Exploring Life Trajectories

From their visualization to the identification of typical sequences

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

Journée MONEITHS, Lyon, 21 février 2013

Outline

Outline

Section outline

1. Sequences and sequence analysis
   - About sequence data
   - What is sequence analysis (SA)?

Sequence data

Sequence data

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

- Example:

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

- Life trajectories described as chronological sequence data
  - Time order of the elements
  - Categorical longitudinal data

Sequence data

Sequence data

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

- Example:

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

- Life trajectories described as chronological sequence data
  - Time order of the elements
  - Categorical longitudinal data

Sequence data

Sequence data

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

- Example:

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

- Life trajectories described as chronological sequence data
  - Time order of the elements
  - Categorical longitudinal data

Sequence data

Sequence data

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

- Example:

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

- Life trajectories described as chronological sequence data
  - Time order of the elements
  - Categorical longitudinal data

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	$\cdots$
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.

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
- (Statistical) match of data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

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
- (Statistical) match of data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

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
- (Statistical) match of data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

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
- (Statistical) match of data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

State sequences: an example

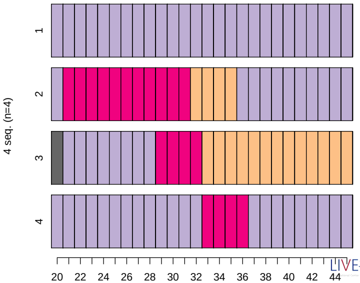
- Occupational state sequences, ages 20 to 45 (from SHP)

  FT = Full Time, PT = Part Time, AH = At Home, ED = Education, ...

```
  Sequence
1 FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-F
2 FT-AH-AH-AH-AH-AH-AH-AH-AH-AH-AH-AH-PT-PT-PT-PT-FT-FT-FT-FT-FT-FT-FT-FT-FT-F
3 ED-FT-FT-FT-FT-FT-FT-FT-FT-AH-AH-AH-AH-PT-PT-PT-PT-PT-PT-PT-PT-PT-PT-PT-PT-P
4 FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-AH-AH-AH-AH-FT-FT-FT-FT-FT-FT-FT-FT-F
```

- Compact representation

```
    Sequence
[1] (FT,26)
[2] (FT,1)-(AH,11)-(PT,4)-(FT,10)
[3] (ED,1)-(FT,8)-(AH,4)-(PT,13)
[4] (FT,13)-(AH,4)-(FT,9)
```



State sequences: an example

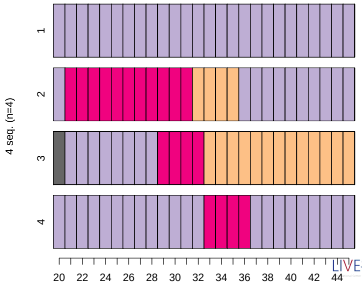
- Occupational state sequences, ages 20 to 45 (from SHP)

  FT = Full Time, PT = Part Time, AH = At Home, ED = Education, ...

```
  Sequence
1 FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-F
2 FT-AH-AH-AH-AH-AH-AH-AH-AH-AH-AH-AH-PT-PT-PT-PT-FT-FT-FT-FT-FT-FT-FT-FT-FT-F
3 ED-FT-FT-FT-FT-FT-FT-FT-FT-AH-AH-AH-AH-PT-PT-PT-PT-PT-PT-PT-PT-PT-PT-PT-PT-P
4 FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-FT-AH-AH-AH-AH-FT-FT-FT-FT-FT-FT-FT-FT-F
```

- Compact representation

```
    Sequence
[1] (FT,26)
[2] (FT,1)-(AH,11)-(PT,4)-(FT,10)
[3] (ED,1)-(FT,8)-(AH,4)-(PT,13)
[4] (FT,13)-(AH,4)-(FT,9)
```



State versus event sequences

- Life course trajectories described by chronological 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

- Life course trajectories described by chronological 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

- Life course trajectories described by chronological 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

- Life course trajectories described by chronological 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

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

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
- Popularized in social sciences by Abbott (Abbott and Forrest, 1986)
- Some other important names: Elzinga (2003, 2010), Halpin (2010), Piccarreta and Billari (2007), Grelet (2002), Rousset et al. (2011) and the TraMineR team (Gabadinho, Studer, Bürgin, ...). SA in social sciences inspired from bioinformatics and other fields (Sankoff and Kruskal, 1983).

What kind of questions may SA answer to?

- 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)?
- How are sequences organized?
  - 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?

What kind of questions may SA answer to?

- 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)?
- How are sequences organized?
  - 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?

What kind of questions may SA answer to?

- 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)?
- How are sequences organized?
  - 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?

What kind of questions may SA answer to?

- 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)?
- How are sequences organized?
  - 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?

What kind of questions may SA answer to?

- 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)?

- How are sequences organized?
  - 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?

What kind of questions may SA answer to?

- 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)?

- How are sequences organized?
  - 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?

What kind of questions may SA answer to?

- 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)?

- How are sequences organized?
  - 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

The data: Swiss occupational trajectories

- Data about Swiss occupational trajectories
  - From the 2002 biographical survey by the Swiss Household Panel http://www.swisspanel.ch
  - 1503 trajectories from 20 to 45 years old (26 years)
  - Same data as in Widmer and Ritschard (2009)

The data: Swiss occupational trajectories

- Data about Swiss occupational trajectories
  - From the 2002 biographical survey by the Swiss Household Panel http://www.swisspanel.ch
  - 1503 trajectories from 20 to 45 years old (26 years)
  - Same data as in Widmer and Ritschard (2009)

The data: Swiss occupational trajectories

- Data about Swiss occupational trajectories
  - From the 2002 biographical survey by the Swiss Household Panel http://www.swisspanel.ch
  - 1503 trajectories from 20 to 45 years old (26 years)
  - Same data as in Widmer and Ritschard (2009)

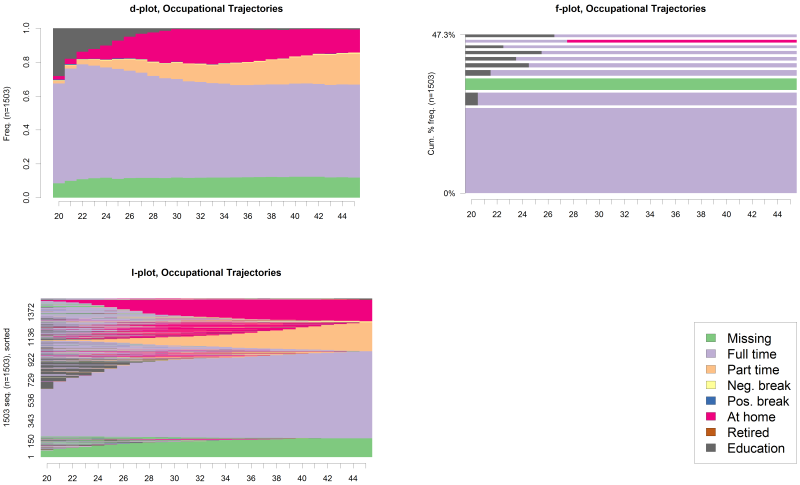
The data: Swiss occupational trajectories

- Data about Swiss occupational trajectories
  - From the 2002 biographical survey by the Swiss Household Panel http://www.swisspanel.ch
  - 1503 trajectories from 20 to 45 years old (26 years)
  - Same data as in Widmer and Ritschard (2009)

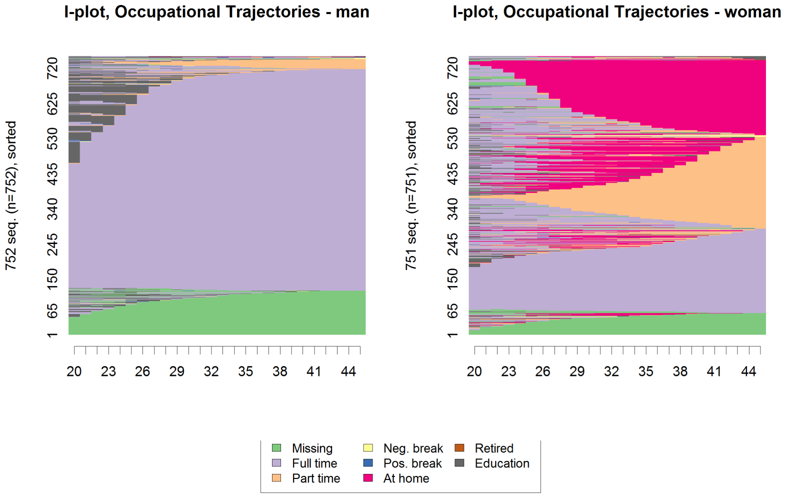
Section outline

2. Some views of Swiss occupational trajectories
   - Rendering transversal and longitudinal characteristics
   - Dissimilarity-based analysis
   - Methods for event sequences

Rendering sequences



Rendering sequences by group (sex)



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	$\cdots$

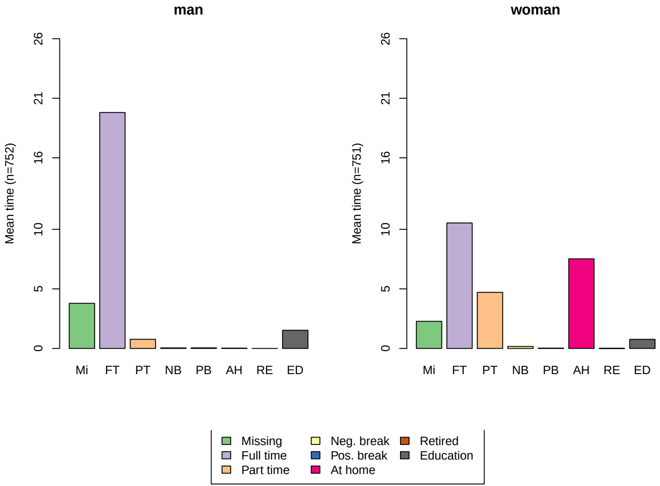
- Summary of longitudinal measures (within entropy, transition rates, mean duration ...)

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, ...

Mean time in each state

```
seqmtplot(seqs.occ, group = seqs$sex)
```

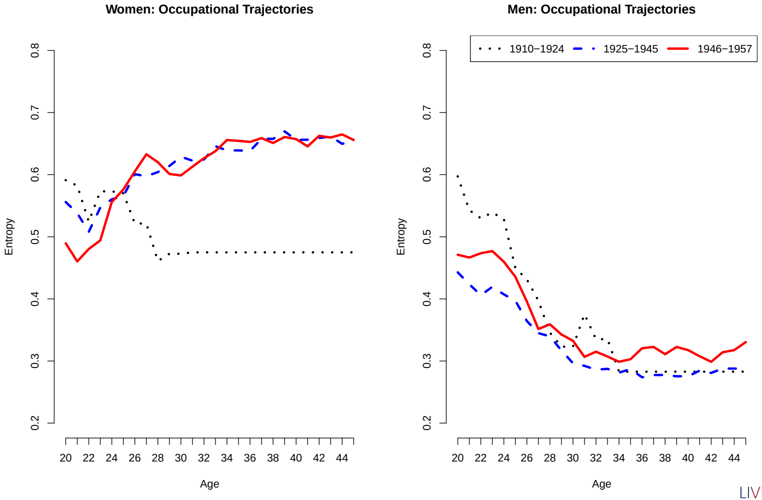


Transition rates

	[-> Mi]	[-> FT]	[-> PT]	[-> NB]	[-> PB]	[-> AH]	[-> RE]	[-> ED]
[Mi ->]	0.969	0.005	0.004	0.001	0.001	0.011	0.000	0.008
[FT ->]	0.003	0.971	0.009	0.001	0.001	0.013	0.000	0.003
[PT ->]	0.005	0.026	0.939	0.001	0.001	0.018	0.000	0.010
[NB ->]	0.040	0.047	0.027	0.880	0.000	0.007	0.000	0.000
[PB ->]	0.105	0.316	0.105	0.000	0.404	0.018	0.000	0.053
[AH ->]	0.003	0.007	0.032	0.000	0.000	0.956	0.000	0.002
[RE ->]	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
[ED ->]	0.044	0.236	0.045	0.001	0.002	0.006	0.000	0.664

Heterogeneity: Sequence of transversal entropies

Occupational, Women vs Men (example from Widmer and Ritschard 2009)



Number of state transitions (longitudinal)

Section outline

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)
  - Measure the discrepancy between sequences
  - Run self-organizing maps (SOM) on sequences
  - MDS scatterplot representation of 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)
  - Measure the discrepancy between sequences
  - Run self-organizing maps (SOM) on sequences
  - MDS scatterplot representation of sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy

Pairwise dissimilarities between sequences

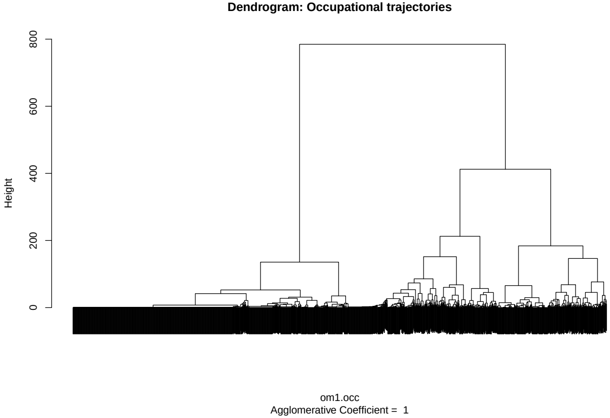
- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Measure the discrepancy between sequences
  -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

Dissimilarity matrix

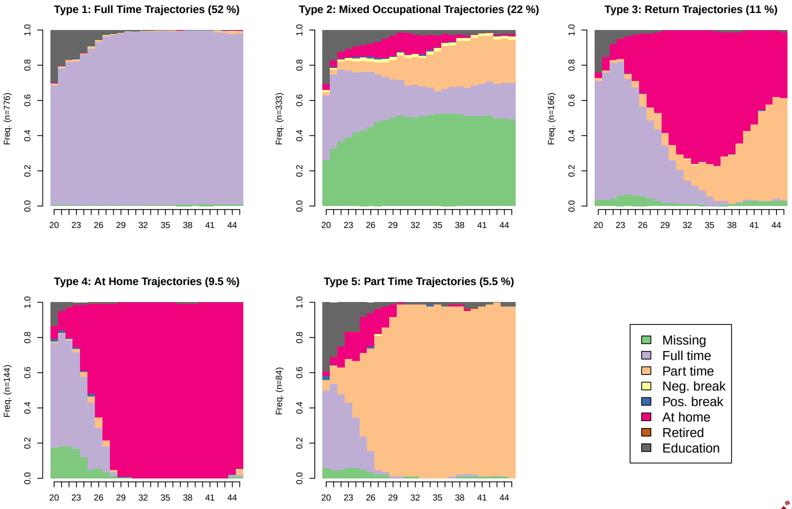
```
print(seqs.occ[1:4, ], format = "SPS")

    Sequence
[1] (FT,26)
[2] (FT,26)
[3] (Mi,6)-(ED,3)-(Mi,17)
[4] (ED,1)-(Mi,3)-(PT,4)-(FT,18)

dm <- seqdist(seqs.occ[1:4, ], method = "LCS")
dm[1:4, 1:4]

     [,1] [,2] [,3] [,4]
[1,]    0    0   52   16
[2,]    0    0   52   16
[3,]   52   52    0   44
[4,]   16   16   44    0
```

Cluster analysis: determining typologies

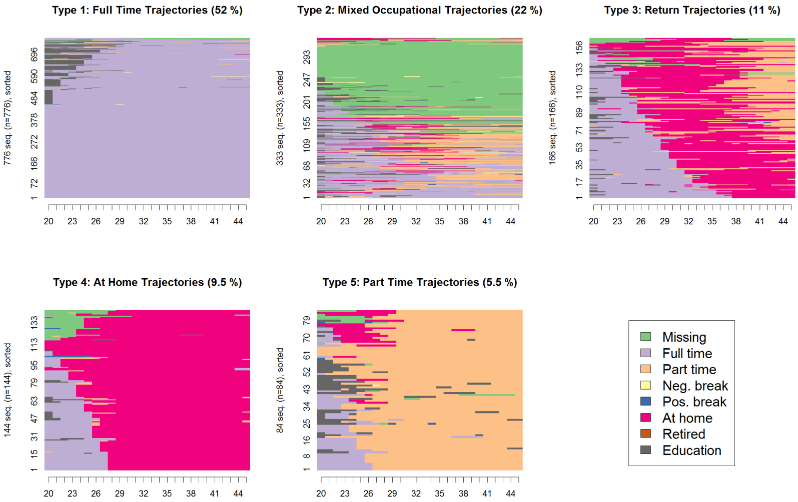
Example from Widmer and Ritschard (2009)



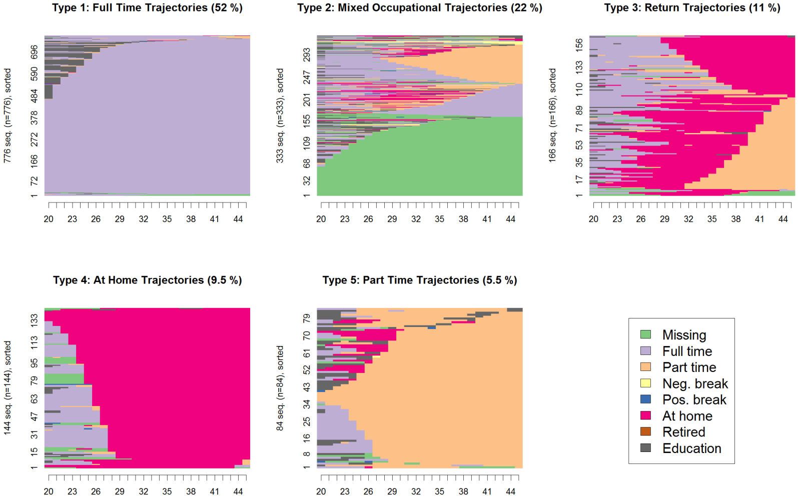
Rendering clusters: d-plots



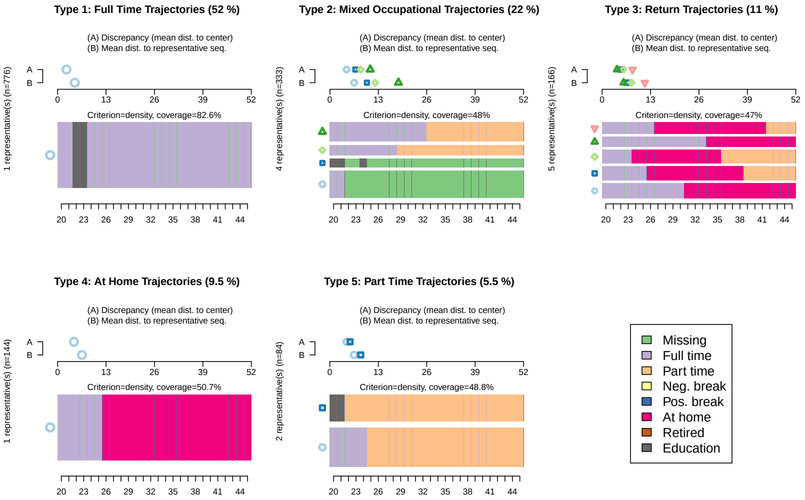
Rendering clusters: i-plots (sorted by 1st MDS factor)



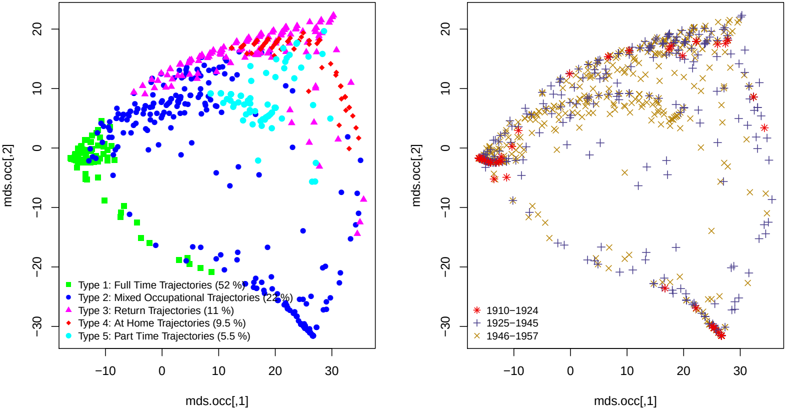
Rendering clusters: i-plots (sorted from end)



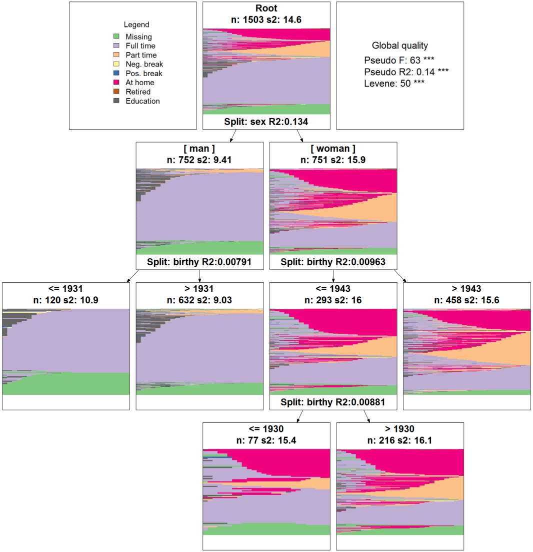
Representative sequences Gabadinho et al. (2011)



MDS: Scatterplot view of sequences



Regression tree (Studer et al., 2011)



Section outline

2. Some views of Swiss occupational trajectories
   - Rendering transversal and longitudinal characteristics
   - Dissimilarity-based analysis
   - Methods for event sequences

Event sequences

- Instead of the successive states, we may consider the transitions between states and more specifically the—possibly simultaneous—events that provoke the transitions.
- Event sequences are more difficult to render because they have no duration!
- Event sequences are of interest for studying the sequencing
  - What are the typical sequencing of life events?
  - Which event sequencing distinguishes men and women? younger and older cohorts?

Event sequences

- Instead of the successive states, we may consider the transitions between states and more specifically the—possibly simultaneous—events that provoke the transitions.

- Event sequences are more difficult to render because they have no duration!

- Event sequences are of interest for studying the sequencing
  - What are the typical sequencing of life events?
  - Which event sequencing distinguishes men and women? younger and older cohorts?

Event sequences

- Instead of the successive states, we may consider the transitions between states and more specifically the—possibly simultaneous—events that provoke the transitions.

- Event sequences are more difficult to render because they have no duration!
- Event sequences are of interest for studying the sequencing
  - What are the typical sequencing of life events?
  - Which event sequencing distinguishes men and women? younger and older cohorts?

Event sequences

- Instead of the successive states, we may consider the transitions between states and more specifically the—possibly simultaneous—events that provoke the transitions.

- Event sequences are more difficult to render because they have no duration!
- Event sequences are of interest for studying the sequencing
  - What are the typical sequencing of life events?
  - Which event sequencing distinguishes men and women? younger and older cohorts?

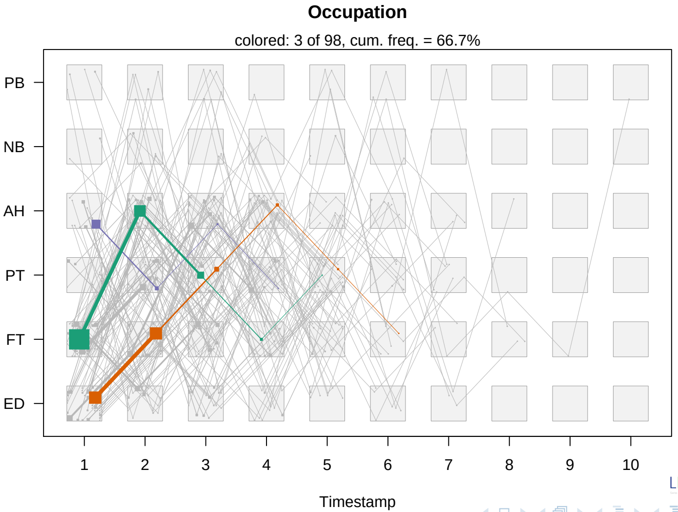
Event sequences

- Instead of the successive states, we may consider the transitions between states and more specifically the—possibly simultaneous—events that provoke the transitions.

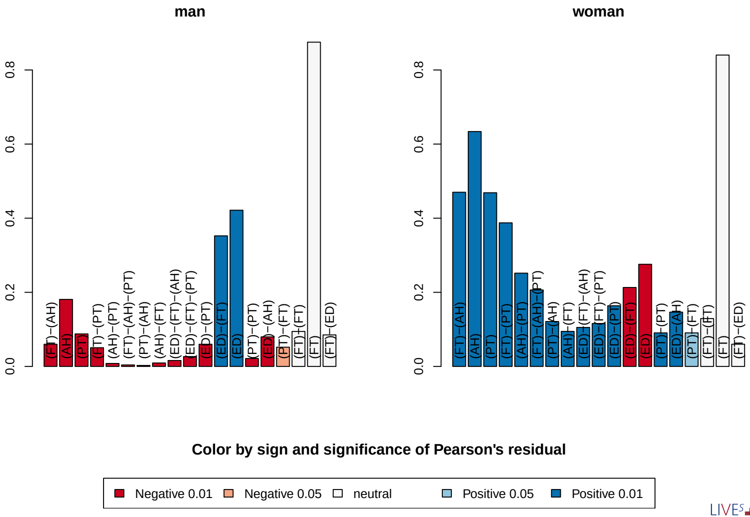
- Event sequences are more difficult to render because they have no duration!
- Event sequences are of interest for studying the sequencing
    - What are the typical sequencing of life events?
    - Which event sequencing distinguishes men and women? younger and older cohorts?

Rendering event sequences

Parallel coordinate plot (Bürgin and Ritschard, 2012)



Event sequences: discriminating sub-sequences



Outline

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

Outline

Conclusion: Limits of sequence analysis

- By focusing on complete trajectories until 45 years => we ignore recent generations
- Most recent birth year is 1957 ($2002 - 45$)
- Other issues:
  - Granularity: year, month, day, ...
  - State definition: should we distinguish {separated, divorced, widowed} or consider a single state? works by Raffaella Piccaretta

Conclusion: Limits of sequence analysis

- By focusing on complete trajectories until 45 years
  => we ignore recent generations
- Most recent birth year is 1957 (2002 – 45)
- Other issues:
  - Granularity: year, month, day, ...
  - State definition: should we distinguish {separated, divorced, widowed} or consider a single state? works by Raffaella Piccaretta

SA Scalability

- SA main bottleneck is dimension of dissimilarity matrix.
  - Dimension depends on number of cases (should not exceed about 10000).
  - Solution: work on a representative sample of the sequences
- Other limitations are
  - Size of alphabet (should be less than 20), especially for graphical rendering, but also computation time.
    - Solution: aggregate elements of alphabet.
  - Sequence length (< 200), affects computation time
    - Solution: change time granularity

SA Scalability

- SA main bottleneck is dimension of dissimilarity matrix.
  - Dimension depends on number of cases (should not exceed about 10000).
  - Solution: work on a representative sample of the sequences
- Other limitations are
  - Size of alphabet (should be less than 20), especially for graphical rendering, but also computation time.
    - Solution: aggregate elements of alphabet
  - Sequence length (< 200), affects computation time
    - Solution: change time granularity

SA Scalability

- SA main bottleneck is dimension of dissimilarity matrix.
  - Dimension depends on number of cases (should not exceed about 10000).
  - Solution: work on a representative sample of the sequences
- Other limitations are
  - Size of alphabet (should be less than 20), especially for graphical rendering, but also computation time.
    - ⇒ Solution: aggregate elements of alphabet
  - Sequence length (< 200), affects computation time
    - ⇒ Solution: change time granularity

SA Scalability

- SA main bottleneck is dimension of dissimilarity matrix.
  - Dimension depends on number of cases (should not exceed about 10000).
  - Solution: work on a representative sample of the sequences
- Other limitations are
  - Size of alphabet (should be less than 20), especially for graphical rendering, but also computation time.
    - Solution: aggregate elements of alphabet.
  - Sequence length (< 200), affects computation time
    - Solution: change time granularity

SA Scalability

- SA main bottleneck is dimension of dissimilarity matrix.
  - Dimension depends on number of cases (should not exceed about 10000).
  - Solution: work on a representative sample of the sequences
- Other limitations are
  - Size of alphabet (should be less than 20), especially for graphical rendering, but also computation time.
    - Solution: aggregate elements of alphabet.
  - Sequence length ($<$ 200), affects computation time
    - Solution: change time granularity

SA Scalability

- SA main bottleneck is dimension of dissimilarity matrix.
  - Dimension depends on number of cases (should not exceed about 10000).
  - Solution: work on a representative sample of the sequences
- Other limitations are
  - Size of alphabet (should be less than 20), especially for graphical rendering, but also computation time.
    - Solution: aggregate elements of alphabet.
  - Sequence length (< 200), affects computation time
    - Solution: change time granularity

SA Scalability

- SA main bottleneck is dimension of dissimilarity matrix.
  - Dimension depends on number of cases (should not exceed about 10000).
  - Solution: work on a representative sample of the sequences
- Other limitations are
  - Size of alphabet (should be less than 20), especially for graphical rendering, but also computation time.
    - Solution: aggregate elements of alphabet.
  - Sequence length (< 200), affects computation time
    - Solution: change time granularity

SA Scalability

- SA main bottleneck is dimension of dissimilarity matrix.
  - Dimension depends on number of cases (should not exceed about 10000).
  - Solution: work on a representative sample of the sequences
- Other limitations are
  - Size of alphabet (should be less than 20), especially for graphical rendering, but also computation time.
    - Solution: aggregate elements of alphabet.
  - Sequence length ($<$ 200), affects computation time
    - Solution: change time granularity

Work in progress? ...

- SA is mainly static analysis of sequences
- Analysis of generating processes (Alexis Gabadinho's thesis)
    - Markov-Chain (MC) models
    - Variable length MC or probabilistic suffix trees (PST)
        - Model of the generating process
        - Likelihood of each sequence for a given model
        - Likelihood-based mining of typical and rare sequences
        - Testing divergence between groups with nested stratified models
- Modeling time evolution of tendencies in sequences (mixed effect model trees of longitudinal ordinal data) (Reto Bürgin's thesis)

Thank you!

References I

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

Abbott, A. and J. Forrest (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

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.

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

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

Bürgin, R. and G. Ritschard (2012). Categorical parallel coordinate plot. In *LaCOSA Lausanne Conference On Sequence Analysis, University of Lausanne, June 6th-8th 2012*, Lausanne. Poster.

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

References II

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

Elzinga, C. H. (2010). Complexity of categorical time series. *Sociological Methods & Research 38*(3), 463–481.

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.

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, and N. S. Müller (2011). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, and 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.

References III

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.

Piccarreta, R. and F. C. Billari (2007). Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*(4), 1061–1078.

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.

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.

References IV

Sankoff, D. and J. B. Kruskal (Eds.) (1983). *Time Warps, String Edits, and Macro-Molecules: The Theory and Practice of Sequence Comparison*. Reading: Addison-Wesley.

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

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.

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 (2007)
  - Various metrics, including original ones based on non-aligning methods
  - Turbulence