Exploring Sequential Data
A Tutorial

Gilbert Ritschard

Institute for Demographic and Life Course Studies, University of Geneva
and NCCR LIVES: Overcoming vulnerability, life course perspectives
http://mephisto.unige.ch/traminer

Discovery Science, Lyon, October 29-31, 2012

Outline

1. Introduction
2. Overview of what sequence analysis can do
3. About TraMineR

Outline

1. Introduction
2. Overview of what sequence analysis can do
3. About TraMineR

Section outline

1. Introduction
   - Objectives
   - About longitudinal data analysis
   - What is sequence analysis (SA)?
     - How does SA compare with other longitudinal methods?
     - Types of categorical sequences
   - What kind of questions may SA answer to?

Objectives of the course

- Methods for extracting knowledge from sequence data
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

Objectives of the course

- Methods for extracting knowledge from sequence data
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

Objectives of the course

- Methods for extracting knowledge from sequence data
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

Objectives of the course

- Methods for extracting knowledge from sequence data
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

Objectives of the course

- Methods for extracting knowledge from sequence data
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

Section outline

1. Introduction
   - Objectives
   - About longitudinal data analysis
   - What is sequence analysis (SA)?
     - How does SA compare with other longitudinal methods?
     - Types of categorical sequences
   - What kind of questions may SA answer to?

About longitudinal data: Sequence data

Sequence data

- Multiple cases (n cases)
- For each case a sorted list of (categorical) values

- Example:

$$\begin{array}{llllllll}
1: & a & a & d & d & c & \\
2: & a & b & b & c & c & d \\
3: & b & c & c & & & \\
. & . & . & . & . & &
\end{array}$$

What is longitudinal data?

Longitudinal data

- Repeated observations on units observed over time (Beck and Katz, 1995).
- “A dataset is longitudinal if it tracks the same type of information on the same subjects at multiple points in time”. (http://www.caldercenter.org/whatis.cfm)
- “The defining feature of longitudinal data is that the multiple observations within subject can be ordered” (Singer and Willett, 2003)

Successive transversal data vs longitudinal data

- Successive transversal observations (same units)

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

- Longitudinal observations

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

Successive transversal data vs longitudinal data

- Successive transversal observations (same units)

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

- Longitudinal observations

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

Repeated independent cross sectional observations

- Successive independent transversal observations

id	t_1	t_2	t_3	···
11	B	.	.	···
12	A	.	.	···
13	B	.	.	···
.	.	.	.	···
21	.	B	.	···
22	.	B	.	···
23	.	B	.	···
.	.	.	.	···
24	.	.	D	···
25	.	.	C	···
26	.	.	A	···
.	.	.	.	···

- This is not longitudinal ...
- but ... sequences of transversal (aggregated) characteristics.

Repeated independent cross sectional observations

- Successive independent transversal observations

id	t_1	t_2	t_3	$\cdots$
11	B	.	.	$\cdots$
12	A	.	.	$\cdots$
13	B	.	.	$\cdots$
.	.	.	.	$\cdots$
21	.	B	.	$\cdots$
22	.	B	.	$\cdots$
23	.	B	.	$\cdots$
.	.	.	.	$\cdots$
24	.	.	D	$\cdots$
25	.	.	C	$\cdots$
26	.	.	A	$\cdots$
.	.	.	.	$\cdots$

- This is not longitudinal ...
- but ... sequences of transversal (aggregated) characteristics.

Longitudinal data: Where do they come from?

- Individual follow-ups: Each important event is recorded as soon as it occurs (medical card, cellular phone, weblogs, ...).
- Panels: Periodic observation of same units
- Retrospective data (biography): Depends on interviewees' memory
- Matching data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

  Examples: Wanner and Delaporte (2001), censuses and population registers, Perroux and Oris (2005), 19th Century Geneva, censuses, acts of marriage, registers of deaths, register of migrations.

- Rotating panels: partial follow up

  e.g.; Swiss Labor Force Survey, SLFS, 5 year-rotating panel (Wernli, 2010).

Longitudinal data: Where do they come from?

- Individual follow-ups: Each important event is recorded as soon as it occurs (medical card, cellular phone, weblogs, ...).
- Panels: Periodic observation of same units
- Retrospective data (biography): Depends on interviewees' memory
- Matching data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

  Examples: Wanner and Delaporte (2001), censuses and population registers, Perroux and Oris (2005), 19th Century Geneva, censuses, acts of marriage, registers of deaths, register of migrations.

- Rotating panels: partial follow up

  e.g.; Swiss Labor Force Survey, SLFS, 5 year-rotating panel (Wernli, 2010).

Longitudinal data: Where do they come from?

- Individual follow-ups: Each important event is recorded as soon as it occurs (medical card, cellular phone, weblogs, ...).
- Panels: Periodic observation of same units
- Retrospective data (biography): Depends on interviewees' memory
- Matching data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

  Examples: Wanner and Delaporte (2001), censuses and population registers, Perroux and Oris (2005), 19th Century Geneva, censuses, acts of marriage, registers of deaths, register of migrations.
- Rotating panels: partial follow up

  e.g.; Swiss Labor Force Survey, SLFS, 5 year-rotating panel (Wernli, 2010).

Longitudinal data: Where do they come from?

- Individual follow-ups: Each important event is recorded as soon as it occurs (medical card, cellular phone, weblogs, ...).
- Panels: Periodic observation of same units
- Retrospective data (biography): Depends on interviewees' memory
- Matching data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

  Examples: Wanner and Delaporte (2001), censuses and population registers, Perroux and Oris (2005), 19th Century Geneva, censuses, acts of marriage, registers of deaths, register of migrations.
- Rotating panels: partial follow up

  e.g.; Swiss Labor Force Survey, SLFS, 5 year-rotating panel (Wernli, 2010)

Longitudinal data: Where do they come from?

- Individual follow-ups: Each important event is recorded as soon as it occurs (medical card, cellular phone, weblogs, ...).
- Panels: Periodic observation of same units
- Retrospective data (biography): Depends on interviewees' memory
- Matching data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

  Examples: Wanner and Delaporte (2001), censuses and population registers, Perroux and Oris (2005), 19th Century Geneva, censuses, acts of marriage, registers of deaths, register of migrations.
- Rotating panels: partial follow up

  e.g.; Swiss Labor Force Survey, SLFS, 5 year-rotating panel (Wernli, 2010)

State sequences: an example

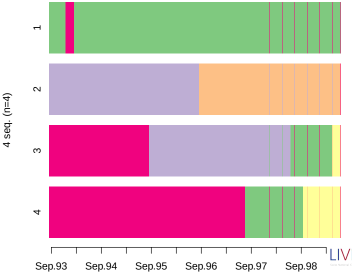
- Transition from school to work, (McVicar and Anyadike-Danes, 2002)

  Monthly states: EM = employment, TR = training, FE = further education, HE = higher education, SC = school, JL = joblessness

```
  Sequence
1 EM-EM-EM-EM-TR-TR-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-E
2 FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-F
3 TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-FE-F
4 TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-T
```

- Compact representation

```
  Sequence
1 (EM,4)-(TR,2)-(EM,64)
2 (FE,36)-(HE,34)
3 (TR,24)-(FE,34)-(EM,10)-(JL,2)
4 (TR,47)-(EM,14)-(JL,9)
```



State sequences: an example

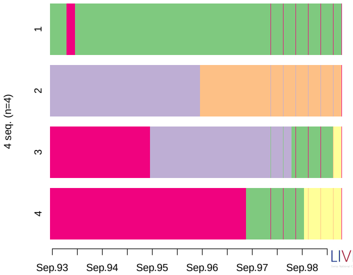
- Transition from school to work, (McVicar and Anyadike-Danes, 2002)

  Monthly states: EM = employment, TR = training, FE = further education, HE = higher education, SC = school, JL = joblessness

```
  Sequence
1 EM-EM-EM-EM-TR-TR-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-E
2 FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-F
3 TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-FE-F
4 TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-T
```

- Compact representation

```
  Sequence
1 (EM,4)-(TR,2)-(EM,64)
2 (FE,36)-(HE,34)
3 (TR,24)-(FE,34)-(EM,10)-(JL,2)
4 (TR,47)-(EM,14)-(JL,9)
```



Section outline

1. Introduction
   - Objectives
   - About longitudinal data analysis
   - What is sequence analysis (SA)?
     - How does SA compare with other longitudinal methods?
     - Types of categorical sequences
   - What kind of questions may SA answer to?

What is sequence analysis (SA)?

- Sequence analysis (SA)
  - concerned by categorical sequences,
  - holistic: interest is in the whole sequence, not just one element in the sequence (unlike survival analysis for example)
- Aim is
  - Characterizing sets of sequences
  - Identifying typical (sequence) patterns
  - Studying relationship with individual characteristics and environment

What is sequence analysis (SA)?

- Sequence analysis (SA)
  - concerned by categorical sequences,
  - holistic: interest is in the whole sequence, not just one element in the sequence (unlike survival analysis for example)
- Aim is
  - Characterizing sets of sequences
  - Identifying typical (sequence) patterns
  - Studying relationship with individual characteristics and environment

What is sequence analysis (SA)?

- Sequence analysis (SA)
  - concerned by categorical sequences,
  - holistic: interest is in the whole sequence, not just one element in the sequence (unlike survival analysis for example)
- Aim is
  - Characterizing sets of sequences
  - Identifying typical (sequence) patterns
  - Studying relationship with individual characteristics and environment

What is sequence analysis (SA)?

- Sequence analysis (SA)
  - concerned by categorical sequences,
  - holistic: interest is in the whole sequence, not just one element in the sequence (unlike survival analysis for example)
- Aim is
  - Characterizing sets of sequences
  - Identifying typical (sequence) patterns
  - Studying relationship with individual characteristics and environment

What is sequence analysis (SA)?

- Sequence analysis (SA)
  - concerned by categorical sequences,
  - holistic: interest is in the whole sequence, not just one element in the sequence (unlike survival analysis for example)
- Aim is
  - Characterizing sets of sequences
  - Identifying typical (sequence) patterns
  - Studying relationship with individual characteristics and environment

What is sequence analysis (SA)?

- Sequence analysis (SA)
  - concerned by categorical sequences,
  - holistic: interest is in the whole sequence, not just one element in the sequence (unlike survival analysis for example)
- Aim is
  - Characterizing sets of sequences
  - Identifying typical (sequence) patterns
  - Studying relationship with individual characteristics and environment

What is sequence analysis (SA)?

- Sequence analysis (SA)
  - concerned by categorical sequences,
  - holistic: interest is in the whole sequence, not just one element in the sequence (unlike survival analysis for example)
- Aim is
  - Characterizing sets of sequences
  - Identifying typical (sequence) patterns
  - Studying relationship with individual characteristics and environment

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Other Longitudinal methods

- Numerical longitudinal data: Essentially modeling approaches
  - Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
    - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
  - Growth curve models (specialized SEM) (McArdle, 2009)
  - But also, distance-based analysis (DTW, ...)
- Categorical longitudinal data
  - Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
  - Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
  - Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
  - Aligning techniques (biology) (Sharma, 2008)

Types of categorical sequences

Nature of sequences

Depends on

- Chronological order?
  - If yes, we can study timing and duration
- Information conveyed by position j in the sequence
  - If position is a time stamp, differences between positions reflect durations
- Nature of the elements of the alphabet
  - states, transitions or events, letters, proteins, ...

Types of categorical sequences

Nature of sequences

Depends on

- Chronological order?
  - If yes, we can study timing and duration.
- Information conveyed by position j in the sequence
  - If position is a time stamp, differences between positions reflect durations.
- Nature of the elements of the alphabet
  - states, transitions or events, letters, proteins, ...

Types of categorical sequences

Nature of sequences

Depends on

- Chronological order?
  - If yes, we can study timing and duration.
- Information conveyed by position j in the sequence
  - If position is a time stamp, differences between positions reflect durations.
- Nature of the elements of the alphabet
  - states, transitions or events, letters, proteins, ...

Types of categorical sequences

Nature of sequences

Depends on

- Chronological order?
  - If yes, we can study timing and duration.
- Information conveyed by position j in the sequence
  - If position is a time stamp, differences between positions reflect durations.
- Nature of the elements of the alphabet
  - states, transitions or events, letters, proteins, ...

Types of categorical sequences

Nature of sequences

Depends on

- Chronological order?
  - If yes, we can study timing and duration.
- Information conveyed by position j in the sequence
  - If position is a time stamp, differences between positions reflect durations.
- Nature of the elements of the alphabet
  - states, transitions or events, letters, proteins, ...

Types of categorical sequences

Nature of sequences

Depends on

- Chronological order?
  - If yes, we can study timing and duration.
- Information conveyed by position j in the sequence
  - If position is a time stamp, differences between positions reflect durations.
- Nature of the elements of the alphabet
  - states, transitions or events, letters, proteins, ...

Types of categorical sequences

Nature of sequences

Depends on

- Chronological order?
  - If yes, we can study timing and duration.
- Information conveyed by position j in the sequence
  - If position is a time stamp, differences between positions reflect durations.
- Nature of the elements of the alphabet
  - states, transitions or events, letters, proteins, ...

State versus event sequences

- An important distinction for chronological sequences is between
  state sequences and event sequences
  - A State, such as ‘living with a partner’ or ‘being unemployed’, lasts the whole unit of time
  - An event, such as ‘moving in with a partner’ or ‘ending education’, does not last but provokes a state change, possibly in conjunction with other events.

State versus event sequences

- An important distinction for chronological sequences is between
  state sequences and event sequences
  - A State, such as ‘living with a partner’ or ‘being unemployed’, lasts the whole unit of time
  - An event, such as ‘moving in with a partner’ or ‘ending education’, does not last but provokes a state change, possibly in conjunction with other events.

State versus event sequences

- An important distinction for chronological sequences is between
  state sequences and event sequences
  - A State, such as 'living with a partner' or 'being unemployed', lasts the whole unit of time
  - An event, such as 'moving in with a partner' or 'ending education', does not last but provokes a state change, possibly in conjunction with other events.

State versus event sequences: examples

Time stamped events

Sandra	Ending education in 1980	Start working in 1980
Jack	Ending education in 1981	Start working in 1982

- There can be simultaneous events (see Sandra)
- Elements at same position do not occur at same time

State sequence view

year	1979	1980	1981	1982	1983
Sandra	Education	Education	Employed	Employed	Employed
Jack	Education	Education	Education	Unemployed	Employed

- Only one state at each observed time
- Position conveys time information: All states at position 2 are states in 1980.

State versus event sequences: examples

Time stamped events

Sandra	Ending education in 1980	Start working in 1980
Jack	Ending education in 1981	Start working in 1982

- There can be simultaneous events (see Sandra)
- Elements at same position do not occur at same time

State sequence view

year	1979	1980	1981	1982	1983
Sandra	Education	Education	Employed	Employed	Employed
Jack	Education	Education	Education	Unemployed	Employed

- Only one state at each observed time
- Position conveys time information: All states at position 2 are states in 1980.

Section outline

1. Introduction
   - Objectives
   - About longitudinal data analysis
   - What is sequence analysis (SA)?
     - How does SA compare with other longitudinal methods?
     - Types of categorical sequences
   - What kind of questions may SA answer to?

Typical questions

- Are there standard sequences, types of sequences?
- How are those standards linked to covariates such as sex, birth cohort, ... ?
- How does some target variable (e.g., social status) depend on the followed sequence (lived trajectory)?
- . . .

Typical questions

- Are there standard sequences, types of sequences?
- How are those standards linked to covariates such as sex, birth cohort, ... ?
- How does some target variable (e.g., social status) depend on the followed sequence (lived trajectory)?
- . . .

Typical questions

- Are there standard sequences, types of sequences?
- How are those standards linked to covariates such as sex, birth cohort, ... ?
- How does some target variable (e.g., social status) depend on the followed sequence (lived trajectory)?
- ...

Typical questions

- Are there standard sequences, types of sequences?
- How are those standards linked to covariates such as sex, birth cohort, ... ?
- How does some target variable (e.g., social status) depend on the followed sequence (lived trajectory)?
- ...

Sequencing, timing and duration

- For chronological sequences (with time dimension)
- SA can answer questions about:
    - Sequencing: Order in which the different elements occur.
    - Timing: When do the different elements occur?
    - Duration: How long do we stay in the successive states?

Sequencing, timing and duration

- For chronological sequences (with time dimension)
- SA can answer questions about:
  - Sequencing: Order in which the different elements occur.
  - Timing: When do the different elements occur?
  - Duration: How long do we stay in the successive states?

Sequencing, timing and duration

- For chronological sequences (with time dimension)
- SA can answer questions about:
  - Sequencing: Order in which the different elements occur.
  - Timing: When do the different elements occur?
  - Duration: How long do we stay in the successive states?

Sequencing, timing and duration

- For chronological sequences (with time dimension)
- SA can answer questions about:
  - Sequencing: Order in which the different elements occur.
  - Timing: When do the different elements occur?
  - Duration: How long do we stay in the successive states?

Sequencing, timing and duration

- For chronological sequences (with time dimension)
- SA can answer questions about:
  - Sequencing: Order in which the different elements occur.
  - Timing: When do the different elements occur?
  - Duration: How long do we stay in the successive states?

Outline

1 Introduction

2 Overview of what sequence analysis can do

3 About TraMineR

Overview of sequence analysis outcomes

Aim:

- Show what kind of results can be obtained
- as well as how to get the results with our TraMineR package for R
- TraMineR: Trajectory Miner for R (Gabadinho et al., 2011)

Overview of sequence analysis outcomes

Aim:

- Show what kind of results can be obtained
- as well as how to get the results with our TraMineR package for R
- TraMineR: Trajectory Miner for R (Gabadinho et al., 2011)

Overview of sequence analysis outcomes

Aim:

- Show what kind of results can be obtained
- as well as how to get the results with our TraMineR package for R
- TraMineR: Trajectory Miner for R (Gabadinho et al., 2011)

Section outline

2 Overview of what sequence analysis can do

- The mvad example dataset
- Creating the state sequence object
- Rendering sequences
- Characterizing set of sequences
- Longitudinal characteristics
- Dissimilarity-based analyses
  - Cluster analysis
  - Discrepancy analysis

The ‘mvad’ data set

- McVicar and Anyadike-Danes (2002)’s study of school to work transition in Northern Ireland.
- dataset distributed with the TraMineR library.
- 712 cases (survey data).
- 72 monthly activity statuses (July 1993-June 1999)
- States are:

EM	Employment
FE	Further education
HE	Higher education
JL	Joblessness
SC	School
TR	Training.

- 14 additional (binary) variables
- The follow-up starts when respondents finished compulsory school (16 years old).

mvad variables

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

The mvad sequences are in STS form

- The `mvad` sequences are organized in STS form, i.e., each sequence is given as a (row) vector of consecutive states.

```
head(mvad[, 17:22])
```

```
          Sep.93     Oct.93     Nov.93     Dec.93   Jan.94   Feb.94
1  employment employment employment employment training training
2          FE         FE         FE         FE       FE       FE
3    training   training   training   training training training
4    training   training   training   training training training
5          FE         FE         FE         FE       FE       FE
6 joblessness   training   training   training training training
```

- There are many other ways of organizing sequences data and TraMineR supports most of them.

The mvad sequences are in STS form

- The `mvad` sequences are organized in STS form, i.e., each sequence is given as a (row) vector of consecutive states.

```
head(mvad[, 17:22])
```

```
       Sep.93     Oct.93     Nov.93     Dec.93   Jan.94   Feb.94
1  employment employment employment employment training training
2          FE         FE         FE         FE       FE       FE
3    training   training   training   training training training
4    training   training   training   training training training
5          FE         FE         FE         FE       FE       FE
6 joblessness   training   training   training training training
```

- There are many other ways of organizing sequences data and TraMineR supports most of them.

Section outline

2 Overview of what sequence analysis can do

- The mvad example dataset
- Creating the state sequence object
- Rendering sequences
- Characterizing set of sequences
- Longitudinal characteristics
- Dissimilarity-based analyses
  - Cluster analysis
  - Discrepancy analysis

Creating the state sequence object

- Most TraMineR functions for state sequences require a state sequence object as input argument.
- The state sequence object contains
  - the sequences
  - and their attributes (alphabet, labels, colors, weights, ...)
- Hence, we first have to create this object

Creating the state sequence object

- Most TraMineR functions for state sequences require a state sequence object as input argument.
- The state sequence object contains
  - the sequences
  - and their attributes (alphabet, labels, colors, weights, ...)
- Hence, we first have to create this object

Creating the state sequence object

- Most TraMineR functions for state sequences require a state sequence object as input argument.
- The state sequence object contains
  - the sequences
  - and their attributes (alphabet, labels, colors, weights, ...)
- Hence, we first have to create this object

Creating the state sequence object

- Most TraMineR functions for state sequences require a state sequence object as input argument.
- The state sequence object contains
  - the sequences
  - and their attributes (alphabet, labels, colors, weights, ...)
- Hence, we first have to create this object

Creating the state sequence object

- Most TraMineR functions for state sequences require a state sequence object as input argument.
- The state sequence object contains
  - the sequences
  - and their attributes (alphabet, labels, colors, weights, ...)
- Hence, we first have to create this object

Starting TraMineR and creating a state sequence object

- Load TraMineR and the mvad data.

```
library(TraMineR)
data(mvad)
```

- Check the alphabet (from Sept 93 to June 99; i.e., positions 17 to 86: We skip July-August 93)

```
(mvad.alph <- seqstatl(mvad[, 17:86]))
```

```
[1] "employment"  "FE"           "HE"           "joblessness" "school"
[6] "training"
```

- Create the ‘state sequence’ object

```
mvad.lab <- c("employment", "further education",
     "higher education", "joblessness", "school",
     "training")
mvad.shortlab <- c("EM", "FE", "HE", "JL", "SC",
     "TR")
mvad.seq <- seqdef(mvad[, 17:86], alphabet = mvad.alph,
     states = mvad.shortlab, labels = mvad.lab, weights = mvad$weight,
     xtstep = 6)
```

Starting TraMineR and creating a state sequence object

- Load TraMineR and the mvad data.

```
library(TraMineR)
data(mvad)
```

- Check the alphabet (from Sept 93 to June 99; i.e., positions 17 to 86: We skip July-August 93)

```
(mvad.alph <- seqstatl(mvad[, 17:86]))
```

```
[1] "employment"  "FE"          "HE"          "joblessness" "school"
[6] "training"
```

- Create the ‘state sequence’ object

```
mvad.lab <- c("employment", "further education",
     "higher education", "joblessness", "school",
     "training")
mvad.shortlab <- c("EM", "FE", "HE", "JL", "SC",
     "TR")
mvad.seq <- seqdef(mvad[, 17:86], alphabet = mvad.alph,
     states = mvad.shortlab, labels = mvad.lab, weights = mvad$weight,
     xtstep = 6)
```

Starting TraMineR and creating a state sequence object

- Load TraMineR and the mvad data.

```
library(TraMineR)
data(mvad)
```

- Check the alphabet (from Sept 93 to June 99; i.e., positions 17 to 86: We skip July-August 93)

```
(mvad.alph <- seqstatl(mvad[, 17:86]))
```

```
[1] "employment"  "FE"          "HE"          "joblessness" "school"
[6] "training"
```

- Create the 'state sequence' object

```
mvad.lab <- c("employment", "further education",
     "higher education", "joblessness", "school",
     "training")
mvad.shortlab <- c("EM", "FE", "HE", "JL", "SC",
     "TR")
mvad.seq <- seqdef(mvad[, 17:86], alphabet = mvad.alph,
     states = mvad.shortlab, labels = mvad.lab, weights = mvad$weight,
     xtstep = 6)
```

Main sequence object attributes and seqdef arguments

Attribute name	Description	Argument	Default	Retrieve/Set
	input format	`informat=`	`"STS"`	
alphabet	list of states	`states=`	from input data	`alphabet()`
cpal	color palette	`cpal=`	from RColorBrewer	`cpal()`
labels	long state labels	`labels=`	from input data	`stlab()`
cnames	position names	`cnames=`	from input data	`names()`
xtstep	jumps between tick marks	`xtstep=`	`1`	
row.names	row (sequence) labels	`id=`	from input data	`rownames()`
weights	optional case weights	`weights=`	`NULL`	
	missing handling	`left=`	`NA`	
	"	`gaps=`	`NA`	
	"	`right=`	`"DEL"`	

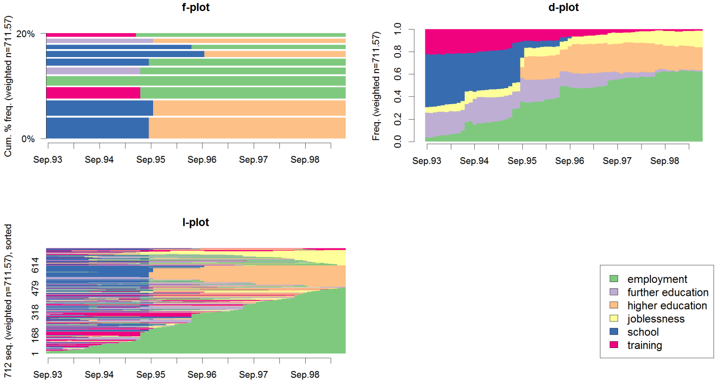
Section outline

2 Overview of what sequence analysis can do

- The mvad example dataset
- Creating the state sequence object
- Rendering sequences
- Characterizing set of sequences
- Longitudinal characteristics
- Dissimilarity-based analyses
    - Cluster analysis
    - Discrepancy analysis

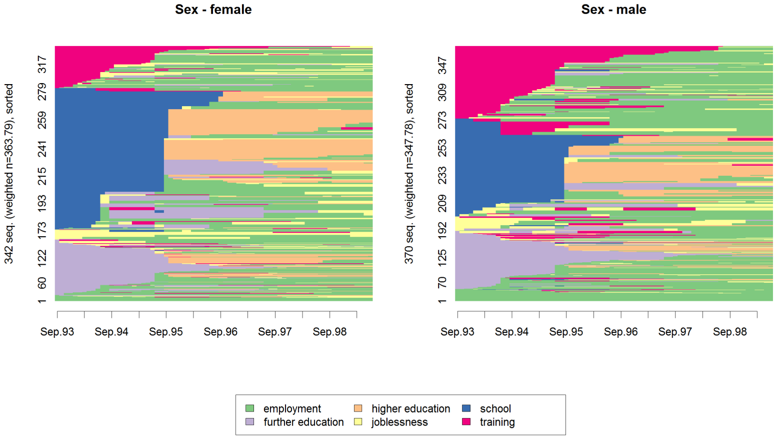
Rendering sequences

```
seqfplot(mvad.seq, withlegend = FALSE, title = "f-plot", border = NA)
seqdplot(mvad.seq, withlegend = FALSE, title = "d-plot", border = NA)
seqIplot(mvad.seq, withlegend = FALSE, title = "I-plot", sortv = "from.end")
seqlegend(mvad.seq, position = "bottomright", fontsize = 1.2)
```



Rendering sequences by group (sex)

```
seqIplot(mvad.seq, group = mvad$male, sortv = "from.start",
    title = "Sex")
```



Section outline

2 Overview of what sequence analysis can do

- The mvad example dataset
- Creating the state sequence object
- Rendering sequences
- Characterizing set of sequences
- Longitudinal characteristics
- Dissimilarity-based analyses
  - Cluster analysis
  - Discrepancy analysis

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: sequence medoid, 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: sequence medoid, 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: sequence medoid, diversity of sequences, ...

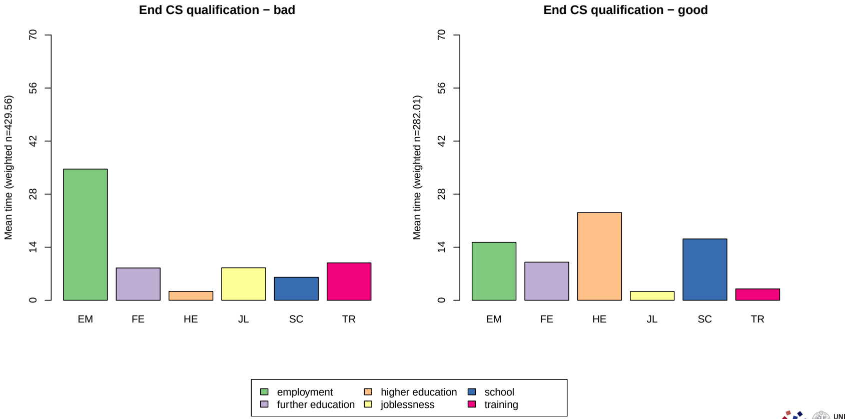
Transition rates

```
round(trate <- seqtrate(mvad.seq), 3)

        [-> EM] [-> FE] [-> HE] [-> JL] [-> SC] [-> TR]
[EM ->]   0.986   0.002   0.003   0.007   0.000   0.002
[FE ->]   0.027   0.950   0.007   0.011   0.001   0.003
[HE ->]   0.010   0.000   0.988   0.001   0.000   0.001
[JL ->]   0.037   0.012   0.002   0.938   0.001   0.010
[SC ->]   0.012   0.008   0.019   0.007   0.950   0.004
[TR ->]   0.037   0.004   0.000   0.015   0.001   0.944
```

Mean time in each state

by qualification gained at end of compulsory school

```
seqmtplot(mvad.seq, group = mvad$gcse5eq, title = "End CS qualification")
```



Sequence of transversal distributions

For bad qualification at end of compulsory school, 9 months

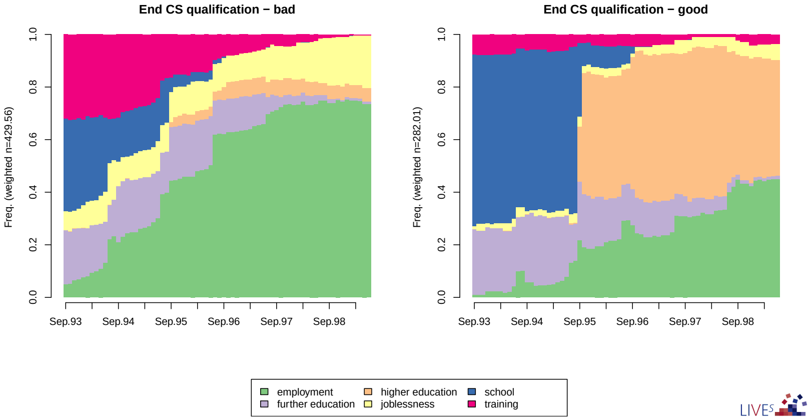
```
seqstatd(mvad.seq[mvad$gcse5eq == "bad", 6:15])
```

```
    [State frequencies]
   Feb.94 Mar.94 Apr.94 May.94 Jun.94 Jul.94 Aug.94 Sep.94 Oct.94 Nov.94
EM   0.08  0.094  0.100   0.11   0.13   0.22   0.23  0.211  0.231  0.244
FE   0.18  0.181  0.176   0.17   0.16   0.13   0.14  0.212  0.211  0.209
HE   0.00  0.000  0.000   0.00   0.00   0.00   0.00  0.000  0.000  0.000
JL   0.10  0.093  0.093   0.11   0.11   0.16   0.15  0.094  0.091  0.084
SC   0.33  0.316  0.316   0.31   0.28   0.17   0.16  0.167  0.171  0.171
TR   0.31  0.316  0.315   0.31   0.32   0.32   0.32  0.316  0.295  0.292

     [Valid states]
   Feb.94 Mar.94 Apr.94 May.94 Jun.94 Jul.94 Aug.94 Sep.94 Oct.94 Nov.94
N     430    430    430    430    430    430    430    430    430    430

     [Entropy index]
   Feb.94 Mar.94 Apr.94 May.94 Jun.94 Jul.94 Aug.94 Sep.94 Oct.94 Nov.94
H    0.82   0.83   0.83   0.84   0.85   0.87   0.87   0.86   0.86   0.86
```

Sequence of transversal distributions (chronogram)

by qualification gained at end of compulsory school

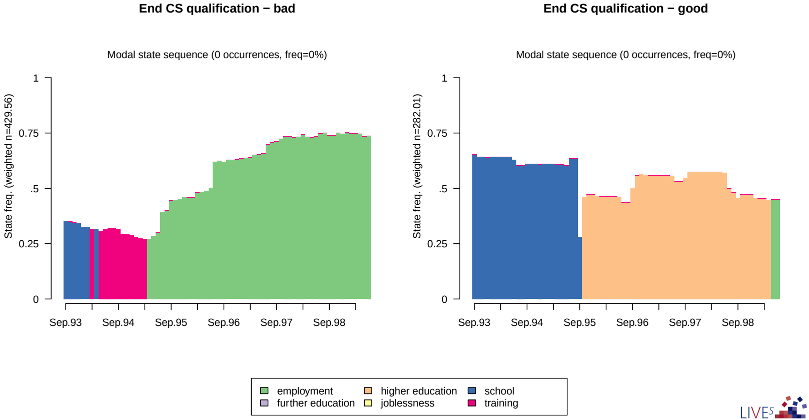
```
seqdplot(mvad.seq, group = mvad$gcse5eq, title = "End CS qualification",
    border = NA)
```



Sequence of modal states

by qualification gained at end of compulsory school

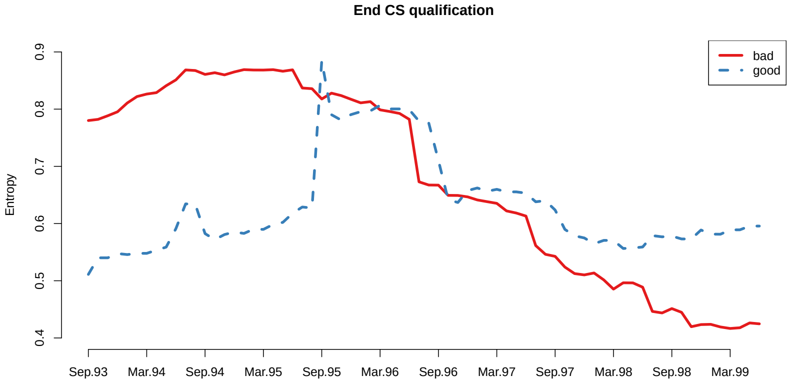
```
seqmsplot(mvad.seq, group = mvad$gcse5eq, title = "End CS qualification",
    border = NA)
```



Transversal entropies

Time evolution of the transversal state diversity

```
seqplot.tentrop(mvad.seq, title = "End CS qualification",
          group = mvad$gcse5eq)
```



Section outline

2 Overview of what sequence analysis can do

- The mvad example dataset
- Creating the state sequence object
- Rendering sequences
- Characterizing set of sequences
- Longitudinal characteristics
- Dissimilarity-based analyses
  - Cluster analysis
  - Discrepancy analysis

Longitudinal Characteristics

- Characteristics of individual sequences

seqlength()	length of the sequence
seqtransn()	number of transitions
seqsubsn()	number of sub-sequences
seqdss()	list of the distinct successive states (DSS)
seqdur()	list of the durations in the states of the DSS
seqistatd()	time in each state (longitudinal distribution)
seqient()	Longitudinal entropy
seqST()	Turbulence (Elzinga and Liefbroer, 2007)
seqici()	Complexity index (Gabadinho et al., 2011)

Distinct successive states and their durations

- SPS format

```
    Sequence
1  (EM,4)-(TR,2)-(EM,64)
2  (FE,36)-(HE,34)
3  (TR,24)-(FE,34)-(EM,10)-(JL,2)
```

- Distinct successive states(DSS)

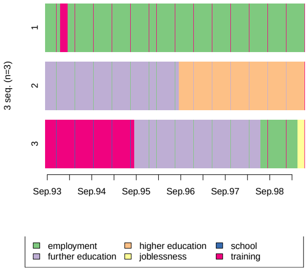
```
seqdss(mvad.seq)[1:3, ]
```

```
    Sequence
1  EM-TR-EM
2  FE-HE
3  TR-FE-EM-JL
```

- Duration in successive states

```
seqdur(mvad.seq)[1:3, 1:5]
```

```
   DUR1 DUR2 DUR3 DUR4 DUR5
1     4    2   64   NA   NA
2    36   34   NA   NA   NA
3    24   34   10    2   NA
```



Distinct successive states and their durations

- SPS format

```
   Sequence
1  (EM,4)-(TR,2)-(EM,64)
2  (FE,36)-(HE,34)
3  (TR,24)-(FE,34)-(EM,10)-(JL,2)
```

- Distinct successive states(DSS)

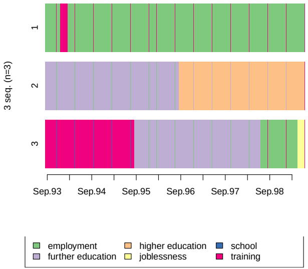
  seqdss(mvad.seq)[1:3,]

```
   Sequence
1  EM-TR-EM
2  FE-HE
3  TR-FE-EM-JL
```

- Duration in successive states

  seqdur(mvad.seq)[1:3, 1:5]

```
   DUR1 DUR2 DUR3 DUR4 DUR5
1     4    2   64   NA   NA
2    36   34   NA   NA   NA
3    24   34   10    2   NA
```



Distinct successive states and their durations

- SPS format

```
   Sequence
1  (EM,4)-(TR,2)-(EM,64)
2  (FE,36)-(HE,34)
3  (TR,24)-(FE,34)-(EM,10)-(JL,2)
```

- Distinct successive states(DSS)

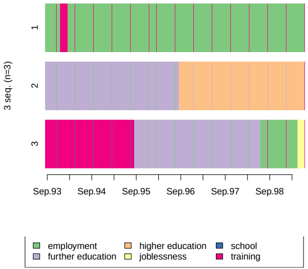
```
seqdss(mvad.seq)[1:3, ]
```

```
   Sequence
1  EM-TR-EM
2  FE-HE
3  TR-FE-EM-JL
```

- Duration in successive states

```
seqdur(mvad.seq)[1:3, 1:5]
```

```
   DUR1 DUR2 DUR3 DUR4 DUR5
1     4    2   64   NA   NA
2    36   34   NA   NA   NA
3    24   34   10    2   NA
```



Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
    - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
    - composite measure based on
        - the number of sub-sequences of the DSS sequence
        - the variance of the durations of the successive states
    - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
    - composite measure based on
        - the number of transitions
        - the longitudinal entropy
    - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
  - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
  - composite measure based on
    - the number of sub-sequences of the DSS sequence
    - the variance of the durations of the successive states
  - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
  - composite measure based on
    - the number of transitions
    - the longitudinal entropy
  - sensitive to state sequencing

Complexity of the sequences

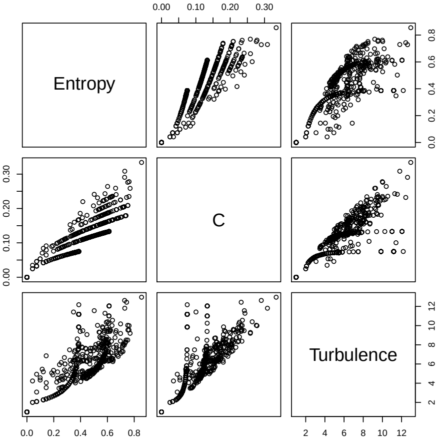
- To evaluate the complexity of a sequence we may consider
- Longitudinal entropy
    - does not account for the sequencing of the states (AABB and ABAB have same entropy)
- Turbulence (Elzinga and Liefbroer, 2007)
    - composite measure based on
        - the number of sub-sequences of the DSS sequence
        - the variance of the durations of the successive states
    - sensitive to state sequencing
- Index of complexity (Gabadinho et al., 2010, 2011)
    - composite measure based on
        - the number of transitions
        - the longitudinal entropy
    - sensitive to state sequencing

Computing the sequence complexity measures

```
mvad.ient <- seqient(mvad.seq)
mvad.cplx <- seqici(mvad.seq)
mvad.turb <- seqST(mvad.seq)
ctab <- data.frame(mvad.ient, mvad.cplx, mvad.turb)
```

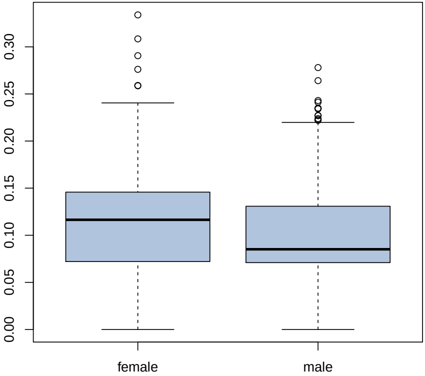
Comparing the measures

```
plot(ctab)
```



Distribution of complexity by sex

```
boxplot(mvad.cplx ~ mvad$male, col = "lightsteelblue")
```



Analyzing how complexity is related to covariates

Regressing complexity on covariates

```
lm.ici <- lm(mvad.cplx ~ male + funemp + gcse5eq, data = mvad)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.109	0.004	28.01	0.000
male	-0.013	0.004	-3.04	0.002
father unemployed	0.007	0.006	1.24	0.216
good ECS grade	0.010	0.005	2.20	0.028

Section outline

2 Overview of what sequence analysis can do
- The mvad example dataset
- Creating the state sequence object
- Rendering sequences
- Characterizing set of sequences
- Longitudinal characteristics
- Dissimilarity-based analyses
  - Cluster analysis
  - Discrepancy analysis

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Summary of available distances

Distance	Method	Position-wise	Additional arguments
Count of common attributes			
Simple Hamming	`HAM`	Yes	
Longest Common Prefix	`LCP`	Yes	
Longest Common Suffix	`RLCP`	Yes	
Longest Common Subsequence	`LCS`	No	
Edit distances			
Optimal Matching	`OM`	No	Insertion/deletion costs (`indel`) and substitution costs matrix (`sm`)
Hamming	`HAM`	Yes	substitution costs matrix (`sm`)
Dynamic Hamming	`DHD`	Yes	substitution costs matrix (`sm`)

Other distances

- There exist many other distances not yet implemented in TraMineR.
    - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
    - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
    - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
    - Distances accounting for the common future (Rousset et al., 2011)
    - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
    - OM of transitions instead of states (Biemann, 2011)
- Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).

Other distances

- There exist many other distances not yet implemented in TraMineR.
  - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
  - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
  - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
  - Distances accounting for the common future (Rousset et al., 2011)
  - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
  - OM of transitions instead of states (Biemann, 2011)
- Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).

Other distances

- There exist many other distances not yet implemented in TraMineR.
    - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
    - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
    - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
    - Distances accounting for the common future (Rousset et al., 2011)
    - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
    - OM of transitions instead of states (Biemann, 2011)
- Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).

Other distances

- There exist many other distances not yet implemented in TraMineR.
    - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
    - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
    - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
    - Distances accounting for the common future (Rousset et al., 2011)
    - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
    - OM of transitions instead of states (Biemann, 2011)
- Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).

Other distances

- There exist many other distances not yet implemented in TraMineR.
  - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
  - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
  - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
  - Distances accounting for the common future (Rousset et al., 2011)
  - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
  - OM of transitions instead of states (Biemann, 2011)
- Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).

Other distances

- There exist many other distances not yet implemented in TraMineR.
    - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
    - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
    - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
    - Distances accounting for the common future (Rousset et al., 2011)
    - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
    - OM of transitions instead of states (Biemann, 2011)
- Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).

Other distances

- There exist many other distances not yet implemented in TraMineR.
  - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
  - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
  - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
  - Distances accounting for the common future (Rousset et al., 2011)
  - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
  - OM of transitions instead of states (Biemann, 2011)
- Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).

Other distances

- There exist many other distances not yet implemented in TraMineR.
  - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
  - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
  - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
  - Distances accounting for the common future (Rousset et al., 2011)
  - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
  - OM of transitions instead of states (Biemann, 2011)
- Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).

Dissimilarity matrix

- TraMineR provides the seqdist function

```
## OM distances with custom indel and substitution
## costs used by McVicar and Anyadike-Danes (2012).
subm.custom  <- matrix(
         c(0,1,1,2,1,1,
           1,0,1,2,1,2,
           1,1,0,3,1,2,
           2,2,3,0,3,1,
           1,1,1,3,0,2,
           1,2,2,1,2,0),
           nrow = 6, ncol = 6, byrow = TRUE,
           dimnames = list(mvad.shortlab, mvad.shortlab))
mvad.dist <- seqdist(mvad.seq, method="OM", indel=4, sm=subm.custom)
dim(mvad.dist)
```

```
[1] 712 712
```

Dissimilarity matrix

```
print(mvad.seq[1:4, ], format = "SPS")
```

```
  Sequence
1 (EM,4)-(TR,2)-(EM,64)
2 (FE,36)-(HE,34)
3 (TR,24)-(FE,34)-(EM,10)-(JL,2)
4 (TR,47)-(EM,14)-(JL,9)
```

```
mvad.dist[1:4, 1:6]
```

```
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0   72   60   63   72   33
[2,]   72    0   86  135   11  104
[3,]   60   86    0   71   97   49
[4,]   63  135   71    0  135   32
```

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - `agnes()`: agglomerative nesting (average, UPGMA, WPGMA, ward, beta-flexible, ...)
    - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - ✓ `agnes()`: agglomerative nesting (average, UPGMA, WPGMA, ward, beta-flexible, ...)
    - ✓ `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - `agnes()`: agglomerative nesting (average, UPGMA WPGMA, ward, beta-flexible, ...)
    - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - `agnes()`: agglomerative nesting (average, UPGMA WPGMA, ward, beta-flexible, ...)
    - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - `agnes()`: agglomerative nesting (average, UPGMA WPGMA, ward, beta-flexible, ...)
    - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - `agnes()`: agglomerative nesting (average, UPGMA WPGMA, ward, beta-flexible, ...)
    - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - `agnes()`: agglomerative nesting (average, UPGMA WPGMA, ward, beta-flexible, ...)
    - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - `agnes()`: agglomerative nesting (average, UPGMA WPGMA, ward, beta-flexible, ...)
    - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

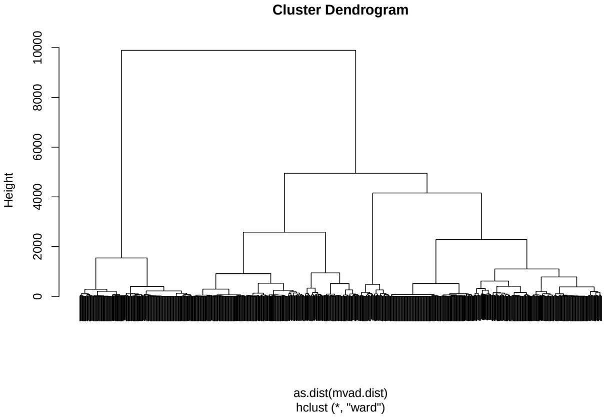
- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
    - `hclust()` base function (can account for weights)
    - Package `cluster` (does not accept weights!):
        - `agnes()`: agglomerative nesting (average, UPGMA WPGMA, ward, beta-flexible, ...)
        - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
    - Packages: `cluster`, `fastclust`, `flashClust`, ...
    - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
- For hierarchical clustering
  - `hclust()` base function (can account for weights)
  - Package `cluster` (does not accept weights!):
    - `agnes()`: agglomerative nesting (average, UPGMA WPGMA, ward, beta-flexible, ...)
    - `diana()`: divisive partitioning
- For PAM and other direct partitioning methods
  - Packages: `cluster`, `fastclust`, `flashClust`, ...
  - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Example: Hierarchical clustering (Ward)

```
mvad.clusterward <- hclust(as.dist(mvad.dist), method = "ward",
     members = mvad$weight)
plot(mvad.clusterward, labels = FALSE)
```



PAM clustering

- PAM much faster, but must set *a priori* number k of clusters.
- `WeightedCluster` offers nice tools to help selecting k.
- $k = 4$ was found to be good choice.
- PAM with function `wcKMedoids` from `WeightedCluster`

```
library(WeightedCluster)
set.seed(4)
pam.mvad <- wcKMedoids(mvad.dist, k = 4, weight = mvad$weight)
```

- Cluster membership is in `pam.mvad$clustering`

```
mvad.cl4 <- pam.mvad$clustering
table(mvad.cl4)
```

```
mvad.cl4
 66 467 607 641
190 305 160  57
```

PAM clustering

- PAM much faster, but must set *a priori* number k of clusters.
- `WeightedCluster` offers nice tools to help selecting k.
- $k = 4$ was found to be good choice.
- PAM with function `wcKMedoids` from `WeightedCluster`

```
library(WeightedCluster)
set.seed(4)
pam.mvad <- wcKMedoids(mvad.dist, k = 4, weight = mvad$weight)
```

- Cluster membership is in `pam.mvad$clustering`

```
mvad.cl4 <- pam.mvad$clustering
table(mvad.cl4)
```

```
mvad.cl4
 66 467 607 641
190 305 160  57
```

PAM clustering

- PAM much faster, but must set *a priori* number k of clusters.
- `WeightedCluster` offers nice tools to help selecting k.
- $k = 4$ was found to be good choice.
- PAM with function `wcKMedoids` from `WeightedCluster`

```
library(WeightedCluster)
set.seed(4)
pam.mvad <- wcKMedoids(mvad.dist, k = 4, weight = mvad$weight)
```

- Cluster membership is in `pam.mvad$clustering`

```
mvad.cl4 <- pam.mvad$clustering
table(mvad.cl4)
```

```
mvad.cl4
 66 467 607 641
190 305 160  57
```

PAM clustering

- PAM much faster, but must set *a priori* number k of clusters.
- `WeightedCluster` offers nice tools to help selecting k.
- $k = 4$ was found to be good choice.
- PAM with function `wcKMedoids` from `WeightedCluster`

```
library(WeightedCluster)
set.seed(4)
pam.mvad <- wcKMedoids(mvad.dist, k = 4, weight = mvad$weight)
```

- Cluster membership is in `pam.mvad$clustering`

```
mvad.cl4 <- pam.mvad$clustering
table(mvad.cl4)
```

```
mvad.cl4
 66 467 607 641
190 305 160  57
```

PAM clustering

- PAM much faster, but must set *a priori* number k of clusters.
- `WeightedCluster` offers nice tools to help selecting k.
- $k = 4$ was found to be good choice.
- PAM with function `wcKMedoids` from `WeightedCluster`

  ```
  library(WeightedCluster)
  set.seed(4)
  pam.mvad <- wcKMedoids(mvad.dist, k = 4, weight = mvad$weight)
  ```

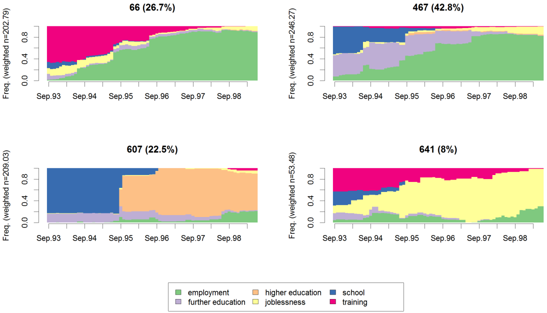
- Cluster membership is in `pam.mvad$clustering`

  ```
  mvad.cl4 <- pam.mvad$clustering
  table(mvad.cl4)
  ```

```
mvad.cl4
 66 467 607 641
190 305 160  57
```

Labeling the PAM clusters

```
seqdplot(mvad.seq, group = group.p(mvad.cl4), border = NA)
```

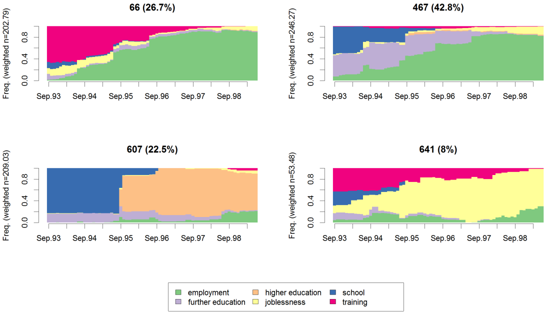


- Rearranging cluster order and defining labels

```
cl4.labels <- c("FE-Employment", "Training-Employment", "Education",
    "Joblessness")
mvad.cl4.factor <- factor(mvad.cl4, levels = c(467, 66, 607,
    641), labels = cl4.labels)
```

Labeling the PAM clusters

```
seqdplot(mvad.seq, group = group.p(mvad.cl4), border = NA)
```

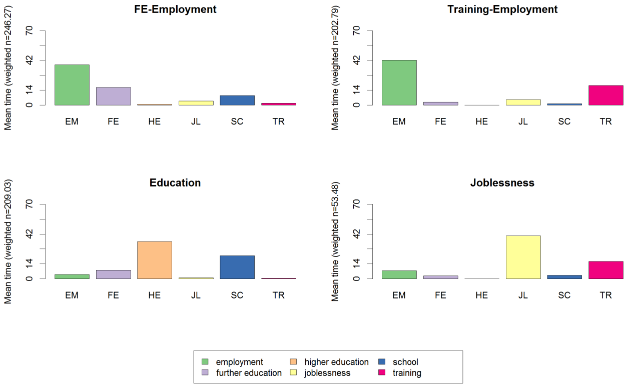


- Rearranging cluster order and defining labels

```
cl4.labels <- c("FE-Employment", "Training-Employment", "Education",
    "Joblessness")
mvad.cl4.factor <- factor(mvad.cl4, levels = c(467, 66, 607,
    641), labels = cl4.labels)
```

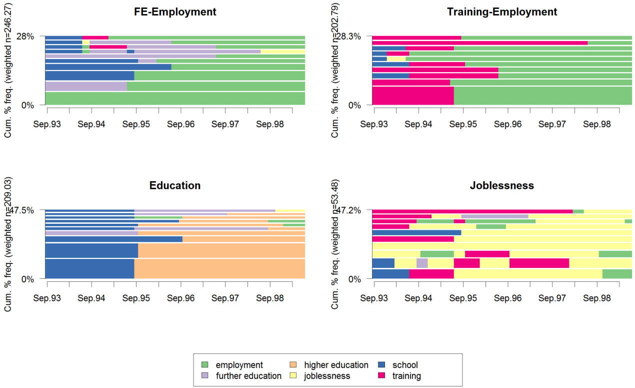
Mean time in each state

```
seqmtplot(mvad.seq, group = mvad.cl4.factor)
```



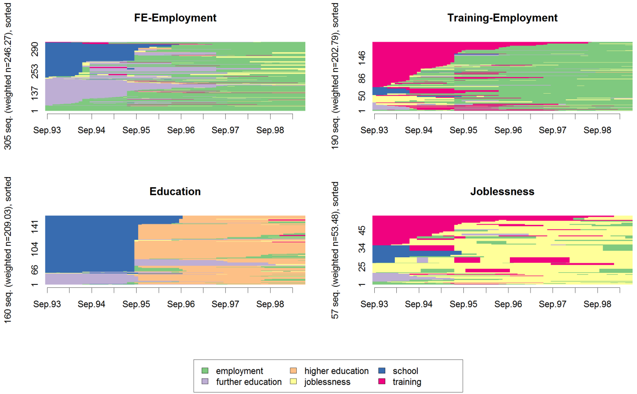
Most frequent sequences

```
seqfplot(mvad.seq, group = mvad.cl4.factor, border = NA)
```



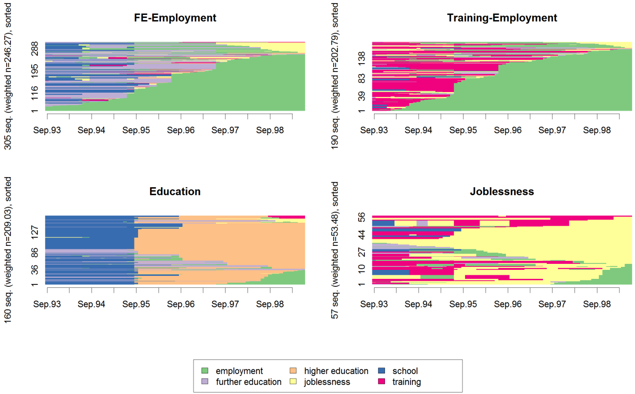
Individual sequences (sorted by states from start)

```
seqIplot(mvad.seq, group = mvad.cl4.factor, sortv = "from.start")
```



Sorted by states from the end

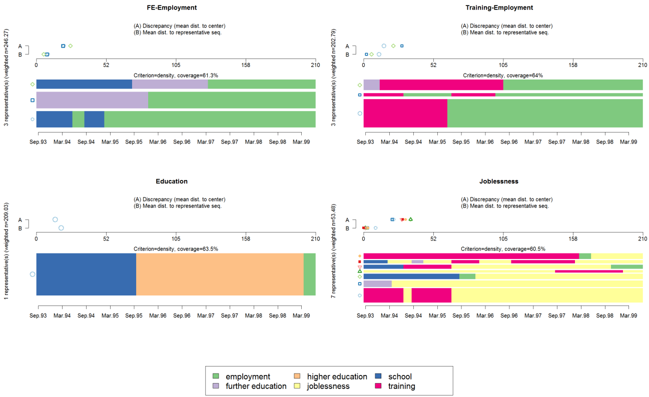
```
seqIplot(mvad.seq, group = mvad.cl4.factor, sortv = "from.end")
```



Representative sequences (Gabadinho et al., 2011)

Smallest set of patterns with given percentage of sequences in their neighborhood

```
seqrplot(mvad.seq, group = mvad.cl4.factor, dist.matrix = mvad.dist,
    trep = 0.6, sim = 0.15, border = NA, cex.legend = 1.5)
```



Discrepancy of sequences

- Sum of squares SS can be expressed in terms of distances between pairs

$$\begin{aligned} SS &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij} \end{aligned}$$

- Setting d_{ij} equal to OM, LCP, LCS ... distance, we get SS.
- From which we can measure the dispersion with the pseudo-variance SS/n.
- And run ANOVA analyses (Studer et al., 2011, 2010, 2009).

Discrepancy of sequences

- Sum of squares SS can be expressed in terms of distances between pairs

$$
\begin{aligned}
SS &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}
\end{aligned}
$$

- Setting d_{ij} equal to OM, LCP, LCS ... distance, we get SS.
- From which we can measure the dispersion with the pseudo-variance SS/n.
- And run ANOVA analyses (Studer et al., 2011, 2010, 2009).

Discrepancy of sequences

- Sum of squares SS can be expressed in terms of distances between pairs

$$\begin{aligned} SS &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij} \end{aligned}$$

- Setting d_{ij} equal to OM, LCP, LCS ... distance, we get SS.
- From which we can measure the dispersion with the pseudo-variance SS/n.
- And run ANOVA analyses (Studer et al., 2011, 2010, 2009).

Discrepancy of sequences

- Sum of squares SS can be expressed in terms of distances between pairs

$$SS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}$$

- Setting d_{ij} equal to OM, LCP, LCS ... distance, we get SS.
- From which we can measure the dispersion with the pseudo-variance SS/n.
- And run ANOVA analyses (Studer et al., 2011, 2010, 2009).

Computing the dispersion

- For the whole set of sequences

```
dissvar(mvad.dist)
```

```
[1] 32.06
```

- By cluster (dissvar.grp from library TraMineRextras)

```
data.frame(Dispersion = dissvar.grp(mvad.dist, group = mvad.cl4.factor))
```

```
                    Dispersion
FE-Employment            18.60
Training-Employment      17.89
Education                15.90
Joblessness              27.14
```

Analysis of sequence discrepancy

- Running an ANOVA-like analysis for `gcse5eq`

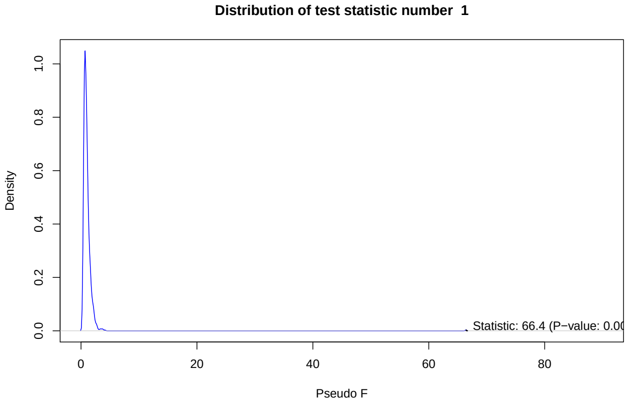
```
da <- dissassoc(mvad.dist, group = mvad$gcse5eq, R = 1000)

print(da)
```

ANOVA output

```
Pseudo ANOVA table:
           SS  df     MSE
Exp      1952   1  1952.4
Res     20871 710    29.4
Total   22823 711    32.1

Test values  (p-values based on 1000 permutation):
                 t0 p.value
Pseudo F   66.41934   0.001
Pseudo Fbf 67.37188   0.001
Pseudo R2   0.08555   0.001
Bartlett    0.14693   0.339
Levene      0.77397   0.403

Inconclusive intervals:
0.00383  <  0.01  <  0.0162
0.03649  <  0.05  <  0.0635

Discrepancy per level:
        n discrepancy
bad   452       29.76
good  260       28.53
Total 712       32.06
```

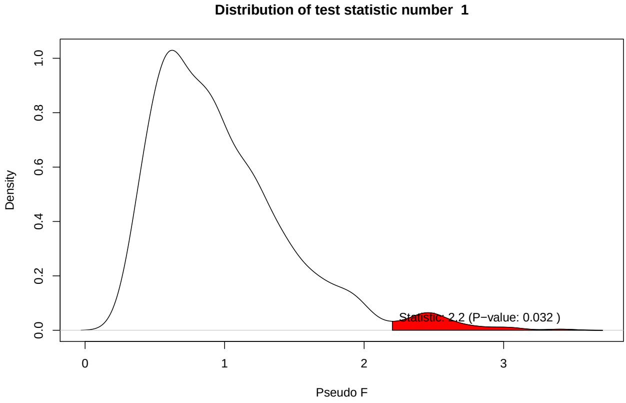
Distribution of pseudo F, gcse5eq

```
hist(da, col = "blue", xlim = c(0, 90))
```



Distribution of pseudo F, livboth

```
da.lb <- dissassoc(mvad.dist, group = mvad$livboth, R = 1000)
hist(da.lb)
```



Differences over time

- How do differences between groups vary over time?
- At which age do trajectories most differ across birth cohorts?
- Compute R^2 for short sliding windows (length 2)
- We get thus a sequence of R^2, which can be plotted
- Similarly, we can plot series of
  - total within (residual) discrepancy (SS_W)
  - within discrepancy of each group (SS_G)

Differences over time

- How do differences between groups vary over time?
- At which age do trajectories most differ across birth cohorts?
- Compute R^2 for short sliding windows (length 2)
- We get thus a sequence of R^2, which can be plotted
- Similarly, we can plot series of
  - total within (residual) discrepancy (SS_W)
  - within discrepancy of each group (SS_G)

Differences over time

- How do differences between groups vary over time?
- At which age do trajectories most differ across birth cohorts?
- Compute R^2 for short sliding windows (length 2)
- We get thus a sequence of R^2, which can be plotted
- Similarly, we can plot series of
    - total within (residual) discrepancy (SS_W)
    - within discrepancy of each group (SS_G)

Differences over time

Grade at end of compulsory school

```
mvad.diff <- seqdiff(mvad.seq, group = mvad$gcse5eq)
```

```
mvad.diff$stat[c(1, 13, 25, 37), ]
```

```
       Pseudo F Pseudo Fbf Pseudo R2 Bartlett  Levene
Sep.93    41.46      44.64   0.05520  9.87187  76.271
Sep.94    72.00      77.42   0.09213  9.49256 104.501
Sep.95    50.52      50.37   0.06646  0.06569   1.041
Sep.96   104.80     103.06   0.12869  0.76633   2.748
```

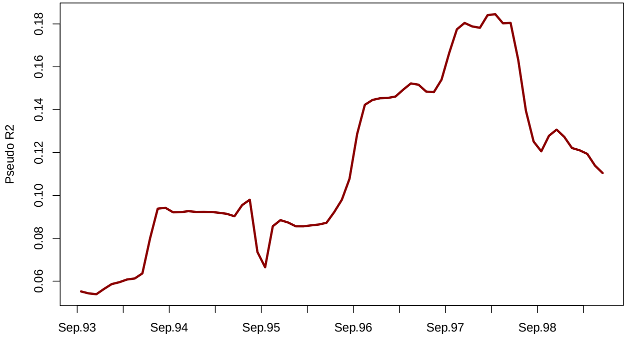
```
mvad.diff$discrepancy[c(1, 13, 25, 37), ]
```

```
          bad   good  Total
Sep.93 0.3620 0.2561 0.3387
Sep.94 0.3876 0.2761 0.3783
Sep.95 0.3590 0.3691 0.3888
Sep.96 0.2862 0.3147 0.3415
```

Plotting R-squares over time

Grade at end of compulsory school

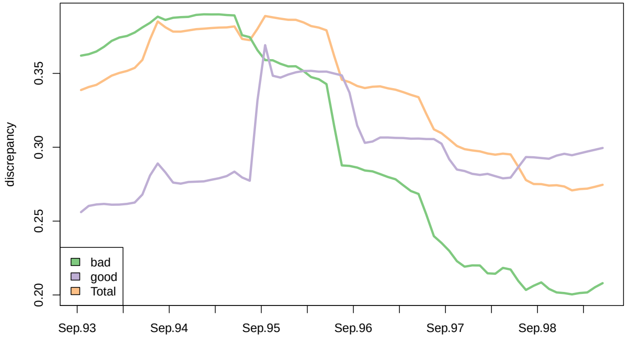
```
plot(mvad.diff, lwd = 3, col = "darkred", xtstep = 6)
```



Plotting within discrepancies over time

Grade at end of compulsory school

```
plot(mvad.diff, lwd = 3, stat = "discrepancy", xtstep = 6,
     legendposition = "bottomleft")
```



Tree structured discrepancy analysis

- Objective: Find the most important predictors and their interactions.
- Iteratively segment the cases using values of covariates (predictors)
- Such that groups be as homogenous as possible.

- At each step, we select the covariate and split with highest R^2.
- Significance of split is assessed through a permutation F test.
- Growing stops when the selected split is not significant.

Tree structured discrepancy analysis

- Objective: Find the most important predictors and their interactions.
- Iteratively segment the cases using values of covariates (predictors)
- Such that groups be as homogenous as possible.

- At each step, we select the covariate and split with highest R^2.
- Significance of split is assessed through a permutation F test.
- Growing stops when the selected split is not significant.

Growing the tree

```
dt <- seqtree(mvad.seq ~ male + Grammar + funemp + gcse5eq +
     fmpr + livboth, weighted = FALSE, data = mvad, diss = mvad.dist,
     R = 5000)

print(dt, gap = 3)
```

Tree in text form

```
Dissimilarity tree:
 Parameters: minSize=35.6, maxdepth=5, R=5000, pval=0.01
 Formula: mvad.seq ~ male + Grammar + funemp + gcse5eq + fmpr + livboth
 Global R2: 0.12

 Fitted tree:

 |-- Root  (n: 712 disc: 32)
   |-> gcse5eq 0.086
       |-- [ bad ]  (n: 452 disc: 30)
         |-> funemp 0.017
          |-- [ no ]  (n: 362 disc: 28)
            |-> male 0.014
             |-- [ female ]    (n: 146 disc: 31)[(FE,2)-(EM,68)] *
             |-- [ male ]    (n: 216 disc: 25)[(EM,70)] *
          |-- [ yes ]    (n: 90 disc: 36)[(EM,70)] *
       |-- [ good ]  (n: 260 disc: 29)
         |-> Grammar 0.048
          |-- [ no ]    (n: 183 disc: 30)[(FE,22)-(EM,48)] *
          |-- [ yes ]    (n: 77 disc: 21)[(SC,25)-(HE,45)] *
```

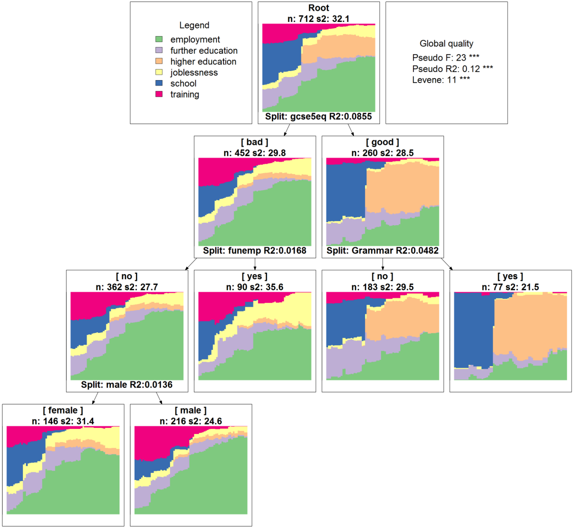
Graphical tree

- The graphical rendering uses Graphviz http://www.graphviz.org/

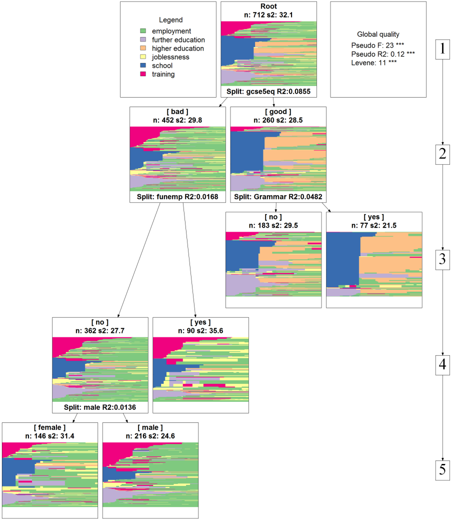
```
R> seqtreedisplay(dt, filename = "fg_mvadseqtree.png",
+        type = "d", border = NA)
```

- The plot is produced as a png file and displayed with the default program associated to this extension.

Graphical Tree



Graphical Tree, using I-plots and showdepth=TRUE



Outline

1. Introduction
2. Overview of what sequence analysis can do
3. About TraMineR

TraMineR: What is it?

TraMineR

- Trajectory Miner in R: a toolbox for exploring, rendering and analyzing categorical sequence data

- Developed within the SNF (Swiss National Fund for Scientific Research) project Mining event histories 1/2007-1/2011
- ... development goes on within IP 14 methodological module of the NCCR LIVES: Overcoming vulnerability: Life course perspectives (http://www.lives-nccr.ch) .

TraMineR, Who?

- Under supervision of a scientific committee:
  - Gilbert Ritschard (Statistics for social sciences)
  - Alexis Gabadinho (Demography)
  - Nicolas S. Müller (Sociology, Computer science)
  - Matthias Studer (Economics, Sociology)
- Additional members of the development team:
  - Reto Bürgin (Statistics)
  - Emmanuel Rousseaux (KDD and Computer science)

  both PhD students within NCCR LIVES IP-14

TraMineR: Where and why in R?

- Package for the free open source R statistical environment
  - freely available on the CRAN (Comprehensive R Archive Network) http://cran.r-project.org

```
R> install.packages("TraMineR", dependencies=TRUE)
```

- TraMineR runs in R, it can straightforwardly be combined with other R commands and libraries. For example:
  - dissimilarities obtained with TraMineR can be inputted to already optimized processes for clustering, MDS, self-organizing maps, ...
  - TraMineR 's plots can be used to render clustering results;
  - complexity indexes can be used as dependent or explanatory variables in linear and non-linear regression, ...

TraMineR: Where and why in R?

- Package for the free open source R statistical environment
  - freely available on the CRAN (Comprehensive R Archive Network) http://cran.r-project.org

```
R> install.packages("TraMineR", dependencies=TRUE)
```

- TraMineR runs in R, it can straightforwardly be combined with other R commands and libraries. For example:
  - dissimilarities obtained with TraMineR can be inputted to already optimized processes for clustering, MDS, self-organizing maps, ...
  - TraMineR 's plots can be used to render clustering results;
  - complexity indexes can be used as dependent or explanatory variables in linear and non-linear regression, ...

TraMineR's features

- Handling of longitudinal data and conversion between various sequence formats
- Plotting sequences (distribution plot, frequency plot, index plot and more)
- Individual longitudinal characteristics of sequences (length, time in each state, longitudinal entropy, turbulence, complexity and more)
- Sequence of transversal characteristics by position (transversal state distribution, transversal entropy, modal state)
- Other aggregated characteristics (transition rates, average duration in each state, sequence frequency)
- Dissimilarities between pairs of sequences (Optimal matching, Longest common subsequence, Hamming, Dynamic Hamming, Multichannel and more)
- Representative sequences and discrepancy measure of a set of sequences
- ANOVA-like analysis and regression tree of sequences
- Rendering and highlighting frequent event sequences
- Extracting frequent event subsequences
- Identifying most discriminating event subsequences
- Association rules between subsequences

Other programs for sequence analysis

- Optimize (Abbott, 1997)
  - Computes optimal matching distances
  - No longer supported
- TDA (Rohwer and Pötter, 2002)
  - free statistical software, computes optimal matching distances
- Stata, SQ-Ados (Brzinsky-Fay et al., 2006)
  - free, but licence required for Stata
  - optimal matching distances, visualization and a few more
  - See also the add-ons by Brendan Halpin
    http://teaching.sociology.ul.ie/seqanal/
- CHESA free program by Elzinga (2007a)
  - Various metrics, including original ones based on non-aligning methods
  - Turbulence

Thank you!

References I

Abbott, A. (1997). Optimize. http://home.uchicago.edu/˜aabbott/om.html.

Aisenbrey, S. and A. E. Fasang (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods and Research 38*(3), 430–462.

Beck, N. and J. N. Katz (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review 89*, 634–647.

Bejerano, G. and G. Yona (2001). Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics 17*(1), 23–43.

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.

Biemann, T. (2011). A transition-oriented approach to optimal matching. *Sociological Methodology 41*(1), 195–221.

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.

References II

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

Elzinga, C. H. (2003). Sequence similarity: A non-aligning technique. *Sociological Methods and Research 31*, 214–231.

Elzinga, C. H. (2007a). CHESA 2.1 User manual. User guide, Dept of Social Science Research Methods, Vrije Universiteit, Amsterdam.

Elzinga, C. H. (2007b). Sequence analysis: Metric representations of categorical time series. Manuscript, Dept of Social Science Research Methods, Vrije Universiteit, Amsterdam.

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.

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. New York: Cambridge University Press.

Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1–37.

References III

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.

Gabadinho, A., G. Ritschard, M. Studer, et N. S. Müller (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des nouvelles technologies de l'information RNTI E-19*, 61–66.

Gabadinho, A., G. Ritschard, M. Studer, et N. S. Müller (2011). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, et J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Volume 128 of *Communications in Computer and Information Science (CCIS)*, pp. 94–106. Springer-Verlag.

Gauthier, J.-A., E. D. Widmer, P. Bucher, and C. Notredame (2009). How much does it cost? Optimization of costs in sequence analysis of social science data. *Sociological Methods and Research 38*, 197–231.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

References IV

Grelet, Y. (2002). Des typologies de parcours: Méthodes et usages. Notes de travail Génération 92, Céreq, Paris.

Halpin, B. (2010). Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods and Research 38*(3), 365–388.

Hedeker, D. (2007). Multilevel models for ordinal and nominal variables. In J. de Leeuw and E. Meijer (Eds.), *Multilevel Models for Ordinal and Nominal Variables*, Chapter 6, pp. 239–276. Springer.

Hollister, M. (2009). Is Optimal Matching Suboptimal? *Sociological Methods Research 38*(2), 235–264.

Massoni, S., M. Olteanu, and P. Rousset (2009). Career-path analysis using optimal matching and self-organizing maps. In *Advances in Self-Organizing Maps: 7th International Workshop, WSOM 2009, St. Augustine, FL, USA, June 8-10, 2009*, Volume 5629 of *Lecture Notes in Computer Science*, pp. 154–162. Berlin: Springer.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology 60*, 577–605.

References V

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.

Müller, N. S. (2011). *Inégalités sociales et effets cumulés au cours de la vie: concepts et méthodes*, Volume SES-764 of *Collection des thèses*. Université de Genève, Faculté des sciences économiques et sociales.

Oh, S.-J. and J.-Y. Kim (2004). A hierarchical clustering algorithm for categorical sequence data. *Information Processing Letters 91*(3), 135–140.

Perroux, O. et M. Oris (2005). Présentation de la base de données de la population de Genève de 1816 à 1843. Séminaire statistique sciences sociales, Université de Genève.

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.

Rohwer, G. and U. Pötter (2002). TDA user's manual. Software, Ruhr-Universität Bochum, Fakultät für Sozialwissenschaften, Bochum.

References VI

Rousset, P., J.-F. Giret, and Y. Grelet (2011). Les parcours d'insertion des jeunes: une analyse longitudinale basée sur les cartes de kohonen. Net.Doc 82, Céreq.

Sharma, K. R. (2008). *Bioinformatics – Sequence Alignment and Markov Models*. New York: McGraw-Hill.

Singer, J. D. and J. B. Willett (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.

Studer, M. (2012a). *Étude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles*, Volume SES-777 of *Collection des thèses*. Université de Genève, Faculté des sciences économiques et sociales.

Studer, M. (2012b). *WeightedCluster: Clustering of Weighted Data*. R package version 0.9.

Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2009). Analyse de dissimilarités par arbre d'induction. *Revue des nouvelles technologies de l'information RNTI E-15*, 7–18.

References VII

Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2010). Discrepancy analysis of complex objects using dissimilarities. In F. Guillet, G. Ritschard, D. A. Zighed, et H. Briand (Eds.), *Advances in Knowledge Discovery and Management*, Volume 292 of *Studies in Computational Intelligence*, pp. 3–19. Berlin : Springer.

Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research 40*(3), 471–510.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data*. New York: Springer.

Wanner, P. et E. Delaporte (2001). Reconstitution de trajectoires de vie à partir des données de l'état civil (BEVNAT). une étude de faisabilité. Rapport de recherche, Forum Suisse des Migrations.

Wernli, B. (2010). A Swiss survey landscape for communication research. In *Università della Svizzera Italiana, USI, Lugano, 2010, June 15, Institute of Communication and Health*.

References VIII

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.