternative measure proposed by Elzinga and Liefbroer (2007) which is sensitive to the sequencing.
- It is based on
  - the number $\phi(x)$ of subsequences of distinct states that can be extracted from the sequence of distinct consecutive states `x=S-U-M-C` (16 sub-sequences) more turbulent than `y=S-U-S-C` (15 sub-sequences)
  - the variance of time t_i spent in each distinct state i S/10-U/2-M/132 is less turbulent trajectory than S/48-U/48-M/48

Turbulence

- Entropy does not account for the state sequencing
- Turbulence: alternative measure proposed by Elzinga and Liefbroer (2007) which is sensitive to the sequencing.
- It is based on
  - the number $\phi(x)$ of subsequences of distinct states that can be extracted from the sequence of distinct consecutive states
    `x=S-U-M-C` (16 sub-sequences) more turbulent than
    `y=S-U-S-C` (15 sub-sequences)
  - the variance of time t_i spent in each distinct state i
    S/10-U/2-M/132 is less turbulent trajectory than
    S/48-U/48-M/48

Turbulence

- Entropy does not account for the state sequencing
- Turbulence: alternative measure proposed by Elzinga and Liefbroer (2007) which is sensitive to the sequencing.
- It is based on
    - the number $\phi(x)$ of subsequences of distinct states that can be extracted from the sequence of distinct consecutive states `x=S-U-M-C` (16 sub-sequences) more turbulent than `y=S-U-S-C` (15 sub-sequences)
    - the variance of time t_i spent in each distinct state i S/10-U/2-M/132 is less turbulent trajectory than S/48-U/48-M/48

Turbulence

- Entropy does not account for the state sequencing
- Turbulence: alternative measure proposed by Elzinga and Liefbroer (2007) which is sensitive to the sequencing.
- It is based on
  - the number $\phi(x)$ of subsequences of distinct states that can be extracted from the sequence of distinct consecutive states x=S-U-M-C (16 sub-sequences) more turbulent than y=S-U-S-C (15 sub-sequences)
  - the variance of time t_i spent in each distinct state i S/10-U/2-M/132 is less turbulent trajectory than S/48-U/48-M/48

Turbulence (continued)

- We need the sequence of distinct consecutive states (DSS)
- In SPS format, a state sequence is represented by the sequence of distinct states with their associated durations.

```
R> print(mvad.seq[1, ], format = "SPS")

    Sequence
[1] (EM,4)-(TR,2)-(EM,64)
```

- The DSS for the previous sequence is

```
R> seqdss(mvad.seq[1, ])

  Sequence
1 EM-TR-EM
```

- The number of sub-sequences of the above DSS is

```
R> seqsubsn(mvad.seq[1, ], DSS = TRUE)

  Subseq.
1       7
```

Turbulence (continued)

- We need the sequence of distinct consecutive states (DSS)
- In SPS format, a state sequence is represented by the sequence of distinct states with their associated durations.

```
R> print(mvad.seq[1, ], format = "SPS")

    Sequence
[1] (EM,4)-(TR,2)-(EM,64)
```

- The DSS for the previous sequence is

```
R> seqdss(mvad.seq[1, ])

  Sequence
1 EM-TR-EM
```

- The number of sub-sequences of the above DSS is

```
R> seqsubsn(mvad.seq[1, ], DSS = TRUE)

  Subseq.
1       7
```

Turbulence (continued)

- We need the sequence of distinct consecutive states (DSS)
- In SPS format, a state sequence is represented by the sequence of distinct states with their associated durations.

```
R> print(mvad.seq[1, ], format = "SPS")

    Sequence
[1] (EM,4)-(TR,2)-(EM,64)
```

- The DSS for the previous sequence is

```
R> seqdss(mvad.seq[1, ])

  Sequence
1 EM-TR-EM
```

- The number of sub-sequences of the above DSS is

```
R> seqsubsn(mvad.seq[1, ], DSS = TRUE)

  Subseq.
1       7
```

Turbulence: formula

- Formula for a sequence x

$$T(x) = \log_2 \left(\phi(x) \frac{s^2_{t,max}(x) + 1}{s^2_t(x) + 1} \right)$$

- where s^2_t is the variance of the time spent in each distinct states and $s^2_{t,max}$ is the maximal value that this variance can reach for the given sequence length.
- This maximum is

$$s^2_{t,max} = (n-1)(1-\bar{t})$$

- where $\bar{t}$ is the mean of the consecutive time spent in each distinct state:

$$\bar{t} = \frac{\text{sequence length}}{\text{number of distinct consecutive states}}$$

Turbulence: formula

- Formula for a sequence x

$$T(x) = \log_2 \left(\phi(x) \frac{s^2_{t,max}(x) + 1}{s^2_t(x) + 1} \right)$$

- where s^2_t is the variance of the time spent in each distinct states and $s^2_{t,max}$ is the maximal value that this variance can reach for the given sequence length.
- This maximum is

$$s^2_{t,max} = (n-1)(1-\overline{t})$$

- where $\overline{t}$ is the mean of the consecutive time spent in each distinct state:

$$\overline{t} = \frac{\text{sequence length}}{\text{number of distinct consecutive states}}$$

Computing the turbulence

- `seqST()` computes the turbulence of the provided sequences.
- Displaying turbulences of 6 first sequences

```
R> mvad.turb <- seqST(mvad.seq)
R> mvad.turb[1:6]

[1]  3.076599 11.176173  6.411073  4.807756  5.517962  4.987055
```

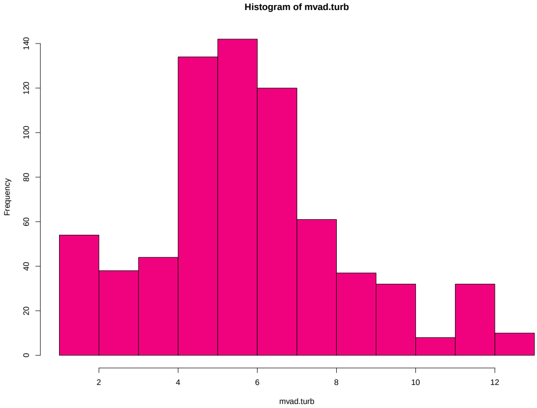
- The measure is not normalized

```
R> min(mvad.turb)

[1] 1

R> max(mvad.turb)

[1] 12.95858
```

Turbulence - Histogram

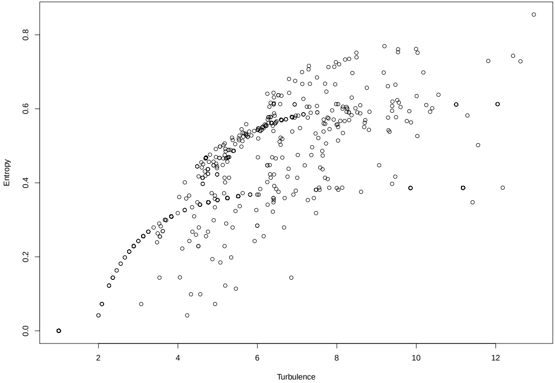
- Distribution of turbulence among the *mvad* sequences

```
R> hist(mvad.turb, col = attr(mvad.seq, "cpal")[6])
```



Comparing Turbulence and Longitudinal Entropy

```
R> plot(mvad.turb, mvad.ient, xlab = "Turbulence", ylab = "Entropy")
```



References I

Abbott, A. and A. Tsay (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research 29*(1), 3–33. (With discussion, pp 34-76).

Berchtold, A. and A. E. Raftery (2002). The mixture transition distribution model for high-order Markov chains and non-gaussian time series. *Statistical Science 17*(3), 328–356.

Billari, F. C. (2001). The analysis of early life courses: Complex description of the transition to adulthood. *Journal of Population Research 18*(2), 119–142.

Brzinsky-Fay, C., U. Kohler, and M. Luniak (2006). Sequence analysis with Stata. *The Stata Journal 6*(4), 435–460.

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.

References II

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2008). Mining sequence data in R with TraMineR: A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva. (TraMineR is on CRAN the Comprehensive R Archive Network).

Hosmer, D. W. and S. Lemeshow (1999). *Applied Survival Analysis, Regression Modeling of Time to Event Data*. New York: Wiley.

Hothorn, T., K. Hornik, and A. Zeileis (2006). party: A laboratory for recursive part(y)itioning. User's manual.

McVicar, D. and M. Anyadike-Danes (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A 165*(2), 317–334.

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.

References III

Ritschard, G., A. Gabadinho, M. Studer, and N. S. Müller (2009). Converting between various sequence representations. In Z. Ras and A. Dardzinska (Eds.), *Advances in Data Management*, Volume 223 of *Studies in Computational Intelligence*, pp. 155–175. Berlin: Springer.

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European Sociological Review 17*(2), 119–144.