HUMBOLDT-UNIVERSITÄT ZU BERLIN

UNIVERSITÀ DEGLI STUDI DI MILANO

Sequence Analysis for Social Science

Anette Fasang and Emanuela Struffolino

PART 1

Population Dynamics and Health Program (PDHP)
University of Michigan
Feb. 9th, 2022

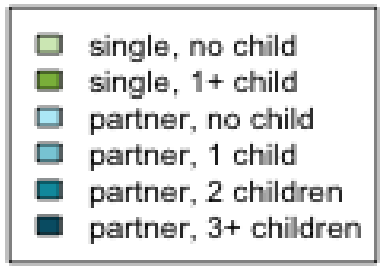
Outlook part 1

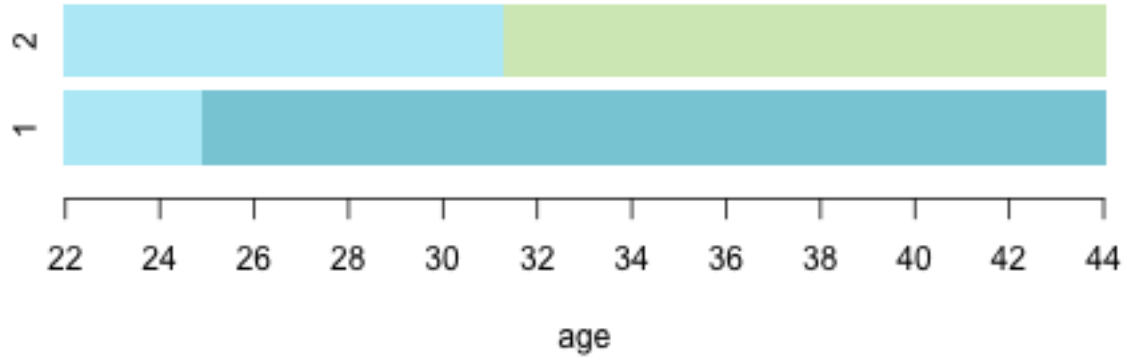
1. Sequence analysis in context: theory and methods

2. Basic Sequence Terminology

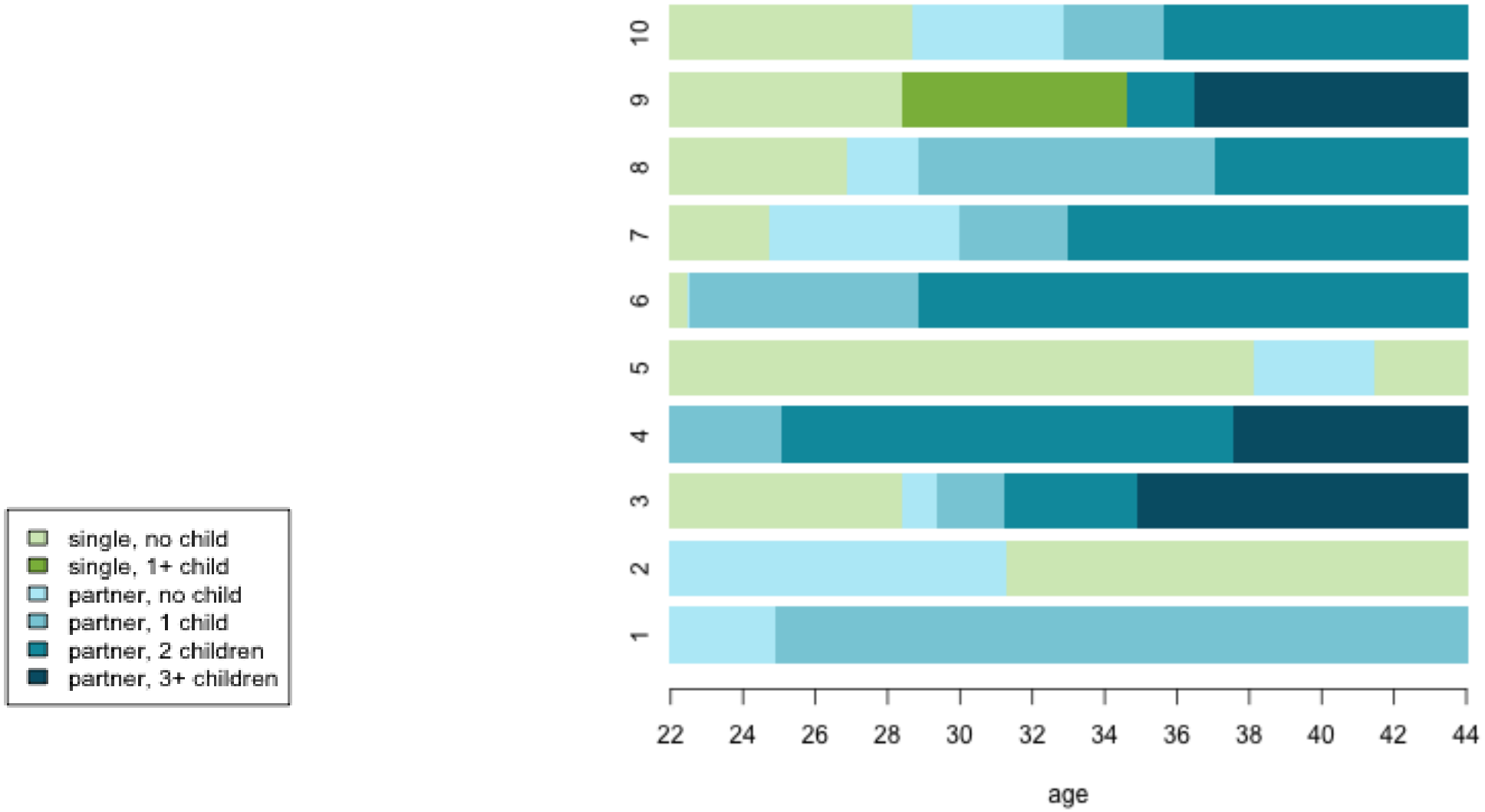
1a. Sequence analysis in context: theory

Sequences in the Social Sciences
Example: family formation



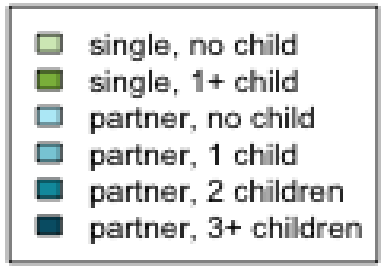


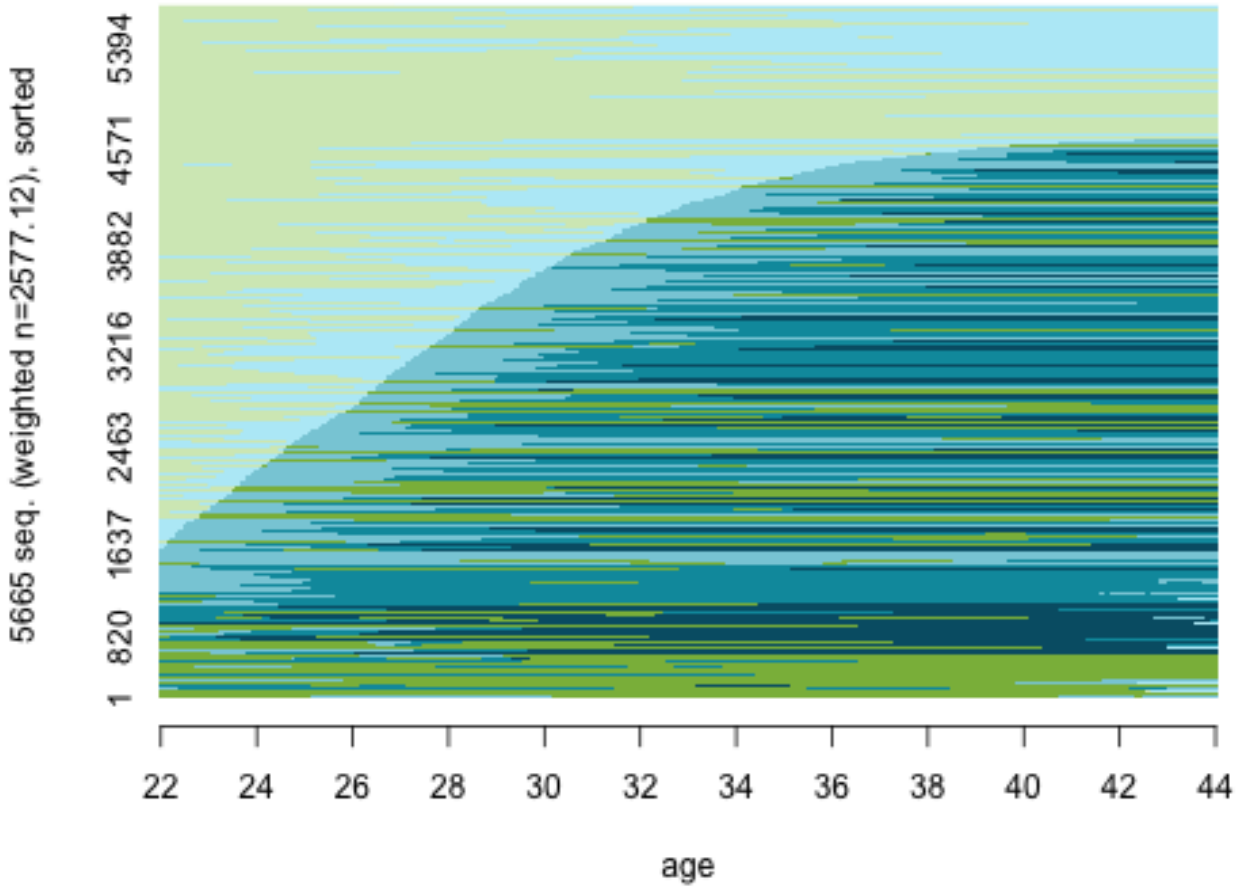
Family formation sequences



Family formation sequences

NLSY 79, women born 1957-1964

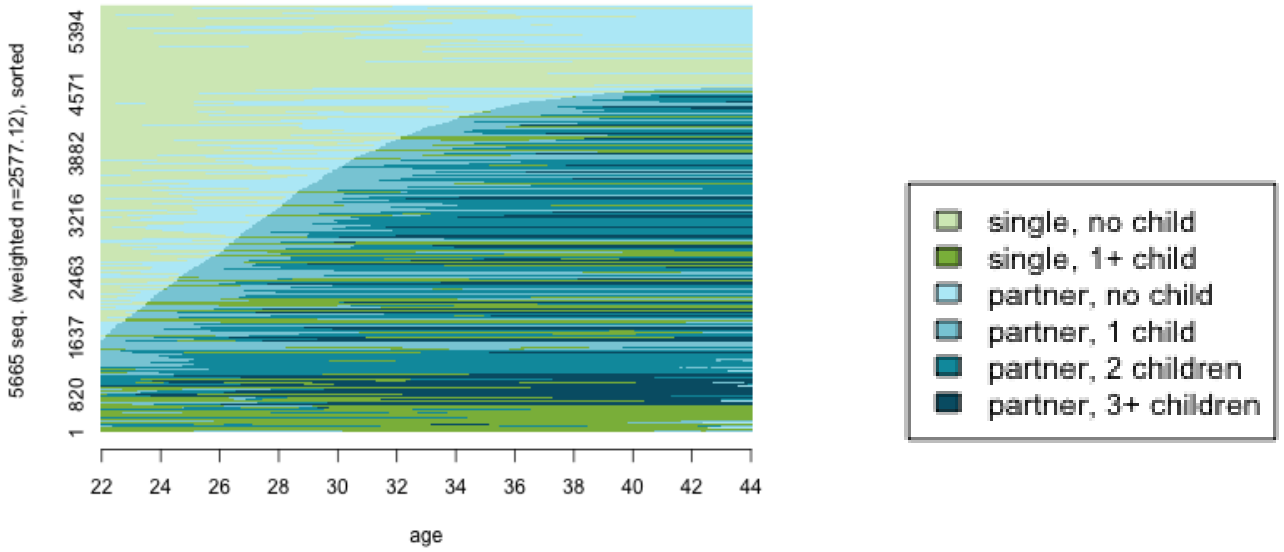


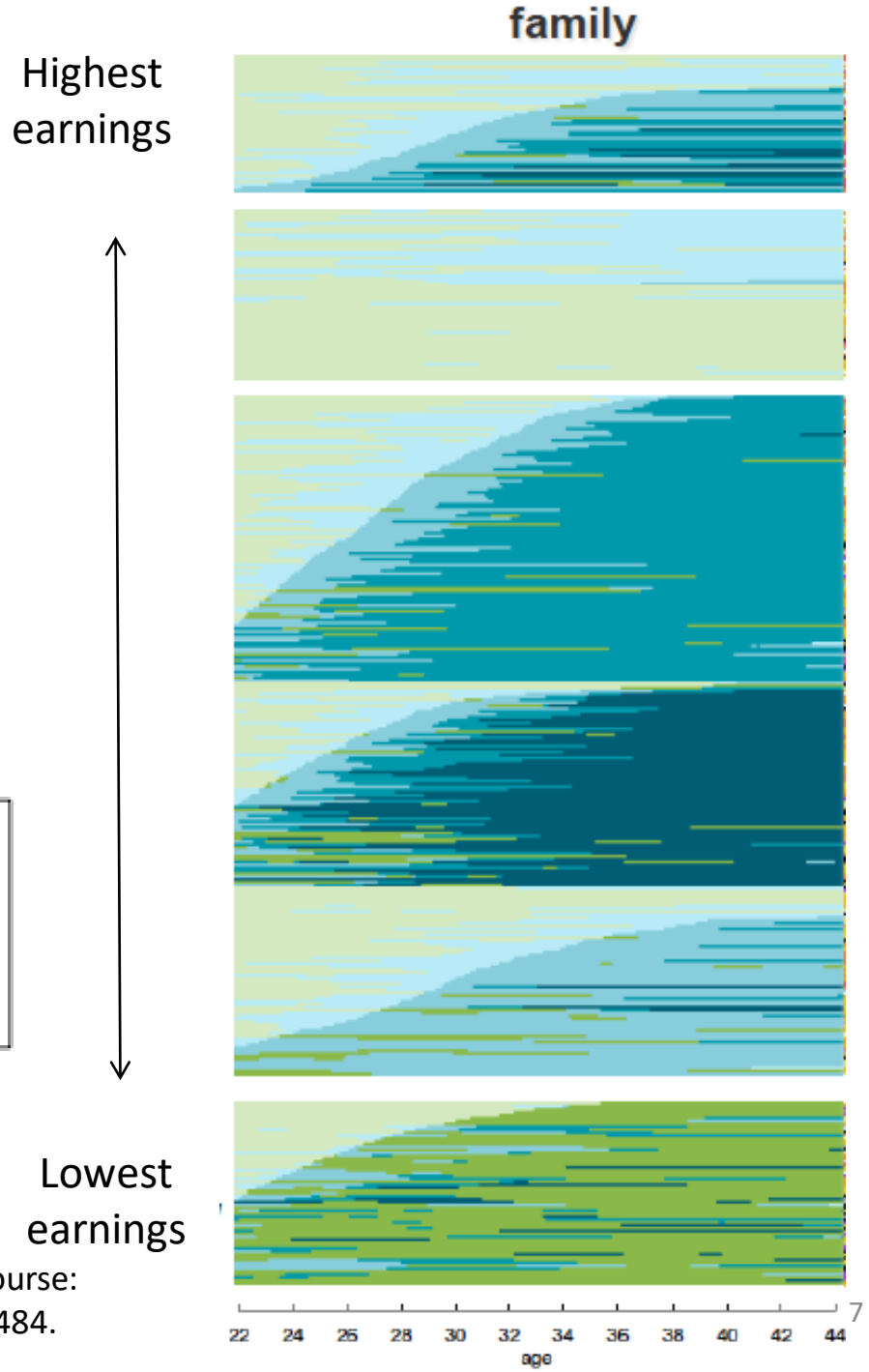


Sequences sorted by age of first birth

Family formation sequences

Goal:
group individuals with similar trajectories („ideal types“) to assess determinants and correlates of family formation processes





Aisenbrey, S., & Fasang, A. (2017). The interplay of work and family trajectories over the life course: Germany and the United States in comparison. American Journal of Sociology, 122(5), 1448-1484.

Sequence analysis in context: theory

"Time Matters": Process / mechanisms as the fundamental building blocks of sociological analysis (*Processual Sociology*, Abbott, 2016):

"… social reality happens in sequences of actions located within constraining or enabling structures" (Abbott 1992)

"Narrative Positivism" → process-based

"General Linear Model" → variable-based

Processual Social Science

"By a processual approach, I mean an approach that presumes that everything in the social world is continuously in the process of making, remaking and unmaking itself (and other things), instant by instant." (p. x)

"The world of the processual approach is a world of events. Individuals and social entities are not the elements of social life, but are patterns and regularities defined on lineages of successive events."

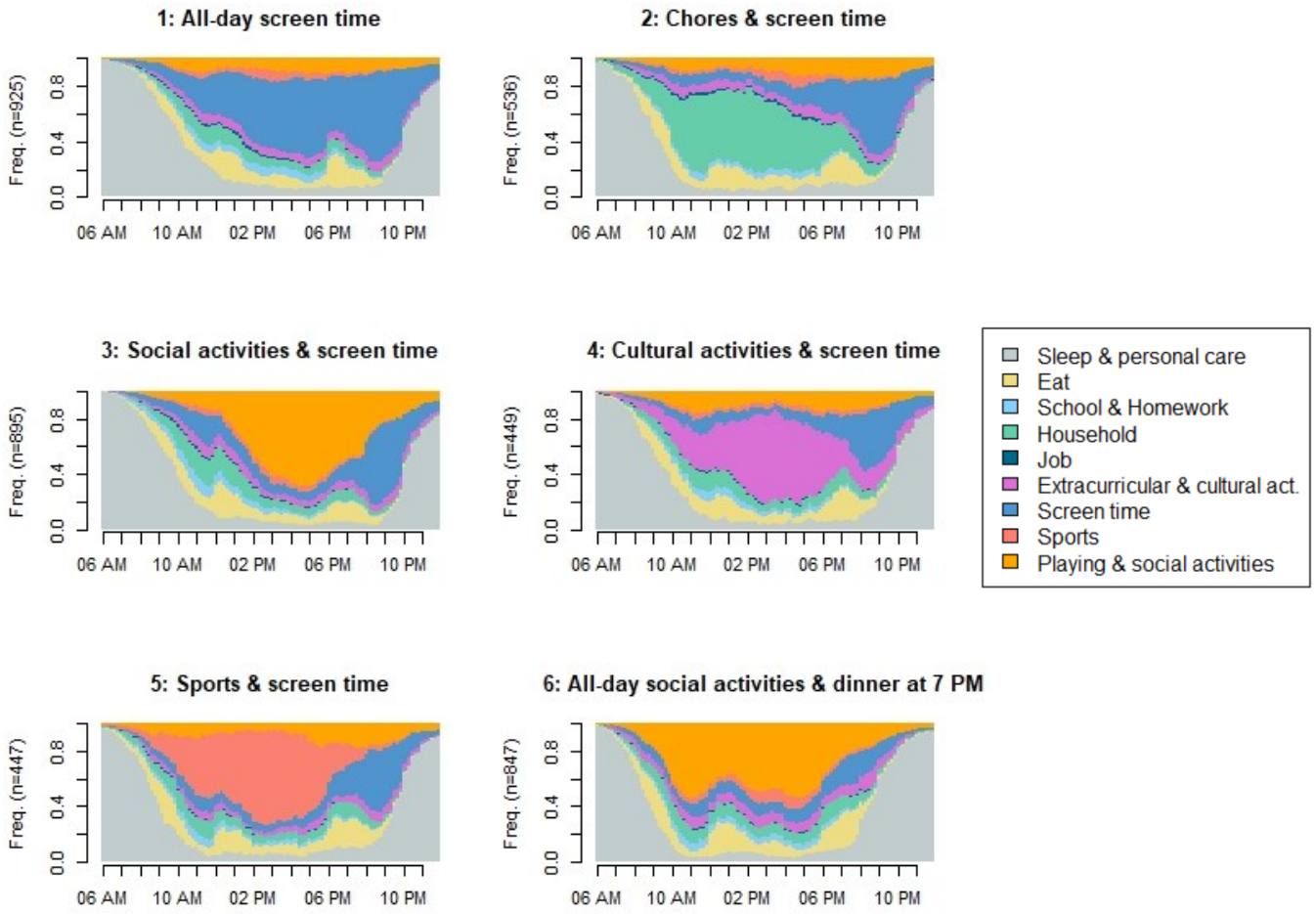
"[…] **process outcomes** are long run stabilities established by myriads of individual events." (p. 176) "[…] it is the whole walk that is the outcome."

Andrew Abbott (2016) "*Processual Sociology*"

Most common fields of application

- Life course sociology
- (Family) demography
- Labor market and career research
- Social stratification
- Aging and retirement
- Welfare state and social policy
- Time use

But: many different applications: regional patterns of lynching, holocaust survivors, cultural sociology, regime changes in political science/history, Jaina monks in India, …



Life Course Paradigm (Elder et al. 2003, Mayer 2009)

- Development as a life long process
- Time: Timing and sequencing of life course processes matter for their correlates and consequences
- Place: Macro-structural contexts/social policies shape life course processes
- Life course norms about the appropriate timing and sequencing of events
- Cumulative Advantage and Dis-advantage (Dannefer 1987, DiPriete and Eirich 2006)

Core theoretical concepts

- Timing and sequencing/order of events
- Same process – different speed? (Fasang & Raab 2014)
- Zig-zag processes / back and forth movements
- Instability, volatility, precarity
- Cumulative advantage/disadvantage (CAD)
- Path-dependency
- “Turning points”
- “Ideal types” of processes
- ……

1b.Sequence analysis in context: methods

Aims of longitudinal data analysis

- **Event history analysis**
  Whether and/or when do events occur?
- **Panel and growth curve/group-based trajectory models**
  How does one outcome change over time (metric/binary)?

Variable-based

- **Sequence analysis**
  How do processes of categorical states develop over time?
  Analyzing trajectories/stories as a whole

Process-based

Sequence analysis in context: methods

	Event history analysis	**Panel Regression**	**Sequence analysis**
Theoretical concept	transition, duration	change (binary/metric)	trajectory (categorical)
Scientific tradition			
Assumption about data generation			
Objective to identify…..			
Broader theoretical assumptions			

Sequence analysis in context: methods

	Event history analysis	**Panel Regression**	**Sequence analysis**
Theoretical concept	transition, duration	change (binary/metric)	trajectory (categorical)
Scientific tradition	**stochastic** data modeling tradition		narrative positivism, algorithmic tradition
Assumption about data generation			
Objective to identify…..			
Broader theoretical assumptions			

Sequence analysis in context: methods

	Event history analysis	**Panel Regression**	**Sequence analysis**
Theoretical concept	transition, duration	change (binary/metric)	trajectory (categorical)
Scientific tradition	**stochastic** data modeling tradition		narrative positivism, algorithmic tradition
Assumption about data generation	stochastic process, **causality**		**none** **black box**
Objective to identify…..			
Broader theoretical assumptions			

Sequence analysis in context: methods

	Event history analysis	**Panel Regression**	**Sequence analysis**
Theoretical concept	transition, duration	change (binary/metric)	trajectory (categorical)
Scientific tradition	**stochastic** data modeling tradition		narrative positivism, algorithmic tradition
Assumption about data generation	stochastic process, **causality**		**none** **black box**
Objective to identify…..	timing of single transitions/durations	probability /change	patterns of sequential equivalence
Broader theoretical assumptions			

Sequence analysis in context: methods

	Event history analysis	**Panel Regression**	**Sequence analysis**
Theoretical concept	transition, duration	change (binary/metric)	trajectory (categorical)
Scientific tradition	**stochastic** data modeling tradition		narrative positivism, algorithmic tradition
Assumption about data generation	stochastic process, **causality**		**none** **black box**
Objective to identify…..	timing of single transitions/durations	probability /change	patterns of sequential equivalence
Broader theoretical assumptions	Structure → occurrence of / change in events		Structure + agency interact across process → sequence

→ **Complementary methods for different questions**

Sequence analysis

- Sequences **are empirically observed traces of temporally ordered events**
- **Goal:** Analyze regularities in ***categorical sequences***

- ***Categorical sequences*** in the social sciences consist of sequentially linked categorical states that make up a social process (‘**qualitative**’ states)
- → Sequence analysis is at the intersection of qualitative and quantitative methodology: it is the “**quantitative**” analysis of sequences of “**qualitative**” states

Examples of categorical state sequences

- **employment careers** (Widmer and Ritschard, 2009): full time employed, unemployed, on family leave, permanent or fixed term employment
- **pathways to adulthood** (Bras, Liefbroer, and Elzinga, 2010): living with parents, living alone, living with spouse, completing education, internship, employment
- **family formation processes** (Elzinga and Liefborer, 2007): single, married, cohabiting, divorced
- **time use** (Lesnard 2006): sleep, eat, work, leisure, child care, exercise over the course of a day

2. Basic terminology: What is a sequence?

Basic terminology: What is a sequence?

- Sequence: ordered lists of a discrete set of elements
- The set of elements constituting a sequence is called a state space or alphabet A
- A sequence x of length k can be written as

$$x = x_1, x_2, \ldots, x_k \text{ with } x_i \in A$$

- Most social science applications examine recurrent sequences, which allow for repeated occurrence of the same states

Episodes & Transitions

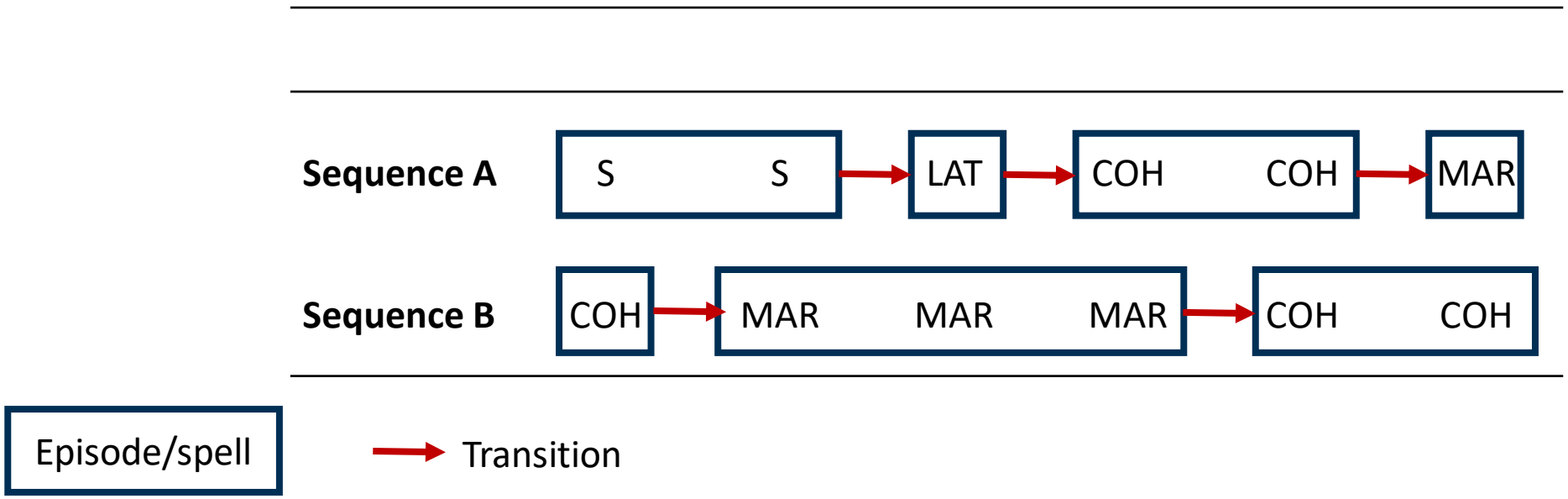
- A single state observed in isolation or series of consecutively repeated states constitute **episodes** or **spells**
- Change of state = **transition** = beginning of a new episode

Sequence A	S	S	LAT	COH	COH	MAR
Sequence B	COH	MAR	MAR	MAR	COH	COH

S= Single, LAT = Living Apart Together (Dating), COH = Cohabiting, MAR = Married

Episodes & Transitions

- A single state observed in isolation or series of consecutively repeated states constitute **episodes** or **spells**
- Change of state = **transition** = beginning of a new episode



Notation – **St**ate **S**equence Format (STS)

Example: Partnership sequence of length *k*=264

LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-S-S-S-S-S-S-LAT-LAT-LAT-LAT- LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LA

T-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT- LAT-LAT-LAT-LAT-LAT-LAT-LAT- LAT-LAT-LAT-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-LAT-

LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LA

T-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-LAT-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-LAT-LAT-LAT-LAT-LAT-L

AT-LAT-LAT-LAT-LAT-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-COH-MAR-MAR-MAR-MAR-MAR-MAR-M

AR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-

MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR

-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MA

R-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR-MAR

Notation – **S**tate **P**ermanence **S**equences (SPS)

Example: Partnership sequence of length k=264

(LAT,13)-(S,6)-(LAT,33)-(S,24)-(LAT,41)-(S,35)-(LAT,10)-(COH,14)-(MAR,88)

Notation – **D**istinct **S**uccessive State Format (DSS)

Example: Partnership sequence of length *k*=264

LAT-S-LAT-S-LAT-S-LAT-COH-MAR

Defining sequences – the alphabet

Substantive and methodological considerations:

- Balance parsimony and detail (usually between 5-10 states)
- Infrequent and/or substantively irrelevant categories can often be combined
- Don't lose sight of your research question
- Start big and reduce
- Large alphabets lead to less stable cluster analysis results, especially when case numbers are small
- Concurrent/overlapping states can be their own category
- Specification of alphabet is very consequential for results (like – which variables do I include in my model?)

Sequence length and time intervals

What defines the starting point of the sequences?

- Process-time: Age or a specific transition, start of process
- Calendar-time: fixed time point
- Comparing individuals at different ages can be problematic
- Time intervals should be calibrated to temporal variation in the sequences: combine to larger units (months, quarters, annual), when not much change is happening, reduce to smaller units otherwise

Sequence length and time intervals

- Alphabet size and sequence length define the number of possible sequence realizations and affect the stability of typologies
- Shorter sequence length can reduce complexity and computation time
- Specification of sequence depends on research question (and available data)

		Sequence length	
		5	**10**
No. of elements in the alphabet	**5**	5^5	5^{10}
		3,125	9,765,625
	10	10^5	10^{10}
		100,000	10,000,000,000

Unequal sequence length and missings

- Most SA applications work with sequences of equal length
- Common strategies:
  - Complete case analysis
  - Add additional missing state to alphabet
  - Replace gaps with valid values using some sort of „imputation“

Things to consider

Computing load:

- Bottleneck: computation of the pairwise dissimilarity matrix
- Samples of up to 10,000 cases and sequences with a length of up to few hundreds seem to be feasible (but it takes some time…)
- Size of the alphabet only has a modest impact; however: visualization suffers from too large alphabets (alphabet length preferably below 10)

Weights:

- Most of the TraMineR functions allow for using weights
  if weights are assigned when defining the sequence object (seqdef-function) they are automatically passed to other TraMineR-functions (e.g., seqplot)
- Weights could also be used to reduce the computing loads if they are used to get rid of doublets in the data

Strengths of sequence analysis

- pattern search of complex processes over time
- holistic perspective on sequences as a whole
- complementary to regression based methods, often useful in combination with other methods

HUMBOLDT-UNIVERSITÄT ZU BERLIN

UNIVERSITÀ DEGLI STUDI DI MILANO

Sequence Analysis for Social Science

Anette Fasang and Emanuela Struffolino

PART 2

Population Dynamics and Health Program (PDHP)
University of Michigan
Feb. 9th, 2022

Outlook part 2

1. Introduction of the dataset

2. Defining and describing sequence data

3. Visualization of sequence

1. Introduction to the dataset

Pairfam

The data for examples come from the German Family Panel (pairfam), release 10.0 Brüderl et al. (2019). A description of the study can be found in Huinink et al. (2011).

We gratefully acknowledge the permission of the pairfam team to share a reduced version of their data to illustrate all techniques presented in the book with real-world survey data.

If you are interested in using the complete data sets please turn to: https://www.pairfam.de/en/

Data format

- TraMineR prefers sequences stored in wide format

id	state1	state2	state3	state4	state5	state6	state7
1	Single	Single	Single	LAT	LAT	LAT	MAR
2	LAT	LAT	Cohab	Cohab	Single	LAT	Cohab

- Regular panel data can be easily prepared for sequence analysis
- If correctly formatted TraMineR can handle episode data as well

Data format

We will use:

- Family formation trajectories
- Yearly data from age 18 to 40
- Sequences of equal length (see backup slides at the end of this pdf for

2. Defining and describing sequence data

Sequences in TraMineR

- TraMineR can handle different input data formats (see: `seqformat`)
- First step of SA in R: `seqdef`-function
  - Most important input: sequence variables
  - Allows to specify:
    - Long and short labels
    - Colors and other settings for plots
    - Weights
    - …
- Alphabet (`seqrecode`) and granularity (`TraMineRextras::seqgranularity`) can also be changed after sequence object has been defined

Time spent in different states & occurrence of episodes

	Time spent in state x in months			*Number of episodes*	
State	*Mean*	*SD*	*Rel. freq.*	*Mean*	*SD*
S	72.5	69.8	0.27	1.6	1.2
LAT	48.0	43.9	0.18	1.8	1.3
COH	48.6	53.3	0.18	1.0	0.8
MAR	95.0	78.9	0.36	0.8	0.5

Time spent in different states & occurrence of episodes

	Time spent in state x in months			*Number of episodes*	
State	*Mean*	*SD*	*Rel. freq.*	*Mean*	*SD*
S	72.5	69.8	0.27	1.6	1.2
LAT	48.0	43.9	0.18	1.8	1.3
COH	48.6	53.3	0.18	1.0	0.8
MAR	95.0	78.9	0.36	0.8	0.5

```
seqmeant(partner.month.seq, serr = TRUE)
```

```
seqmeant(partner.month.seq, prop = TRUE)
```

```
seqmeant(seqdss(partner.month.seq),serr = TRUE)
```

Number of transitions

Granularity	*Mean*	*SD*
Monthly data	5	2.6
Yearly data	4	1.9

```
seqtransn(family.month.seq)
seqtransn(family.year.seq)
```

Transition rates

Transition matrix of sequences stored in *STS* format

	State at							
	Monthly granularity				*Yearly granularity*			
State at	S	LAT	COH	MAR	S	LAT	COH	MAR
S	0.98	0.02	0.00	0.00	0.81	0.14	0.04	0.01
LAT	0.02	0.96	0.02	0.00	0.12	0.68	0.16	0.04
COH	0.00	0.00	0.98	0.01	0.04	0.02	0.80	0.14
MAR	0.00	0.00	0.00	1.00	0.01	0.01	0.00	0.98

Transition matrix of sequences stored in *DSS* format

	State at			
State at	S	LAT	COH	MAR
S	0.00	0.91	0.07	0.02
LAT	0.42	0.00	0.50	0.08
COH	0.20	0.12	0.00	0.68
MAR	0.44	0.46	0.11	0.00

- Transition rates for monthly data rather useless
- Transition rates for yearly data produce more interesting results

- Using DSS instead of STS sequences allows for a focused view on transitions (by definition stability is ignored; main diagonal = 0)

Transition rates

Transition matrix of sequences stored in *STS* format

	State at							
	Monthly granularity				*Yearly granularity*			
State at	S	LAT	COH	MAR	S	LAT	COH	MAR
S	0.98	0.02	0.00	0.00	0.81	0.14	0.04	0.01
LAT	0.02	0.96	0.02	0.00	0.12	0.68	0.16	0.04
COH	0.00	0.00	0.98	0.01	0.04	0.02	0.80	0.14
MAR	0.00	0.00	0.00	1.00	0.01	0.01	0.00	0.98

```
seqtrate(family.month.seq)
seqtrate(family.year.seq)
```

Transition matrix of sequences stored in *DSS* format

	State at			
State at	S	LAT	COH	MAR
S	0.00	0.91	0.07	0.02
LAT	0.42	0.00	0.50	0.08
COH	0.20	0.12	0.00	0.68
MAR	0.44	0.46	0.11	0.00

```
seqtrate(seqdss(family.year.seq))
```

Modal and representative sequences

- Modal sequence:
  Sequence composed of the most prevalent states at each position of the sequence; usually not observed in the data

Granularity	*Modal Sequence*
Monthly data	(S,102)-(MAR,162)
Yearly data	(S,9)-(MAR,13)

```
modal.month.seq <- seqdef(as_tibble(seqmodst(partner.month.seq)))
print(modal.month,seq, format = "SPS")

modal.year.seq <- seqdef(as_tibble(seqmodst(partner.year.seq)))
print(modal.year.seq, format = "SPS")
```

Modal and representative sequences

Set of representative sequences

Sequence	*Coverage*	*Assigned*
(S,1)-(LAT,2)-(MAR,19)	5.7	6.5
(S,20)-(MAR,2)	4.4	25.2
(S,4)-(LAT,1)-(COH,1)-(MAR,16)	3.8	5.3
(LAT,3)-(COH,2)-(MAR,17)	3.1	11.4
(S,2)-(LAT,2)-(COH,3)-(MAR,15)	2.7	17.1
(S,5)-(LAT,2)-(COH,2)-(MAR,13)	2.7	23.5
(COH,2)-(MAR,20)	2.6	3.0
(S,1)-(LAT,5)-(MAR,16)	2.3	8.0
Total Coverage	27.5	100.0

- Representative sequences
  - Observed sequences that represent the data best
  - Require a dissimilarity matrix
  - Different criterions can be used to extract representatives
  - Example on the left:
    - neighborhood density criterion: sequences are considered neighbors if their pairwise dissimilarity falls below a predefined threshold (10% of the maximum possible distance value)
    - subset of nonredundant representative sequences with a total coverage of at least 25%

Modal and representative sequences

```
partner.year.om <- seqdist(partner.year.seq, method="OM", sm="CONSTANT")
partner.year.rep <- seqrep(partner.year.seq, diss=partner.year.om, criterion="density")
summary(partner.year.rep)
```

Set of representative sequences

Sequence	*Coverage*	*Assigned*
(S,1)-(LAT,2)-(MAR,19)	5.7	6.5
(S,20)-(MAR,2)	4.4	25.2
(S,4)-(LAT,1)-(COH,1)-(MAR,16)	3.8	5.3
(LAT,3)-(COH,2)-(MAR,17)	3.1	11.4
(S,2)-(LAT,2)-(COH,3)-(MAR,15)	2.7	17.1
(S,5)-(LAT,2)-(COH,2)-(MAR,13)	2.7	23.5
(COH,2)-(MAR,20)	2.6	3.0
(S,1)-(LAT,5)-(MAR,16)	2.3	8.0
Total Coverage	27.5	100.0

3. Visualization of sequences

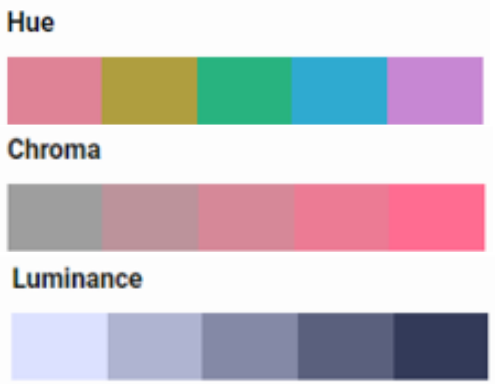
Types of graphs

Tabular vs. graphical inspection

- The complexity of sequence data call for a graphical inspection
- Thorough tabular representations of the data run the risk of presenting an overwhelming amount of information, that might hamper the recognition of regularities
- Even visualizations might be overwhelming
- Two groups of graphs:
  - Data summarization graphs:
    aggregate and summarize the information stored in the sequences
  - Data representation graphs
    render individual sequences instead of aggregate summary measures

Color palettes

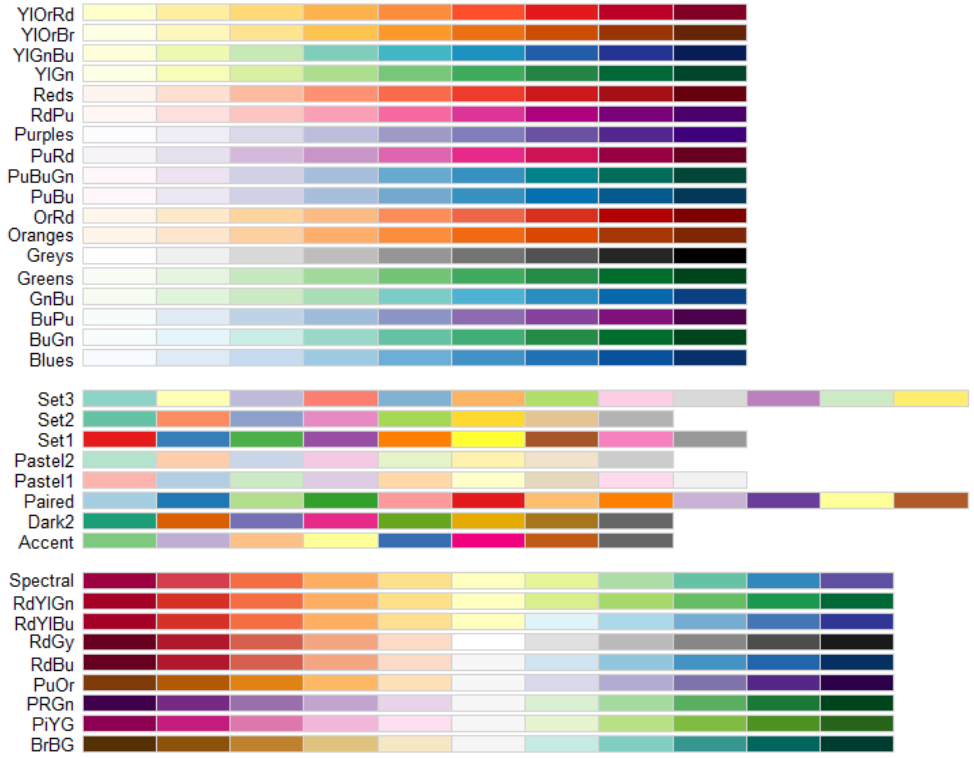
- If colors are not specified TraMineR uses palettes from `RColorBrewer` package
- References for on choosing colors: Zeileis et al. (2009) and Zeileis et al. (2019)
- The R community provides a variety of packages for choosing suitable color palettes (`RColorBrewer, colorspace`)
- Other interesting resources:
  - https://colorbrewer2.org/
  - https://github.com/EmilHvitfeldt/r-color-palettes and https://github.com/EmilHvitfeldt/paletteer



Pre-defined Color palettes

RColorBrewer

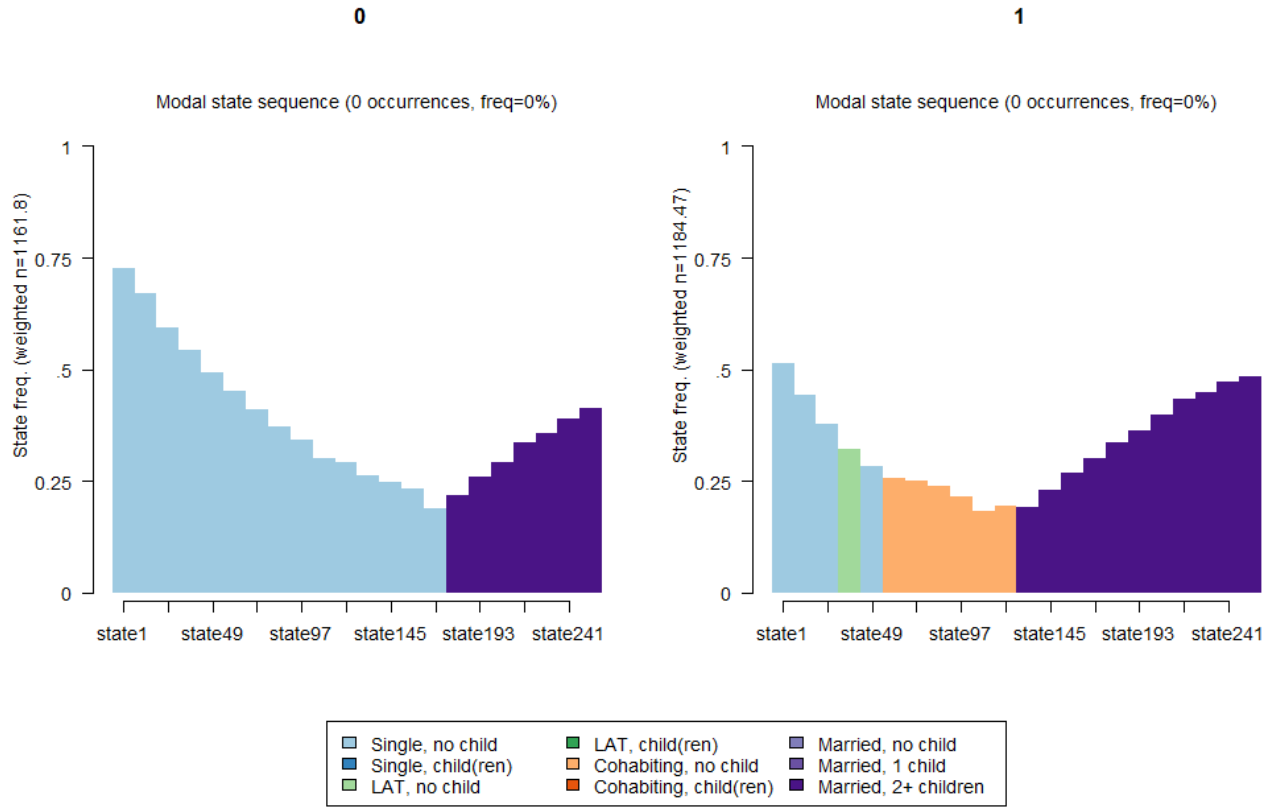
```
display.brewer.all()
```

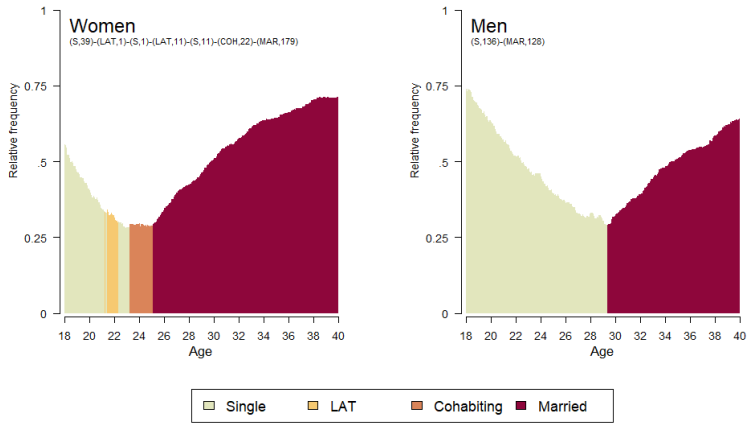


colorspace

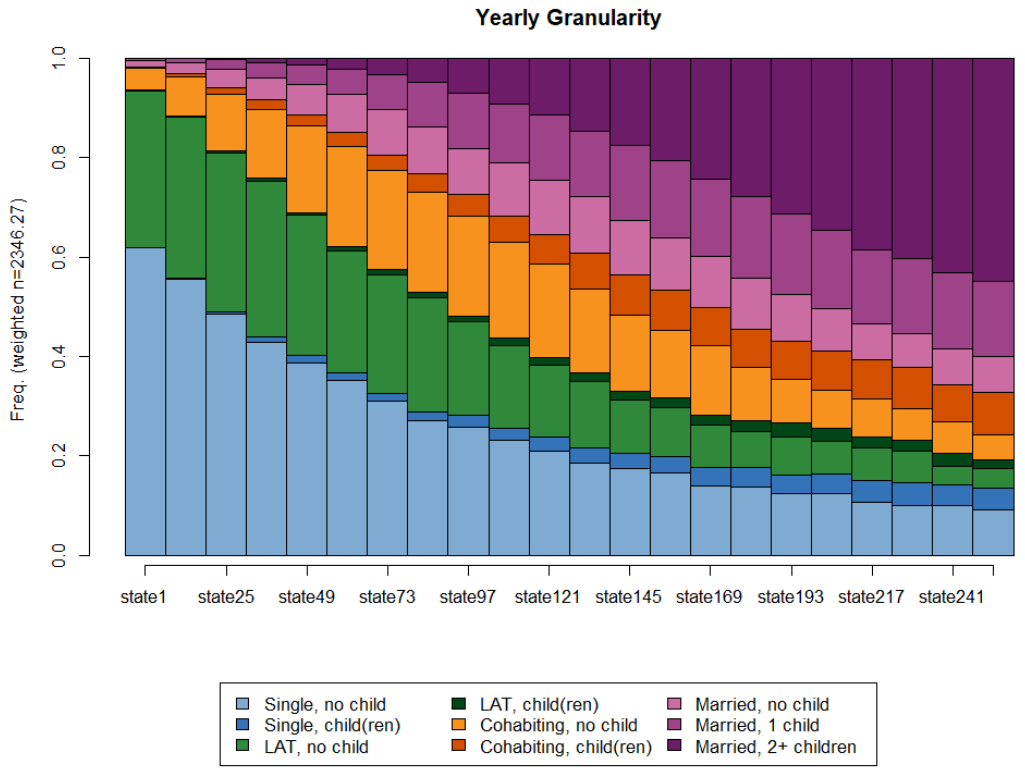
```
hcl_palettes(plot = TRUE)
```

Data summarizing graphs: Modal state plot



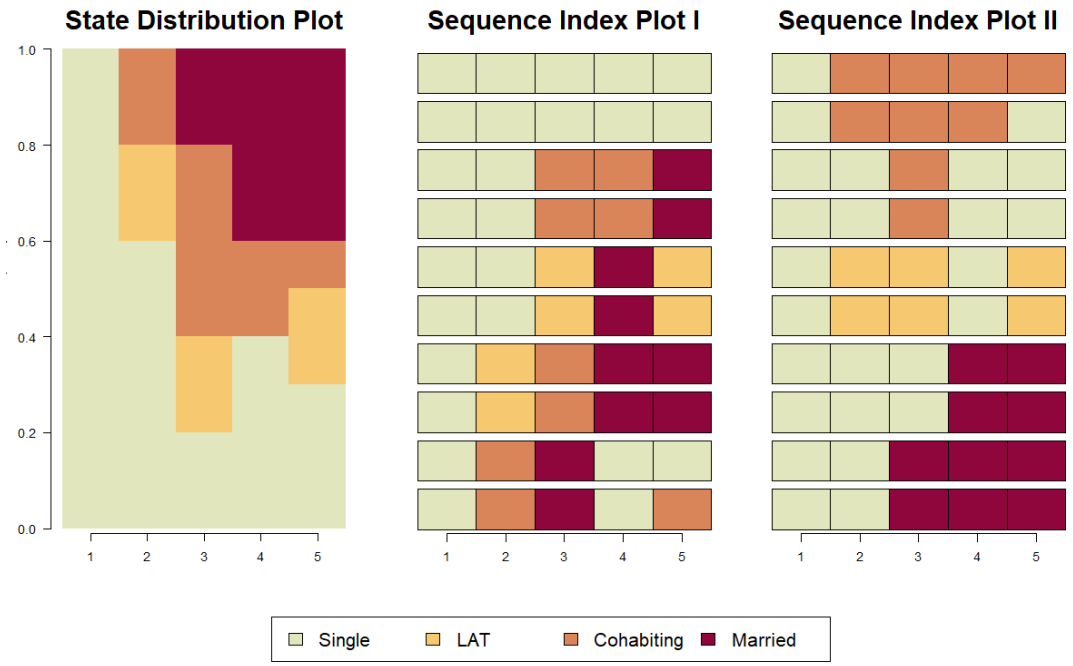


Data summarizing graphs: State distribution




```
seqdplot(family.year.seq, main = "Yearly Granularity")
```

Data representation graphs: Sequence index plot

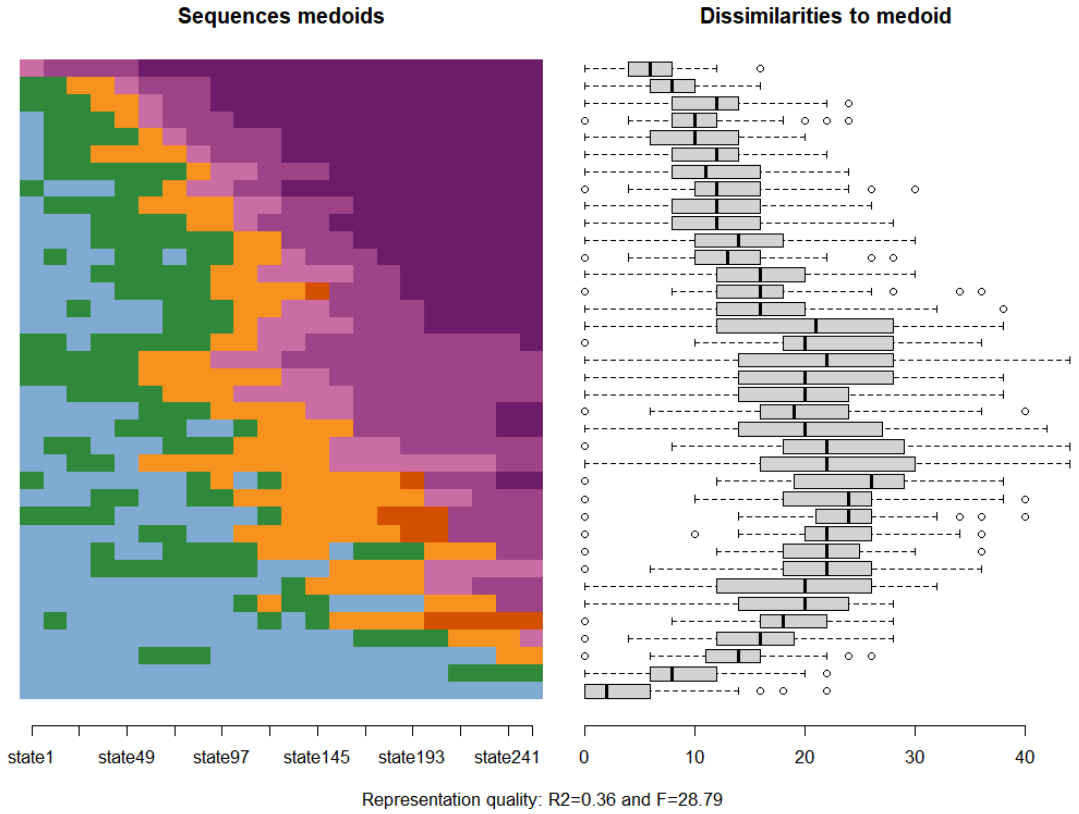


- Distribution plots easy to consume
- But potentially misleading due to aggregation: not strictly longitudinal, just repeated cross-sections
  → requires careful interpretation
- Different sequence sets can produce the same distribution
- Index plots show how the individual sequences unfold
- Index plots only feasible for a limited number of cases (300-400)

Data representation graphs: Relative frequency sequence plot

- The standard sequence index plot are at risk of overplotting when working with larger samples
- Even without overplotting high level of visual complexity
- The relative frequency sequence index plot (Fasang & Liao 2014) addresses these issues
- It renders only a group of representative sequences instead of the full sample
- Procedure:
  - Sequences are sorted (by timing of a specific transition or the score of the first factor obtained by multidimensional scaling
  - After sorting the sequences are divided into equally sized frequency groups (up to 100 seem to be feasible)
  - Only the medoid sequence of the frequency groups are rendered in an index plot

Data representation graphs: Relative frequency sequence plot

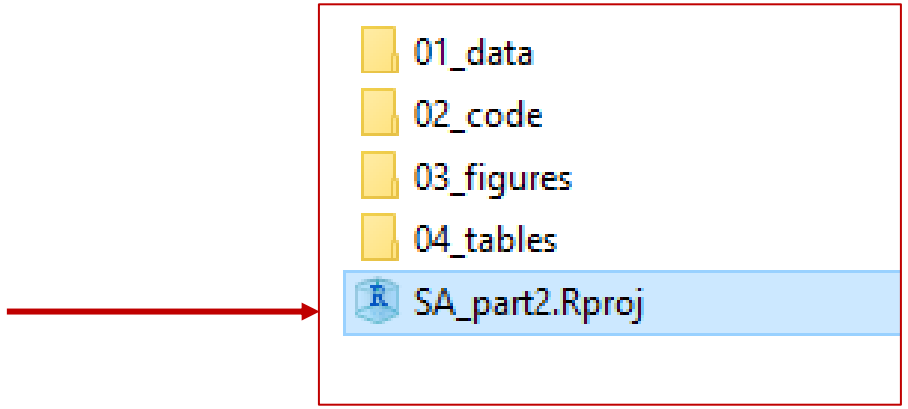



```
# compute distance matrix
family.year.om <- seqdist(family.year.seq,
                          method="OM",
                          sm= "CONSTANT")

# render default plot (TraMineRextras)
seqplot.rf(family.year.seq,
           diss=family.year.om,
           k=37)
```

And now the hands-on!

-----PHDP_SA_workshop\02_hands_on\part2



HUMBOLDT-UNIVERSITÄT ZU BERLIN

UNIVERSITÀ DEGLI STUDI DI MILANO

Sequence Analysis for Social Science

Anette Fasang and Emanuela Struffolino

PART 3

Population Dynamics and Health Program (PDHP)
University of Michigan
Feb. 9th, 2022

Outlook part 3

1. **Complexity and other summary measures (within sequence variation)**
2. **Optimal matching (between sequence dissimilarity)**
3. **Introduction to cluster analysis**

1. Complexity and other summary measures

Summary measures of individual sequences

- ***Unvalued*** (complexity, turbulence, and entropy) and ***valued*** summary measures:
- **Degradation:** proportion of positive to negative changes in sequence (net mobility)
- **Badness**: states weighted by degree of undesirableness of each state and probability to end up in "bad state" (enduring disadvantage and downward mobility)
- **Insecurity**: composite of complexity, degradation and undesirableness
- **Integration into positive state**: probability to end up in a positive state (upward mobility)

Summary indicators on individual sequence variability

Ritschard, G. (2021). Measuring the Nature of Individual Sequences. Sociological Methods & Research, 00491241211036156.

Struffolino, E. and M. Raab (forthcoming). Sequence Analysis. Sage, Series: Quantitative Methods in the Social Sciences

Table 1. Individual Characteristics of Sequences.

Indicator	Symbol	Short Name[a]	Source Citation[b]	Focus on States	Focus on Spells
Basic					
Length	ℓ	Lgth		×	
Number of nonmissing elements	ℓ_v	Nonm		×	
Number of visited states	vn	Visited		×	
Proportion of visited states	vp	Visitp	CBF	×	
Number of transitions	tn	Trans			×
Proportion of transitions	tp	Transp	AG, CBF		×
Number of spells	ℓ_d	Dlgth			×
Mean spell duration	$\bar{d}$	Meand			×
including 0-length spells	$\bar{d}^*$	Meand2	new		×
Recurrence index	ψ	Recu	PBS	×	
Diversity					
Normalized entropy	h_{norm}	Entr	CS	×	
Spell duration standard deviation	s_d	Dustd	CE		×
including 0-length spells	s_d^*	Dustd2	new		×
Complexity					
Number of subsequences of the DSS sequence	ϕ	Nsubs	CE		×
Objective volatility	v	Volat	CBF	×	
Complexity index	c	Cplx	AG	×	
Turbulence	T	Turb	CE		×
including 0-length spells	T^*	turb2	new		×
Normalized turbulence	T_n	Turbn	new		×
including 0-length spells	T_n^*	Turb2n	new		×
Binary					
Proportion of elements of interest	I_{ppos}	Ppos		×	
Normative volatility	I_{nvolat}	Nvolat	CBF		×
Integrative potential	I_{integr}	Integr	CBF, MM	×	
State undesirableness					
Degradation index	I_{degrad}	Degrad	new		×
Badness	I_{bad}	Bad	new	×	
Precarity index	I_{prec}	Prec	RBO	×	×
Insecurity index	I_{insec}	Insec	new	×	×

Note: DSS = distinct successive states.

[a] Names used by the seqindic function of the TraMineR R package.

[b] AG: Gabadinho et al. (2010); CBF: Brzinsky-Fay (2007, 2018); CE Elzinga and Liefbroer (2007); CS: Shannon (1948); MM: Manzoni and Mooi-Reci (2018); PBS: Pelletier, Bignami-Van Assche, and Simard-Gendron (2020); RBO: Ritschard, Bussi, and O'Reilly (2018).

Example: Family life course **complexity** varies much more between European countries than across cohorts

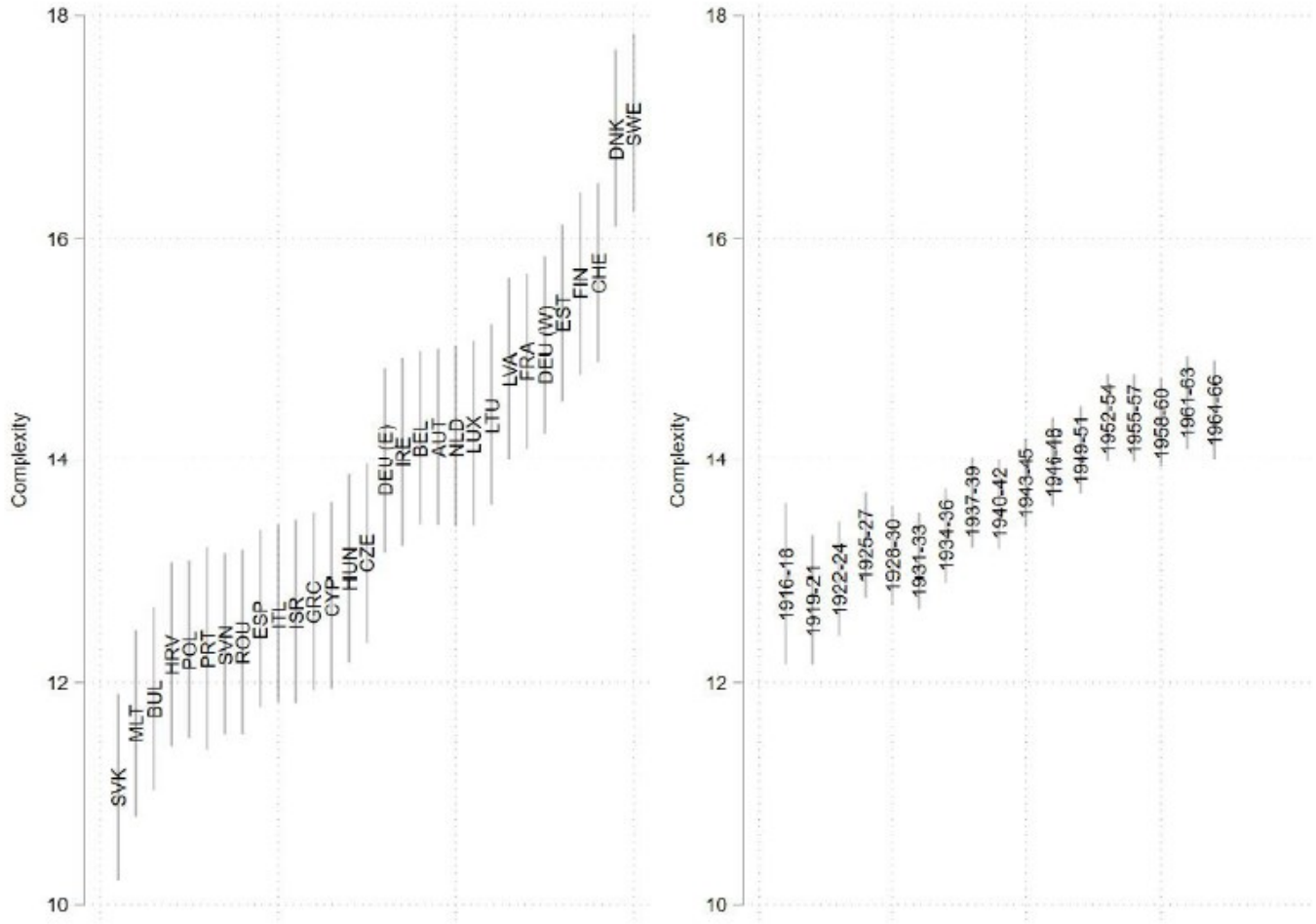
$$C(x) = 100 * \sqrt{\frac{q(x)}{q_{max}} * \frac{h(x)}{h_{max}}},$$

number of transitions in a sequence, *q(x)* divided by the theoretical maximum number of transitions possible, q_{max};
longitudinal entropy of a sequence, *h(x)* divided by the theoretical maximum, h_{max}.

Complexity is minimal in sequences composed of a single state and maximal in sequences that contain each state element with equal durations and have the maximum number of transitions.

Van Winkle, Zachary, and Anette Eva Fasang. "The complexity of employment and family life courses across 20 th century Europe." Demographic Research 44 (2021): 775-810.

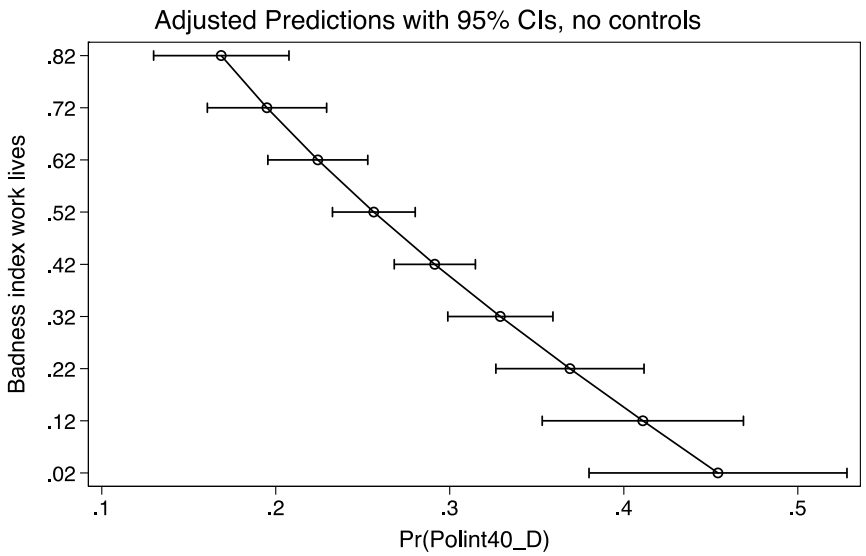
Figure 3: Empirical Bayes estimates of family complexity by cohort and country



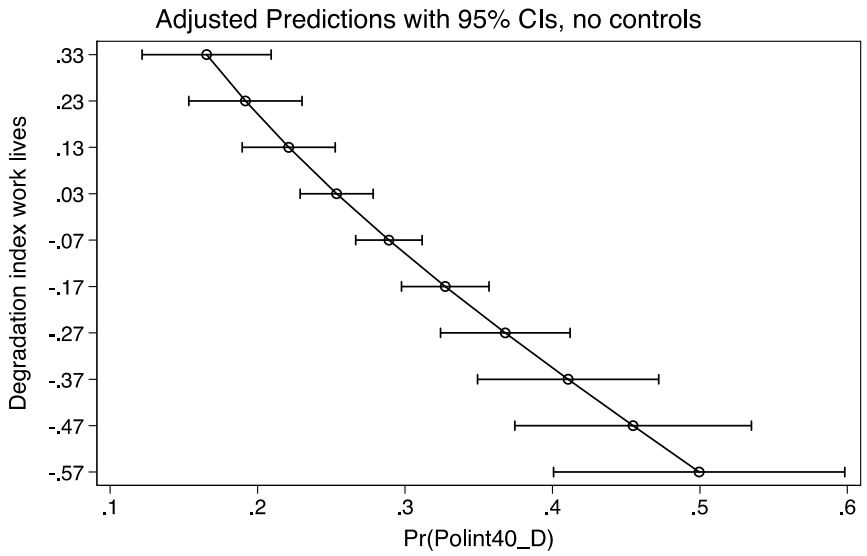
Example: **Badness** and **degradation indices** in work trajectories predict political interest at mid-life

→ Downwardly mobile individuals (high badness & high degradation) have lower political interest at mid-life

Badness



Degradation

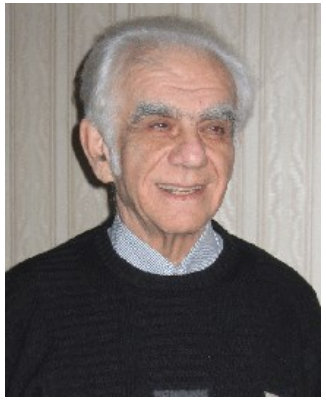


Data: German Socioeconomic Panel (GSOEP), Fasang 2021

2. Optimal Matching (OM)

OM: between sequence dissimilarity

Originally developed by **Vladimir Levensthein** in 1965, has been widely used in the natural sciences for identifying similarity in DNA strings and in computer science as the basis of word- and speech-recognition algorithms.



Andrew Abbott first introduced sequence analysis and OM to the social sciences (1990s)

Optimal Matching

- **Goal**
  Quantify degree of similarity of sequence pairs – actually OM is assessing the dissimilarity

- **Computational Approach**
  Find the “cheapest” way to transform one sequence into another (sequence alignment)

- **Result**
  Pairwise distance matrix for all possible sequence pairs

Typical Steps in OM

1. define sequence: start, end, alphabet of states

2. calculate distances between sequences by applying OM Algorithm → obtain distance matrix

3. further analysis of distances → cluster analysis: typology of trajectories

Alternatives to clustering of distances:

- further analysis of groups/distances (visual, regression based methods)
- distances as direct indicators of dissimilarity (de-standardization) or as input to other techniques (BIC/LRT adaptations for groups comparisons)

Example: distance matrix for four sequences

	1	2	3	4
1	0	1	5	2
2	1	0	0	3
3	5	0	0	2
4	2	3	2	0

Distance is a metric of dissimilarity:

- High values = very dissimilar sequences
- Low values = very similar sequences, 0= identical sequences
- Note: There are axiomatic differences between distance & similarity

Calculating sequence distance with OM

Finding the “cheapest” way to transform one sequence into another = distance between two sequences

For this, OM uses three transformation operations with specific costs:

1. Substituting one state with another → sub costs

2. Inserting/deleting a state → indel costs

•

OM: example of aligning two sequences

Alphabet of States

Unemployment= UE
Education = ED
Internship = I
Employment = EM

Transformation Costs

sub costs = 2
indel costs = 1

Time	t1	t2	t3	t4
ID1	UE	I	EM	EM
ID2	ED	UE	EM	

Cost/distance

2 + 2 + 1 = 5

2 Substitutions + 1 Insertion

Cost specification and sequence dissimilarity

	Insertion-Deletion	Substitution
Preserved	Events	Time
Altered	Time	Events

To emphasize similarity in terms of **timing** of transitions/states:
→ high indel costs & low substitution costs

To emphasize similarity in terms of the **order** of states:
→ low indel costs & high substitution costs

“The assignment of transformation costs haunts all optimal matching analyses”

Katherine Stovel, Mike Savage and Peter Bearman, 1996, “Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970”, *American Journal of Sociology*, 102, p. 394

Yes and No

- some cost specifications not sensible for most social science applications
- many produce similar results
- results often more sensitive to alphabet and clustering
- reviewers usually want to see justification and robustness checks

See: Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. Journal of the Royal Statistical Society: Series A (Statistics in Society), 179(2), 481-511.

See: Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. Journal of the Royal Statistical Society: Series A (Statistics in Society), 179(2), 481-511.

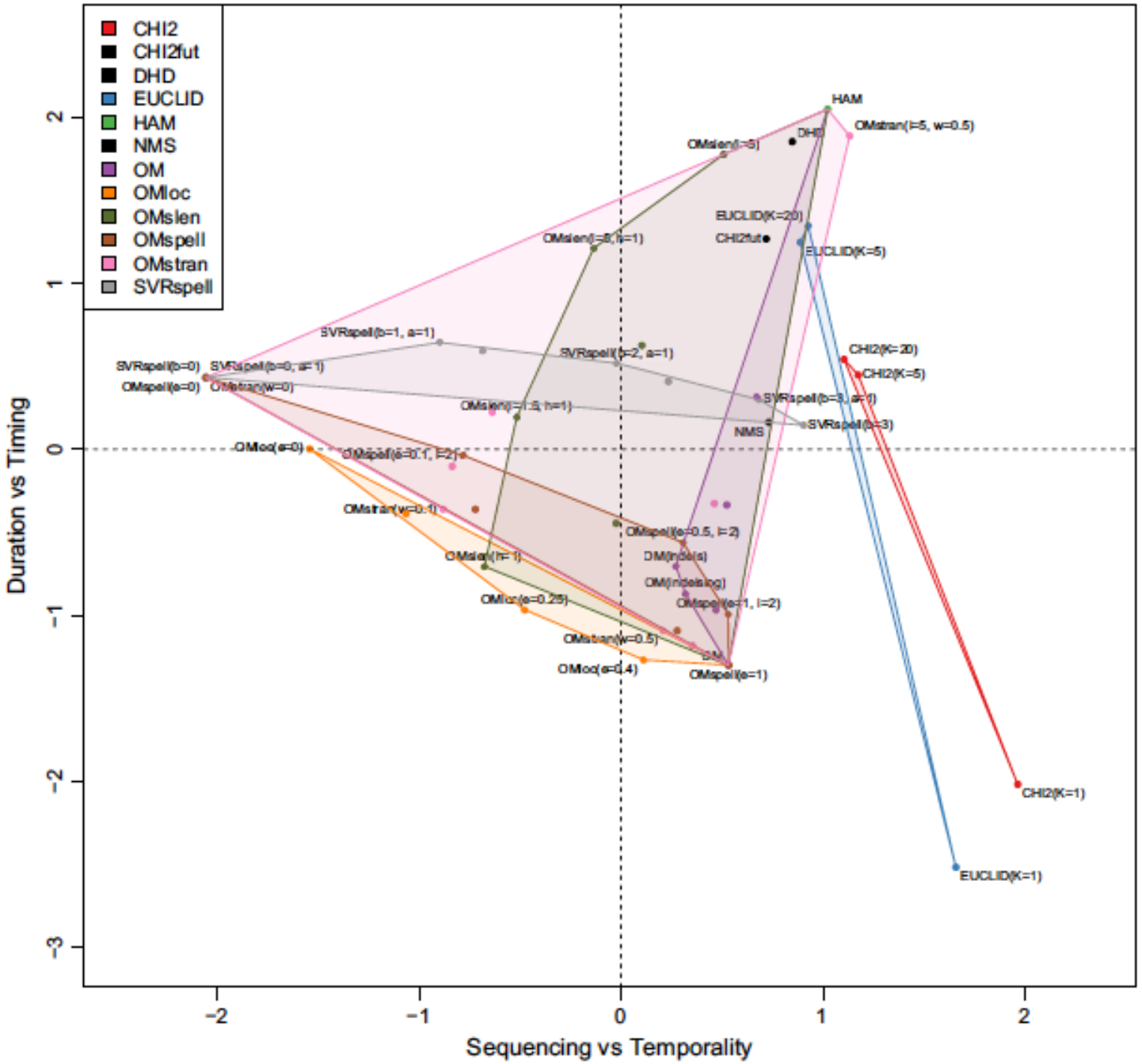


Fig. 1. Scores for state-based simulations

3. Cluster analysis

Objectives

Partition the Sequences into Groups that are:

- As similar to each other as possible
- As dissimilar to sequences in other groups as possible

Cluster Analysis is a tool to simplify complex data:

- This is achieved by ignoring minor differences
  - Produces meaningful typologies
- Danger of ignoring major differences
  - Can lead to false conclusions

Objectives

Description

- Which ideal typical trajectories exist in the data?

Further Analysis Trajectory Types/Clusters

- Why are given ideal types in the data?
  - Institutions, historical circumstances, etc.
- Who is a member or which life course type and why?
  - Race, gender, social class, etc.
- What are the consequences of belonging to a life course type?
  - Part-time employment for women → lower pensions & higher old-age poverty

Cluster analysis: algorithms

- Hierarchical Methods
  - bottom-up, agglomerative-nesting (starts by grouping closest two cases)
  - top-down, divisive (starts by splitting full sample)

- Partitioning around centers
  - k-means (for quantitative data)
  - partitioning around medoids (more general)

Cluster analysis: many different methods, Ward most commonly used

Name	Function	Weights	Interpretation and notes
single	`hclust`	Indep	Merge groups with the closest observations.
complete	`hclust`	Indep	Minimize the diameter (highest distance) of the new group (very sensitive to outliers).
average (ou UPGMA)	`hclust`	Yes	Average distances.
McQuitty (ou WPGMA)	`hclust`	Indep	Sensitive to previous merges.
centroid	`hclust`	Yes	Minimum distance between group medoïds.
median	`hclust`	Indep	Sensitive to previous merges.
ward	`hclust`	Yes	Minimize residual variance.
beta-flexible	`agnes`	No	Use `par.method=0.625` to set $\beta = -0.25$.

`agnes` is available in the `cluster` library. Set `diss=TRUE` to use your own distance matrix!

```r
# apply the clustering algorithm Ward
fam.ward0 <- hclust(as.dist(om.fam), method = "ward.D", members =
family$weight40)
```

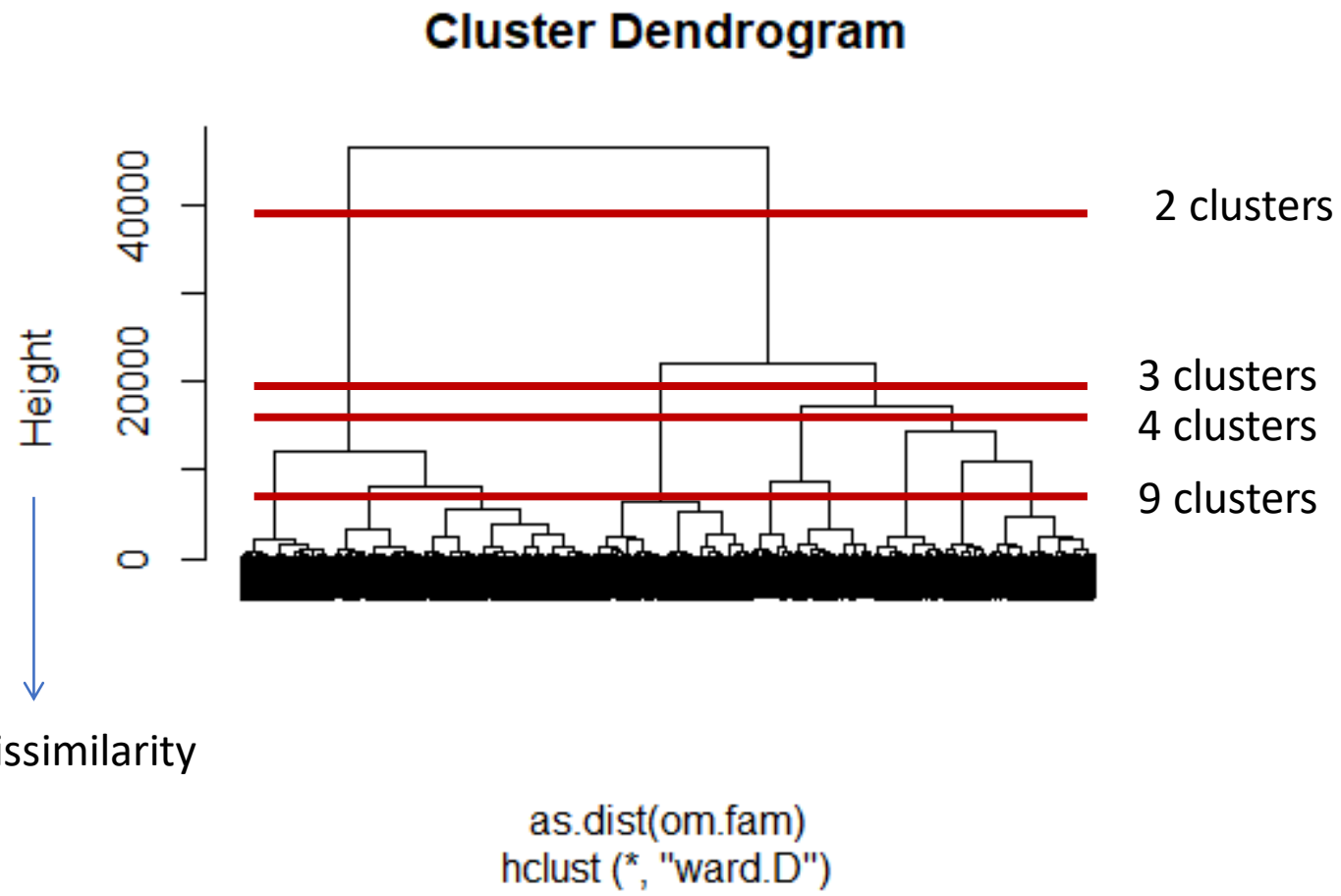
weights

Hierarchical cluster function

Distance matrix

Higherachical method

```r
#plot the dendrogram
plot(fam.ward0, labels = FALSE)
```



But how many clusters to choose?

Is the visual inspection of the dendrogram enough?

How many groups? Cluster cut-off criteria

These tend to favor high numbers of clusters

Name	Abrv.	Interval	Min/Max	Interpretation
Point Biserial Correlation	PBC	$[-1; 1]$	Max	Capacity of the clustering to reproduce the original distance matrix.
Hubert's Gamma	HG	$[-1; 1]$	Max	Capacity of the clustering to reproduce the original distance matrix (Order of magnitude).
Hubert's Somers D	HGSD	$[-1; 1]$	Max	Same as above, taking into account ties in the distance matrix.
Hubert's C	HC	$[0; 1]$	**Min**	Gap between the current quality of clustering and the best possible quality for this distance matrix and number of groups.
Average Silhouette Width	ASW	$[-1; 1]$	Max	Coherence of the assignments. A high coherence indicates high between groups distances and high intra group homogeneity.
Calinski-Harabasz index	CH	$[0; +\infty[$	Max	Pseudo F computed from the distances.
Calinski-Harabasz index	CHsq	$[0; +\infty[$	Max	Idem, using the *squared* distances.
Pseudo R^2	R2	$[0; 1]$	Max	Share of the discrepancy explained by the clustering.
Pseudo R^2	R2sq	$[0; 1]$	Max	Idem, using the *squared* distances.

How many clusters to choose?

```
#test cluster solution quality
fam.wardtest <- as.clustrange(fam.ward0, diss = om.fam,
                              weights =family$weight40, ncluster = 7)
```

hclust outcome

Distance matrix

Nr. of clusters to test

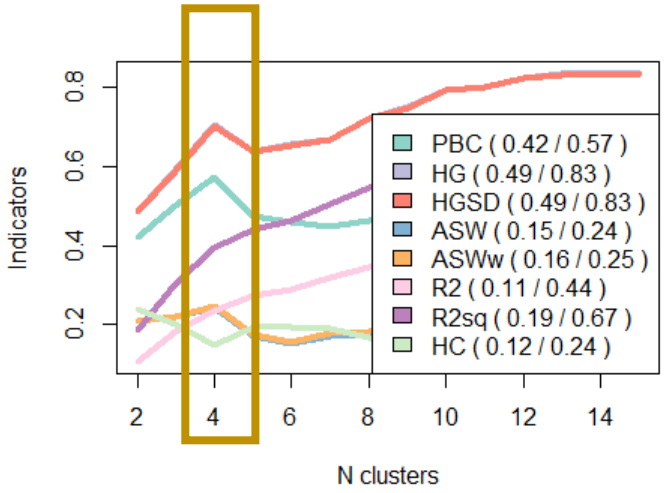
```
#print the quality test for different cluster solutions
fam.wardtest
```

```
##             PBC   HG HGSD  ASW ASWw     CH   R2   CHsq R2sq   HC
## cluster2 0.27 0.30 0.30 0.18 0.18 183.02 0.12 319.45 0.19 0.33
## cluster3 0.44 0.52 0.51 0.21 0.21 163.87 0.19 310.96 0.31 0.24
## cluster4 0.54 0.66 0.66 0.23 0.23 138.52 0.23 282.08 0.38 0.17
## cluster5 0.56 0.71 0.71 0.23 0.23 122.67 0.26 258.69 0.43 0.15
## cluster6 0.49 0.66 0.66 0.16 0.16 112.51 0.29 236.10 0.46 0.19
## cluster7 0.51 0.70 0.70 0.18 0.18 108.39 0.32 237.62 0.51 0.17
```

hclust outcome

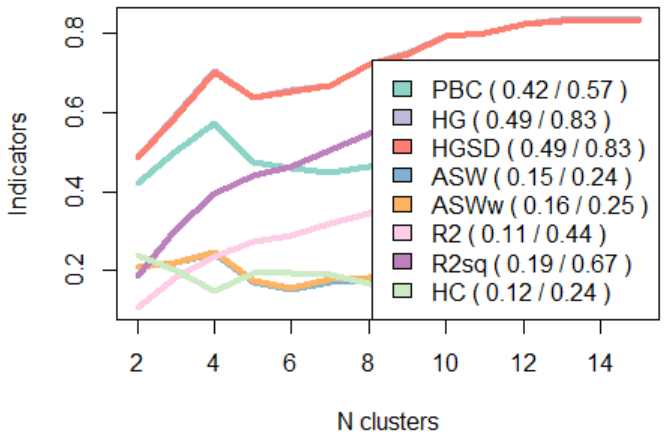
Width of the lines in the plot

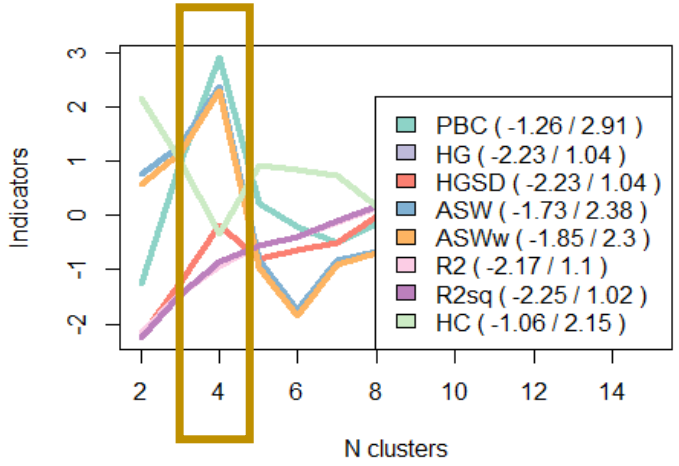
```
#plot the quality criteria
plot(fam.wardtest, lwd = 4)
plot(fam.wardtest, norm = "zscore", lwd = 4)
plot(fam.wardtest, stat = c("ASW", "HC", "PBC"), norm = "zscore", lwd = 4)
```



Normalize the indicators for comparison == shows relative increase/decrease in the values of the indicators

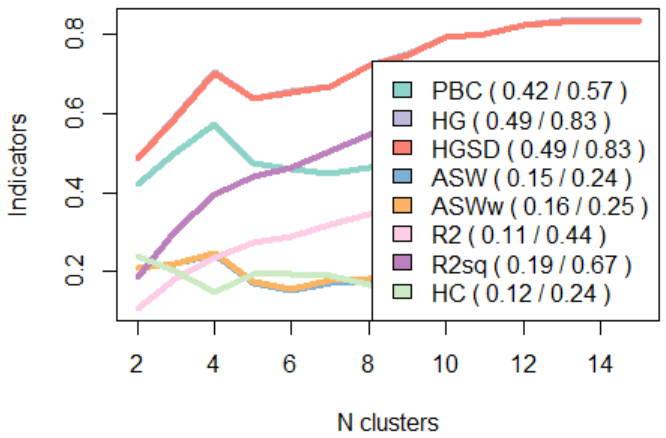
```
#plot the quality criteria
plot(fam.wardtest, lwd = 4)
plot(fam.wardtest, norm = "zscore", lwd = 4)
plot(fam.wardtest, stat = c("ASW", "HC", "PBC"), norm = "zscore", lwd = 4)
```

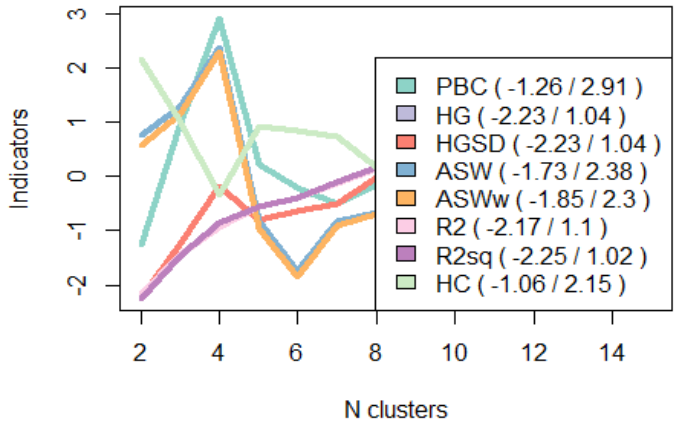


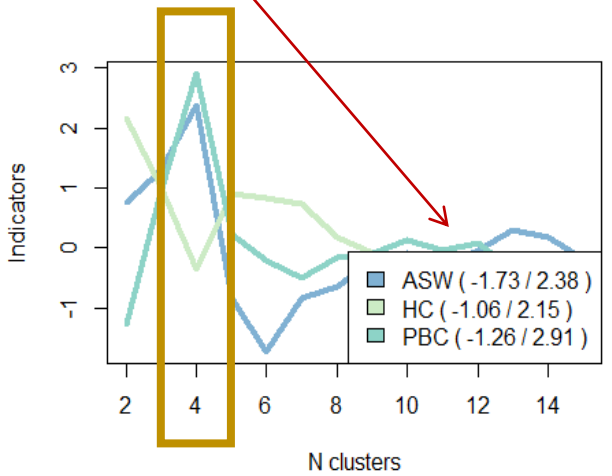


```r
#plot the quality criteria
plot(fam.wardtest,  lwd = 4)
plot(fam.wardtest, norm = "zscore", lwd = 4)
plot(fam.wardtest, stat = c("ASW", "HC", "PBC"), norm = "zscore", lwd = 4)
```

Display selected indicators







```
#plot the AWS silhouette by cluster
summary(silh.ward <- silhouette(fam.ward, dmatrix = om.fam))
```

Distance matrix

cutree object

```
## Silhouette of 1143 units in 4 clusters from silhouette.default(x
= fam.ward, dmatrix = om.fam) :
##  Cluster sizes and average silhouette widths:
##        472        163        295        213
## 0.31092964 0.08809121 0.06326628 0.28219562
## Individual silhouette widths:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.29858  0.07308  0.22855  0.20988  0.37262  0.50624
```

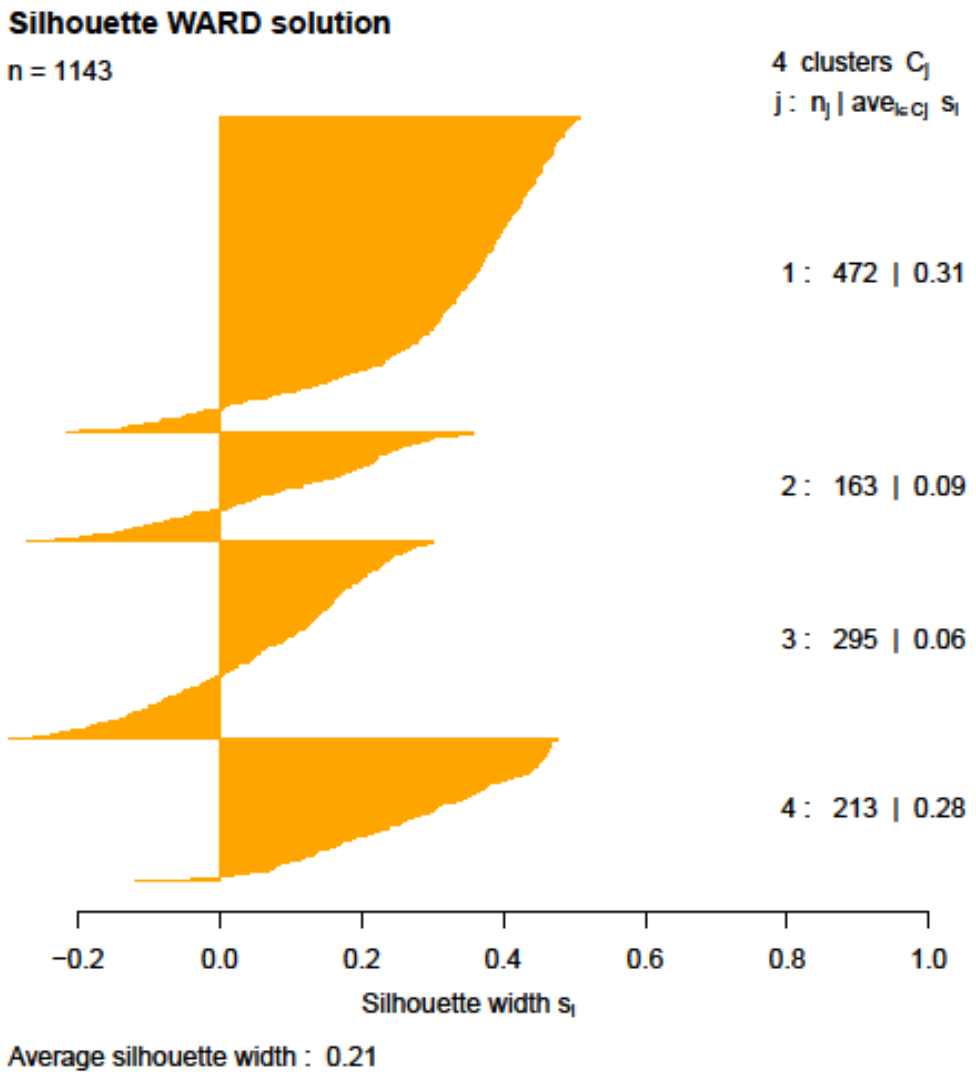
Sil. for each cluster

Statistics on sil. for individuals

```
#plot the AWS silhouette by cluster
pdf('silh-ward.pdf')
plot(silh.ward, main = "Silhouette WARD solution", col="orange")
dev.off()
```

Title of the plot

Color of the plot



Evaluation of cluster quality

- Cluster analysis is a simplification of the data
  - This simplification may be meaningful or not
  - But cluster analysis always produces something!
  - Cluster quality needs to be evaluated

- Several choices need to be made:
  - Choose the cluster analysis algorithm
  - Choose the number of groups to keep
  - Cluster quality measures allow to compare (from a statistical point of view) solutions provided by different algorithms or numbers of groups

Evaluation of cluster quality

“ ***Construct validity****. Another important indicator of the validity of the groups found with sequence analysis is the plausibility of the results and their theoretical interpretability. The validation criterion referring to this is construct validity. Construct validity is based on the logical and empirical relationship among constructs (Babbie 1979). Translated to sequence analysis, we can state that if groupings found with sequence analysis relate to variables as theoretically expected, this indicates construct validity.”* (Aisenbrey and Fasang 2010)

Evaluation of cluster quality

- A good typology should be (Shalizi, 2009):
  - generalizable to other observations
  - generalizable to other properties (variables)
  - linked to a theory

- Cluster analysis is a good descriptive method for the specific data at hand.
- Generalization only with caution!

Recommendations

Building your sequences (coding, alignment, ...):

- Very important, determines what you can find

Clustering procedure

- Try (a lot of) different clustering algorithms
- Select the best solutions according to cluster quality measures
- Interpret them
- Select the one with the best interpretation

Interpretation

- Interpret according to cluster quality
- Do not ignore intra-cluster variability (more problematic when using clusters as dependent or independent variables → classification error)

Outlook: recent developments in SA

- Multichannel Sequence Analysis (MSA)

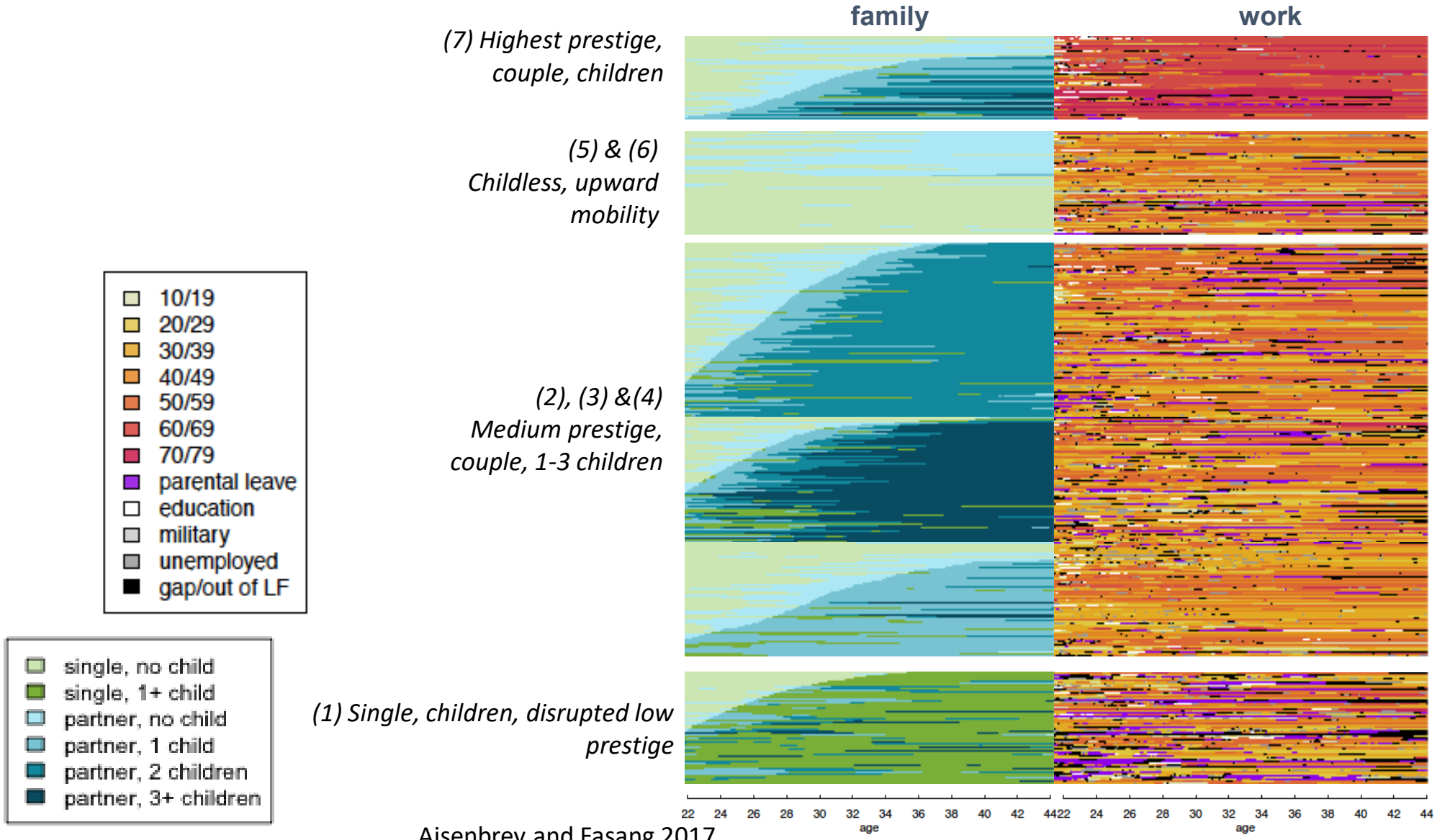
- Additional advances in Sequence Analysis
  - Visualization
  - Combination with other methods to overcome limitations
  - Sequence Analysis multistate model (combination with event history analysis) (Studer et al. 2018)
  - BIC and LRT for sequence comparison (Liao and Fasang 2021)

More on this:

- Sequence Analysis Association: https://sequenceanalysis.org/
- TraMineR: http://traminer.unige.ch/
- Raab, M. & Struffolino, E. (2022). Sequence Analysis. Thousand Oaks, CA: Sage.

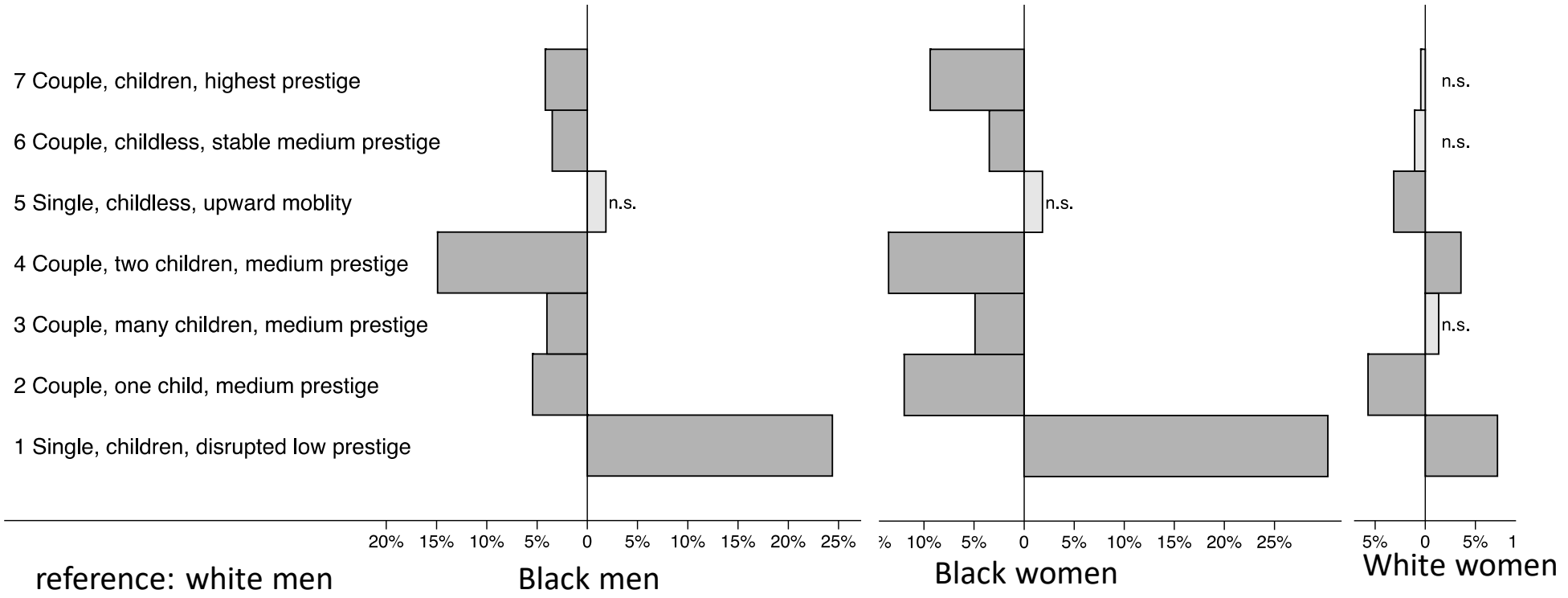
MSA: Seven work-family patterns – a "process measure of inequality"

(sorted by average prestige)



Aisenbrey and Fasang 2017

Typology as dependent variable in logistic regression: race*gender interaction



See also: Fasang, A. E., & Aisenbrey, S. (2021). Uncovering Social Stratification: Intersectional Inequalities in Work and Family Life Courses by Gender and Race. Social Forces.

HUMBOLDT-UNIVERSITÄT ZU BERLIN

UNIVERSITÀ DEGLI STUDI DI MILANO

Sequence Analysis for Social Science

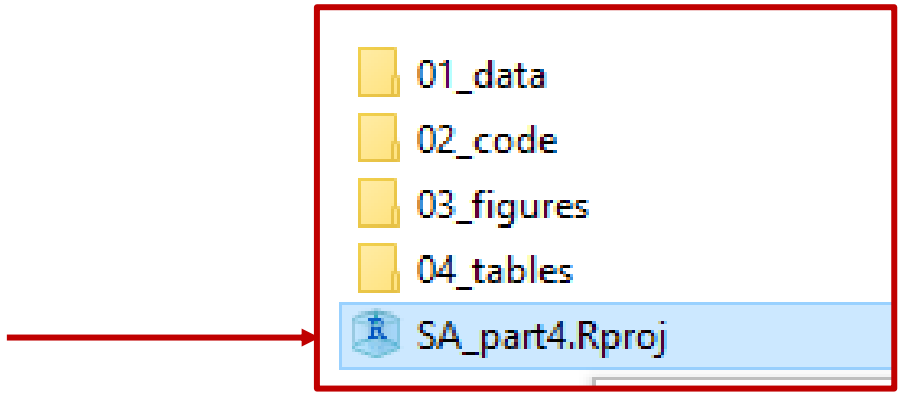
Anette Fasang and Emanuela Struffolino

PART 4

Population Dynamics and Health Program (PDHP)
University of Michigan
Feb. 9th, 2022

And now the hands-on!

-----PHDP_SA_workshop\02_hands_on\part4



https://us.sagepub.com/en-us/nam/sequence-analysis/book272086

https://sa-book.github.io/

OUT IN MAY!

Sequence Analysis - Companion Site

Home | Material for R

Sequence Analysis

Companion website for the little green book

AUTHORS

Marcel Raab

Emanuela Struffolino

AFFILIATIONS

State Institute for Family Research at the University of Bamberg (ifb)

University of Milan, Department of Social and Political Sciences

PUBLISHED

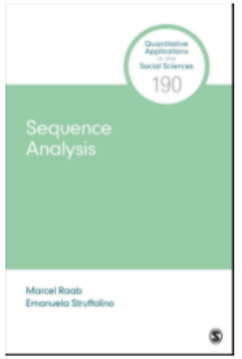
Feb. 1, 2022

CITATION

Raab & Struffolino, 2022

This site is intended to enhance your use of the book **Sequence Analysis** by Marcel Raab & Emanuela Struffolino. On this webpage we provide accompanying material illustrating how to conduct sequence analysis in R using `{TraMineR}` (Gabadinho et al., 2011), `{TraMineRExtras}` (Ritschard et al., 2021), and `{WeightedCluster}` (Studer, 2013).

If you simply want to download the data and code required for reproducing the results can access the chapter-specific zip files on this overview page. In addition to these zip files, the chapter-specific sub-pages provide instructions and some bonus material introducing additional examples and analytical tools.



"This book provides a comprehensive and updated introduction to sequence analysis, I highly recommend it for anyone who wants to learn the topic systematically."

— Tim F. Liao (University of Illinois at Urbana-Champaign)