Introduction to sequence analysis (Lecture 3).

Clustering and Examples

Nicola Barban

University of Bologna

January 20-21 2022

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

dissimilarity measures

- Distance between sequences Different metrics (LCP, LCS, OM, HAM, DHD, ...)
- A dissimilarity is a quantification of how far two objects are. For instance, consider two incomes x and y:
  - $d(x, y) = (x - y)^2$
  - $d(x, y) = |x - y|$
  - $d(A_{x_1,y_1}, B_{x_1,y_2}) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- Optimal Matching, or LCS, DHD, . . . compute distances for categorical trajectories?

Cluster

- Cluster analysis automatically classify different objects in a reduced number of categories.
- It simplifies the large number of distinct sequences in a few different types of trajectories.
- It is used to build a typology of the trajectories. It offers a descriptive approach to analyze the sequences.

Cluster

- Clustering always start from a distance matrix. Usually euclidean distances between variables
- But clustering may be done using a dissimilarity matrix.
- Several methods for agglomerating observations in cluster procedures
- Usually iterative procedure. At every step the most “similar” observations are grouped

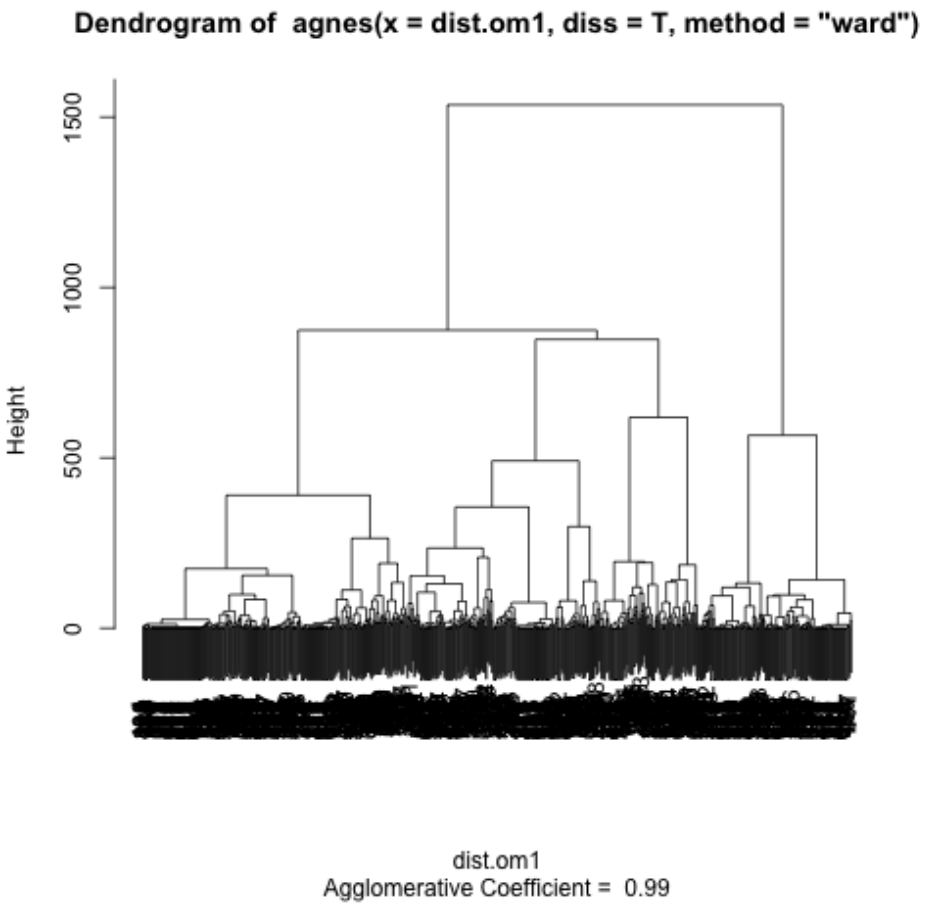
Ward clustering

- Ward is a hierarchical clustering algorithm.
- At each step, it joins together the two less distant groups.
- Ward aims at minimizing the within cluster discrepancy.

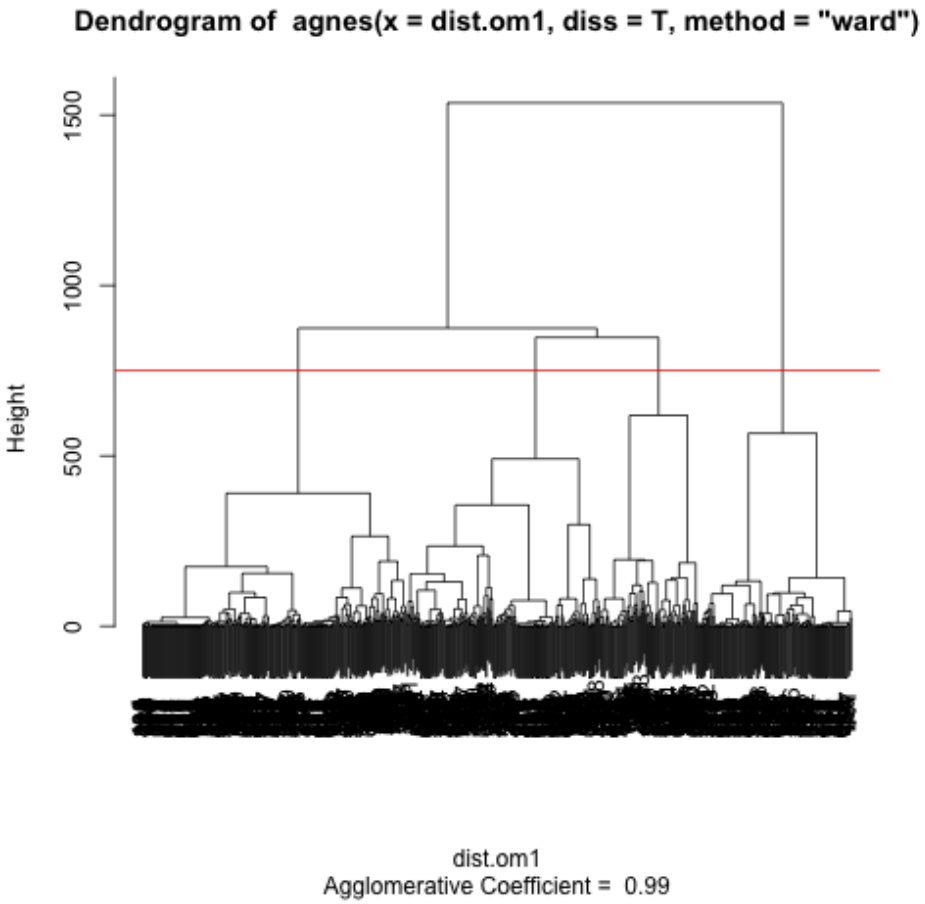
Number of clusters

- The number of clusters needs to be chosen by the researcher
- Several way to do that. No best method
  1. Theory driven. You have some reason to believe that the best number of group is . . .
  2. Description of the clusters. Try different solutions
  3. Dendogram

Dendogram


Dendrogram of agnes(x = dist.om1, diss = T, method = "ward")
Height
0
500
1000
1500
dist.om1
Agglomerative Coefficient = 0.99

Dendogram pruning



Analysis of cluster

- Check the sample size of each cluster. You don't want to have too small clusters
- Check the distribution of clusters. Do you have "residual" clusters
- Try one less clusters. Check distribution
- Be parsimonious.

Medoid

- Clusters can be described by their “center”
- This is called centroid sequence or **medoid**
- What is the sequence that is more “central”?
- “centrality” is equivalent less distance.
- The medoid distance is the sequence that is less distant in average to all the other sequences in the cluster

Medoid 2

- Medoid are **real** sequence
- Easy to describe!
- (S-12)-(C-6)-(M-24)
- (S-6)-(C-03)-(S-09)-(M-12)-(S-12)

Exploring clusters

Three types of graphics:

- Transversal distribution with `seqdplot()`
- Frequency plots with `seqfplot()`
- Individual index-plots `seqiplot()`

Use `group = cluster.membership.factor` to get plots by clusters

Determinants of trajectories

- It is possible to estimate the influence of independent covariates on the probability of belonging to a given cluster (i.e. type of trajectory) rather than another.
- We can fit, for instance, a logistic (multinomial) regression model
- Class membership can be used for further analysis

logistic regression

```
> summary(jobless.reglog)

Call:
glm(formula = jobless ~ male + funemp + gcse5eq, family = bino
    data = mvad)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8116  -0.5948  -0.5813  -0.3565   2.3613

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.64230    0.19297  -8.510  < 2e-16 ***
maleMen     -0.05032    0.22333  -0.225 0.821748
funempyes    0.70083    0.25466   2.752 0.005923 **
gcse5eqyes  -1.03169    0.27872  -3.702 0.000214 ***
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Ä

(Dispersion parameter for binomial family taken to be 1)
```

Example : Family trajectories and Health

From:

Barban (2013) Family Trajectories and Health: A Life Course Perspective. European Journal of Population

What is the association between family trajectories and health?

Lower health outcomes may be associated with:

1. Earlier transitions (*timing*)
2. Number of transitions (*quantum*)
3. Non-normative transitions (*sequencing*)

Moreover, there might be some specific **patterns** of family formation that are associated with lower health outcomes.

Data

- National Longitudinal Study of Adolescent Health **(Add Health)**
- Nationally representative sample of U.S. students in grades 7 through 12 in 1994. Cohort born 1976-1982.
- Four waves: WI 1995; WII 1996; WIII 2001–2002; WIV 2008–2009.
- Sample size: 20,000 students in wave I

For this study I restrict the sample to women who are 30 or older at Wave IV. The final sample size is 2,358

Sequences' state space

Monthly sequence from age 15 to 30
In each month individuals can be classified as:

1. Single (S)
2. Single Parent (SP)
3. Cohabiting (C)
4. Cohabiting parent (CP)
5. Married (M)
6. Married parent (MP)

Distribution of family states

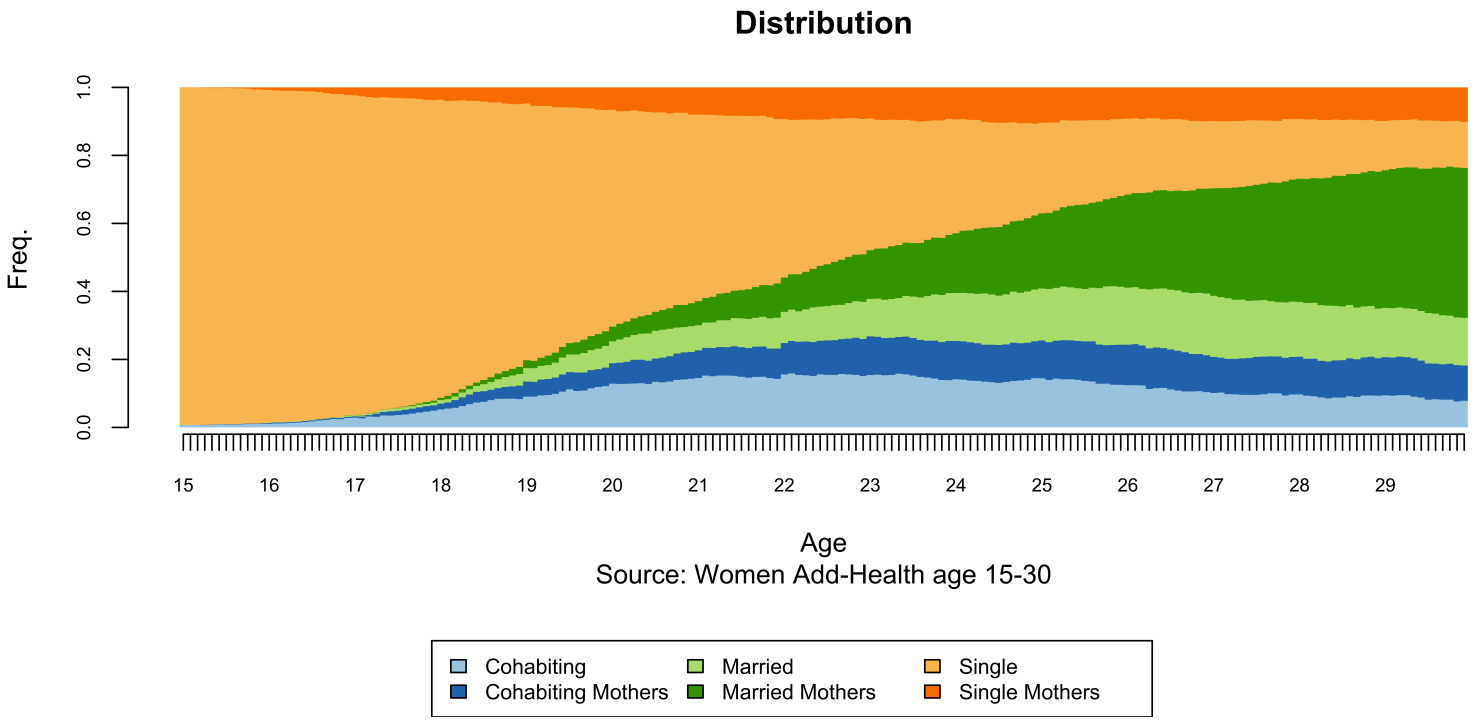


Table: First 10 sequence pattern of transitions in Women 15-30. Weighted frequencies.

		Freq
1	S-C-M-MP	11.46
2	S-M-MP	10.46
3	S-C-M	5.93
4	S-C-CP-MP	4.41
5	S	4.37
6	S-C-S	3.46
7	S-C-S-C-M-MP	3.37
8	S-M	3.15
9	S-C	3.07
10	S-SP-CP-MP	2.77

Data

Data: Add-health. Women 30 or older.
Health Outcomes: (continuous vars)

- Self-reported health
- CES–D Depression scale
- # cigarettes smoked in the last month
- # number of episodes of heavy drinking in the last year (5 or more alcoholic cocktails)

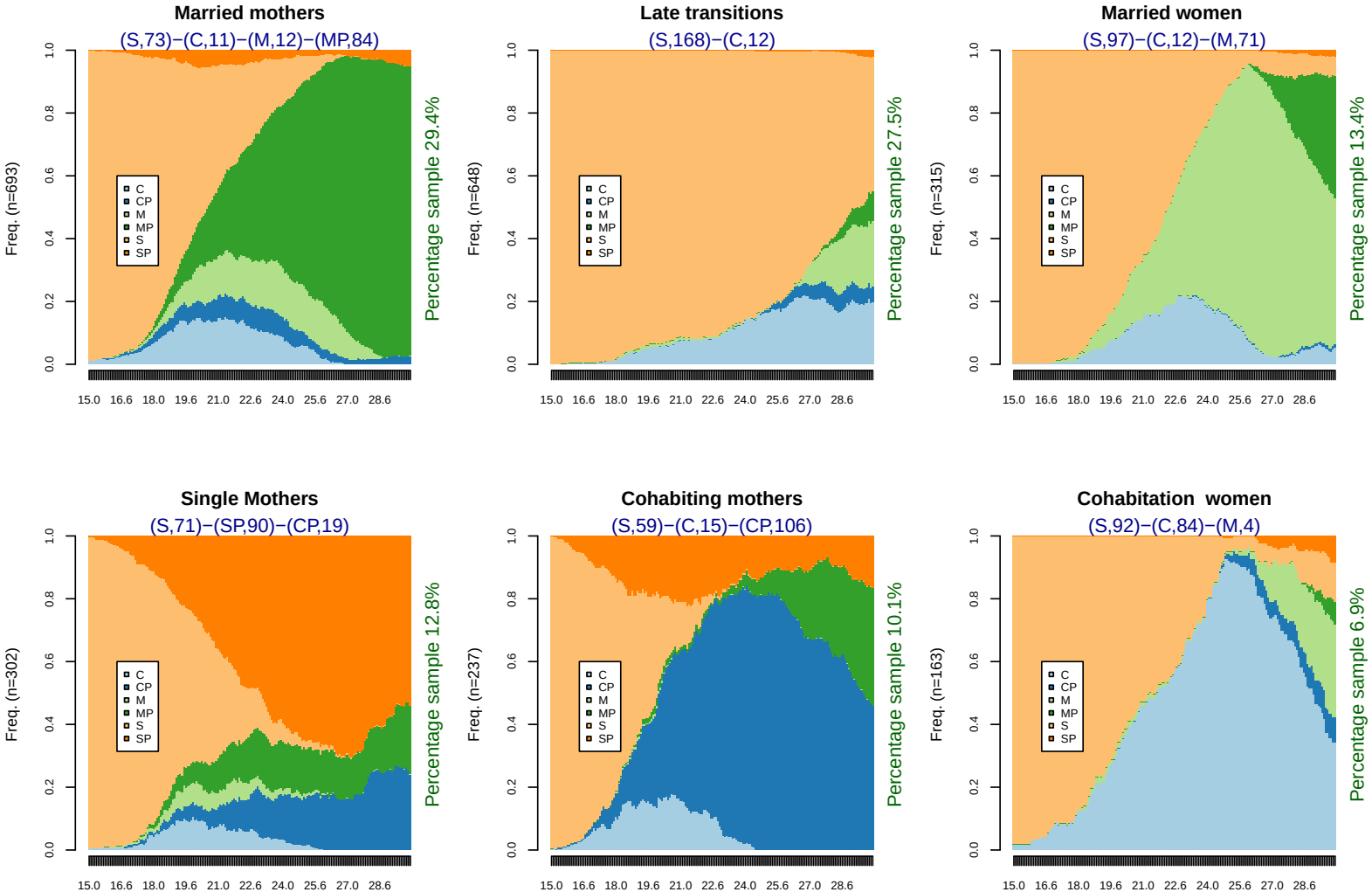
Independent variables:

- Age at first transition (union, child) (*timing*)
- Number of transitions from wave I to wave IV (*quantum*)
- Number of non-normative transitions from wave I to wave IV (norm$\rightarrow$ S–M–MP) (*ordering*)

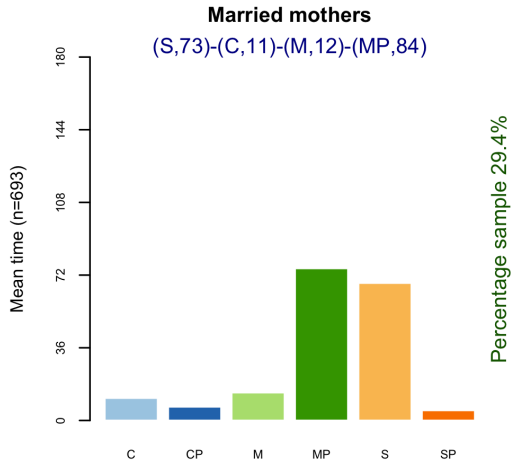
  Controls:

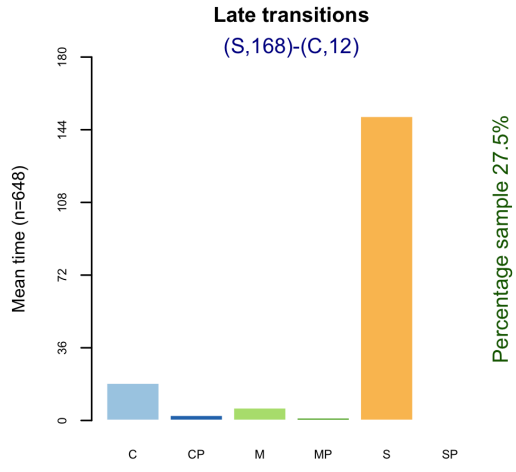
  - Age, Age sq., Race/Ethnicity, Family composition at Wave I, Parent's education.

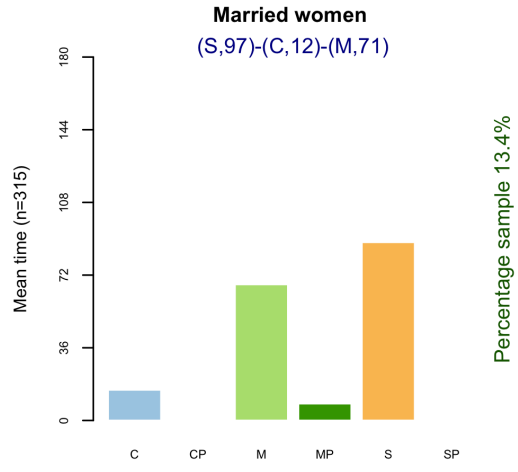
Typologies of trajectories

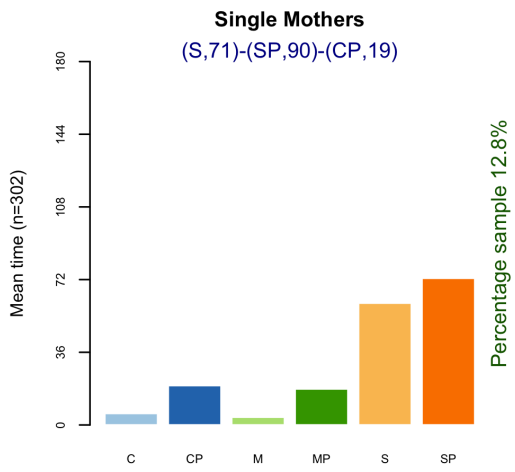


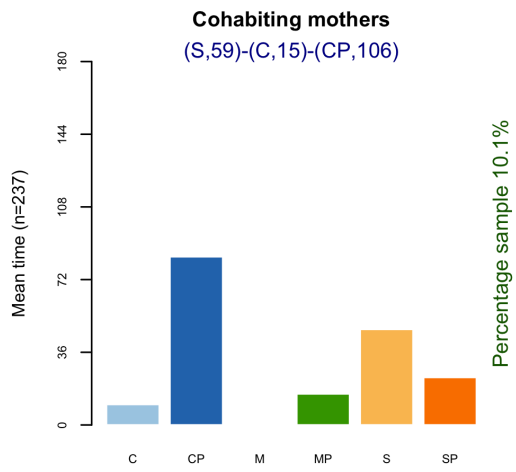
Typologies of trajectories

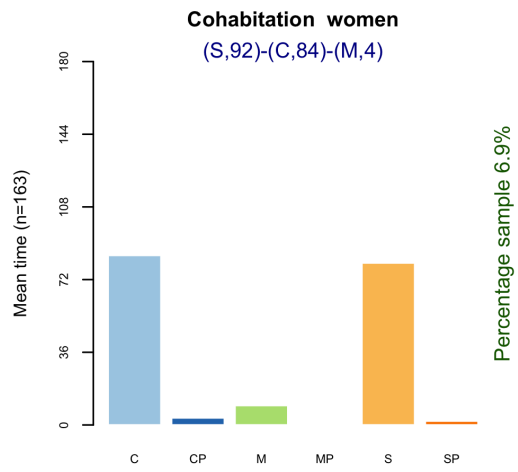












Descriptive statistics

	Married mothers	Late transitions	Married women	Single Mothers	Cohabiting mothers	Cohabiting women
			Union status and parenthood			
Ever married	1.00	0.37	1.00	0.51	0.44	0.44
Ever cohabited	0.70	0.71	0.72	0.83	1.00	1.00
Children	1.00	0.19	0.40	1.00	1.00	0.21
			Age at first transitions			
Age at first transition $<$18	0.58	0.12	0.32	0.63	0.79	0.46
Age at first transition 19-22	0.21	0.09	0.26	0.26	0.19	0.19
Age at first transition 23-25	0.21	0.38	0.42	0.11	0.02	0.36
Age at first transition $>$25	0.00	0.41	0.00	0.00	0.00	0.00
			Quantum and sequencing indicators			
Number of transitions Weave I-IV	3.37	2.41	3.14	3.89	3.79	3.32
Normative transitions	1.78	0.53	1.60	0.74	0.54	0.65
Non-normative transitions	1.59	1.88	1.54	3.15	3.25	2.67
			Compositional characteristics			
Proportion Black	0.1	0.18	0.06	0.34	0.31	0.14
Parents with college degree	0.19	0.27	0.38	0.15	0.07	0.22
Living with parents	0.49	0.56	0.63	0.29	0.26	0.52
Income family W1 (1000$)	41.54	51.92	53.52	33.38	34.59	41.73
Sex before 16	0.38	0.23	0.22	0.43	0.56	0.31

Descriptive statistics

	Married mothers	Late transitions	Married women	Single Mothers	Cohabiting mothers	Cohabiting women
			Union status and parenthood			
Ever married	1.00	0.37	1.00	0.51	0.44	0.44
Ever cohabited	0.70	0.71	0.72	0.83	1.00	1.00
Children	1.00	0.19	0.40	1.00	1.00	0.21
			Age at first transitions			
Age at first transition $<$18	0.58	0.12	0.32	0.63	0.79	0.46
Age at first transition 19-22	0.21	0.09	0.26	0.26	0.19	0.19
Age at first transition 23-25	0.21	0.38	0.42	0.11	0.02	0.36
Age at first transition $>$25	0.00	0.41	0.00	0.00	0.00	0.00
			Quantum and sequencing indicators			
Number of transitions Weave I-IV	3.37	2.41	3.14	3.89	3.79	3.32
Normative transitions	1.78	0.53	1.60	0.74	0.54	0.65
Non-normative transitions	1.59	1.88	1.54	3.15	3.25	2.67
			Compositional characteristics			
Proportion Black	0.1	0.18	0.06	0.34	0.31	0.14
Parents with college degree	0.19	0.27	0.38	0.15	0.07	0.22
Living with parents	0.49	0.56	0.63	0.29	0.26	0.52
Income family W1 (1000$)	41.54	51.92	53.52	33.38	34.59	41.73
Sex before 16	0.38	0.23	0.22	0.43	0.56	0.31

Descriptive statistics (2)

	Married mothers	Late transitions	Married women	Single Mothers	Cohabiting mothers	Cohabiting women
			Health status at Weave I			
Prop. in poor health at WI	0.10	0.08	0.07	0.16	0.10	0.14
Prop. with depression symptoms at WI	0.25	0.23	0.22	0.28	0.29	0.35
Smoking at WI	0.39	0.42	0.35	0.42	0.50	0.46
Heavy drinking at Weave I	0.34	0.39	0.34	0.31	0.32	0.47
			Health status at Weave IV			
Prop. in poor health at WIV	0.09	0.08	0.10	0.13	0.12	0.14
Prop. with depression symptoms at WIV	0.16	0.15	0.13	0.17	0.26	0.23
Smoking at WIV	0.30	0.29	0.20	0.39	0.43	0.37
Heavy drinking at WIV	0.29	0.43	0.38	0.33	0.35	0.52

Methods

Lagged dependent variable model

$$Y_{i2} = \gamma D_i + \rho Y_{i1} + \beta_i X_{i1} + \epsilon_i 2 \tag{1}$$

where:

- Y_{i2} is vector of health indicators measured at Wave IV (Time 2)
- Y_{i1} represents a vector of identical health measures at Wave I (Time 1)
- D_i represents the characteristics of the sequence from Wave I to Wave IV.
- X_{i1} a vector of demographic controls and SES background at Wave I (race; parents' education;)

Results (2)

	(1) Poor Health	(2) Depression	(3) Smoking	(4) Drinking
Late transitions (ref. category)				
Married mother w/o cohabitation	0.105	-0.053	-0.053	-0.715***
	(0.067)	(0.185)	(0.194)	(0.171)
Married women w/o cohabitation	0.086	0.091	-0.416	-0.343
	(0.076)	(0.242)	(0.255)	(0.203)
Single mothers	0.233**	-0.047	0.409	-0.370
	(0.087)	(0.262)	(0.238)	(0.231)
Cohabiting mothers w/o marriage	0.196*	0.672*	0.402	-0.258
	(0.092)	(0.325)	(0.242)	(0.252)
Cohabitation w/o children	0.234*	0.744*	0.257	0.265
	(0.114)	(0.317)	(0.349)	(0.266)
Self-reported health at wave I	0.250***	0.099	0.077	0.022
	(0.030)	(0.083)	(0.083)	(0.074)
Depression WI	0.021***	0.144***	0.003	-0.030*
	(0.006)	(0.016)	(0.018)	(0.014)
Smoking at wave I	0.084	0.099	1.952***	0.422**
	(0.055)	(0.174)	(0.158)	(0.145)

Discussion

- Early childbearing and long cohabitation are associated with negative health outcomes
- Protection effects of marriage on health behaviors
- Cohabitation seems to have no negative effect if short and followed by marriage

Sequence analysis?

Sequence analysis

- seems to be an effective tool for investigating association between trajectories and other outcome
- Attention from transitions to trajectories
- Controlling for previous outcomes is necessary for selection issues
- Not very useful for causality issues but help to highlight disadvantaged situations
- Study can be enlarged to other health indicators (biomarkers) and to multivariate analysis (i.e. How previous smoking affects later drinking or depression)