Introduction to sequence analysis (Lecture 2).

Describing and finding tipical trajectories

Nicola Barban

University of Bologna

January 20-21 2022

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Outline

What do we need?

What are the key ingredients for sequence analysis?

- The **state-space**, (i.e the alphabet from which sequences are constructed) has a finite number of elements and represents all the possible states that an individual can take in each time period.
- **Alphabet A**: a finite set of possible states, for example:
  1. live in the parental home
  2. live alone
  3. live with partner and no children
  4. live with partner and one or more children
  5. live with no partner and one or more children

Time axis

We have to decide the time reference.

- Years
- Months
- Hours in a day
- Years from marriage

Attention! some events can overlap if we have a large time reference. For example a person can cohabit and marry in the same year. What is the order? A possible solution is to set a state Cohabit-Marriage

Length of sequence

Sequence of length k: A ordered list of **k** elements taken from **A**

- Are sequence of the same length?
- Censoring and truncation

How to consider missing values?? with OM a common solution is to set a missing value * state

Categorical time series

- Each individual i can be associated to a variable s_{it} indicating her/his life course status at time t.
- As one can assume that s_{it} takes a finite number of values, trajectories can be described as categorical time series.
- More formally, let us define a discrete-time stochastic process $S_t : t \in T$ with state- space $\Sigma = \{\sigma_1, \ldots \sigma_K\}$ with realizations sit and $i = 1$. The life course trajectory of individual i is described by the sequence $s_i = \{s_{i1} \ldots s_{iT}\}$.

Levenshtein distance

- Information theory and computer science
- **The Levenshtein distance** between two strings is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being **insertion**, **deletion**, or **substitution** of a single character. It is named after Vladimir Levenshtein, who considered this distance in 1965.(Wikipedia)

Biology

- In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

(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)

Social sciences and order of events

- Interest in event sequence is not new.
- Hogan (1978), studied the temporal order of three events - finishing school, getting a first job, and getting married - in the transition to adulthood in the United States.
- He constructed a typology of typical (and normatively sanctioned) and atypical sequences.
- However, without appropriate tools Hogan could only consider a limited number of events, and the timing of event was not taken into account.
- i.e. Person **A** who finished school, got a job immediately, and got married within a year would be considered to belong to the same sequence-type as **B** who took a year to find his first job after finishing school, and stayed unmarried for the next ten years.

Optimal Matching in social sciences

- Andrew Abbott introduce the idea of sequence analysis in social science
- First works late 1980s
- Great intuition! setting **different substitution costs**.
- Not all transitions are equal!
- First applications with costs based on theory

OM

- A set that is composed of three basic operations $\Omega = \{\iota, \delta, \sigma\}$,
  1. ι denotes *insertion* (one state is inserted into the sequence)
  2. δ denotes *deletion* (one state is deleted from the sequence)
  3. σ denotes *substitution* (one state is replaced by another state).
- To each of these elementary operations $\omega_k \in \Omega$, a specific cost can be assigned, $c(\omega_k)$.
- If K basic operations must be performed to transform one sequence into another the transformation cost can be computed as $c(\omega_1, \ldots \omega_K) = \sum_{k=1}^{K} c(\omega_k)$.

Cristicisms

- Operation do not have a sociological meaning
- Cost specification
- Robustness to censoring and truncation
- Time warping. We are comparing events on different time (What is the meaning of substitution?)
- Pure algorithmic method. What is the generating mechanism?
- How to take into account different life domains?
- Comparison to other techniques

“Second Wave” of sequence analysis

Costs specification

- Substitution is equivalent to a deletion followed by an insertion. $c(\sigma) = 2c(\delta)$
- Cost specification can be totally subjective or derived by data
- A possible strategy is to set substitution costs as inversely proportional to transition probabilities

The transition frequency from a to b is:

$$p_{t,t+1}(a,b) = \frac{\sum_{t=1}^{T-1} N_{t,t+1}(a,b)}{\sum_{t=1}^{T-1} N_t(a)} \tag{1}$$

The cost of substituting a for b is
$c(\sigma; a, b) = c(\sigma; b, a) = 2 - p_{t,t+1}(a,b) - p_{t,t+1}(b,a)$ if $a \neq b$.
This cost specification takes into account the occurrence of events, giving more weight to those transitions that are less frequent

Longest Common Subsequences

- In mathematics, a subsequence is a sequence that can be derived from another sequence by deleting some elements without changing the order of the remaining elements. **For example, ABD is a subsequence of ABCDE**
- The longest common subsequence (LCS) problem is to find the longest subsequence common to two sequences.
- *The longer is the common subsequence the more similar are two strings*

Other particular subsequences

- **LCP** (Longest Common Prefix)
- **RLCP** (reversed LCP, i.e. Longest Common Suffix)
- **LCS** (Longest Common Subsequence),

Example

SATURDAY $\longrightarrow$ SUNDAY

Example

S	A	T	U	R	D	A	Y	
S			**U**		**D**	**A**	**Y**	**Length=5**
S			U	N	D	A	Y	

Example

SATURDAY $\longrightarrow$ SUNDAY

Example

S	A	T	U	R	D	A	Y	
					D	**A**	**Y**	**Length=3**
		M	O	N	D	A	Y	

LCS

Try this!

Elzinga measures

- Elzinga developed several measures based on common subsequences.

Pros:

- No need of cost specification
- Works with truncation (without defining a “missing state”)
- Can compare sequences of different measure

Cons:

- Less and less sociological meaning

Hamming distance

- Insertion and Deletion modify the time scale (Time warping)
- i.e. deleting sequence, we compare transition at different age!

Possible solutions:

- **Hamming Distance (1950)** Only substitutions. Measure the number of substitutions (costs may different).
- **Dynamic Hamming Distance (Lesnard)** Specific substitution costs at each position, i.e. Transitions at different ages have different costs.

Note that HAM and DHD apply only to sequences of equal length.

Multichannel optimal matching

- Introduced by Pollock, 2007
- takes into account multiple trajectories simultaneously
- specify costs for different domains

Multichannel optimal matching

Table 2. Substitution cost matrices†

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Employment status										
Self employed	1	0								
Employed	2	0.6	0							
Unemployed	3	1.4	0.8	0						
Retired	4	1.2	1.2	1.2	0					
Maternity leave	5	1.4	0.6	1.4	1.4	0				
Family care	6	1.2	0.8	1.2	0.8	1	0			
Full-time student	7	1.4	0.6	1	1.4	2	1.4	0		
Long term sick	8	1.4	1.4	1.2	0.8	2	1.2	1.4	0	
Government training	9	1.4	0.8	1	1.4	2	1.4	1.4	1.4	0
Housing tenure										
Own outright	1	0								
Own with mortgage	2	0.6	0							
Local authority rent	3	1.4	0.8	0						
Housing association rent	4	1.4	1	0.8	0					
Rent from employer	5	1.4	1	1.4	1.4	0				
Private rent (unfurnished)	6	1.4	1	1.4	1	0.8	0			
Private rent (furnished)	7	1.4	0.8	1.4	1.4	1.2	1	0		
Marital status										
Married	1	0								
Separated	2	1	0							
Divorced	3	0.5	0.6	0						
Widowed	4	0.5	1.4	2	0					
Never married	5	0.5	2	2	2	0				
Responsibility for children aged under 16 years										
Yes	1	0								
No	2	0.5	0							

†The cost of insertion or deletion is 1; hence a substitution cost of 2 will invoke an indel instead of a substitution. This means that there is no need to create a substitution cost which is greater than the sum of an insertion and a deletion.

Multichannel optimal matching

- Alghoritm combine sequences in different domains
- substitution cost equal to sum of different substitutions

OMv

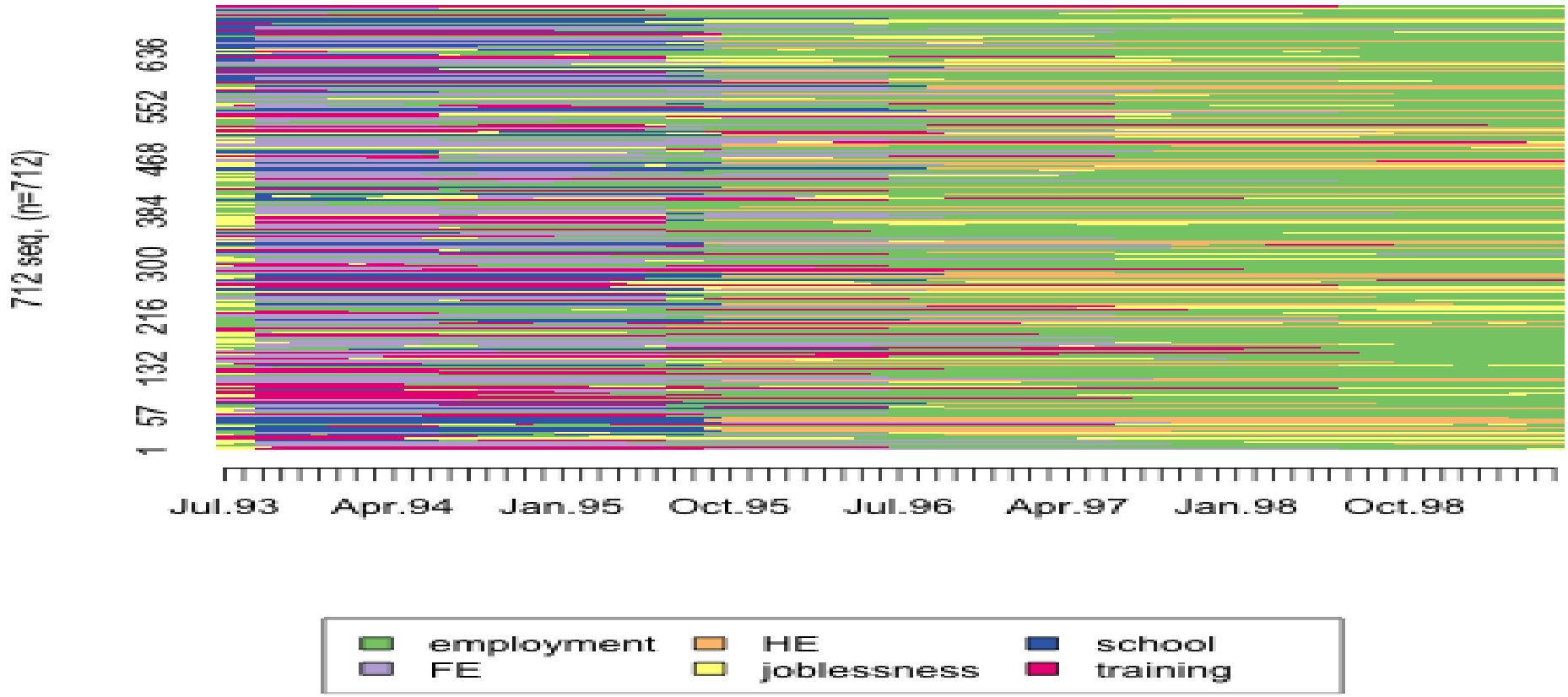
- Recent development by Halpin (2010)
- Modified version of OM, weighting elementary operations inversely with episode length.
- Pairwise distances much lower than the standard algorithm, the more the sequences are composed of long spells in the same state

types of indicators

1. Frequencies of trajectory
2. Transversal statistics
3. Longitudinal statistics

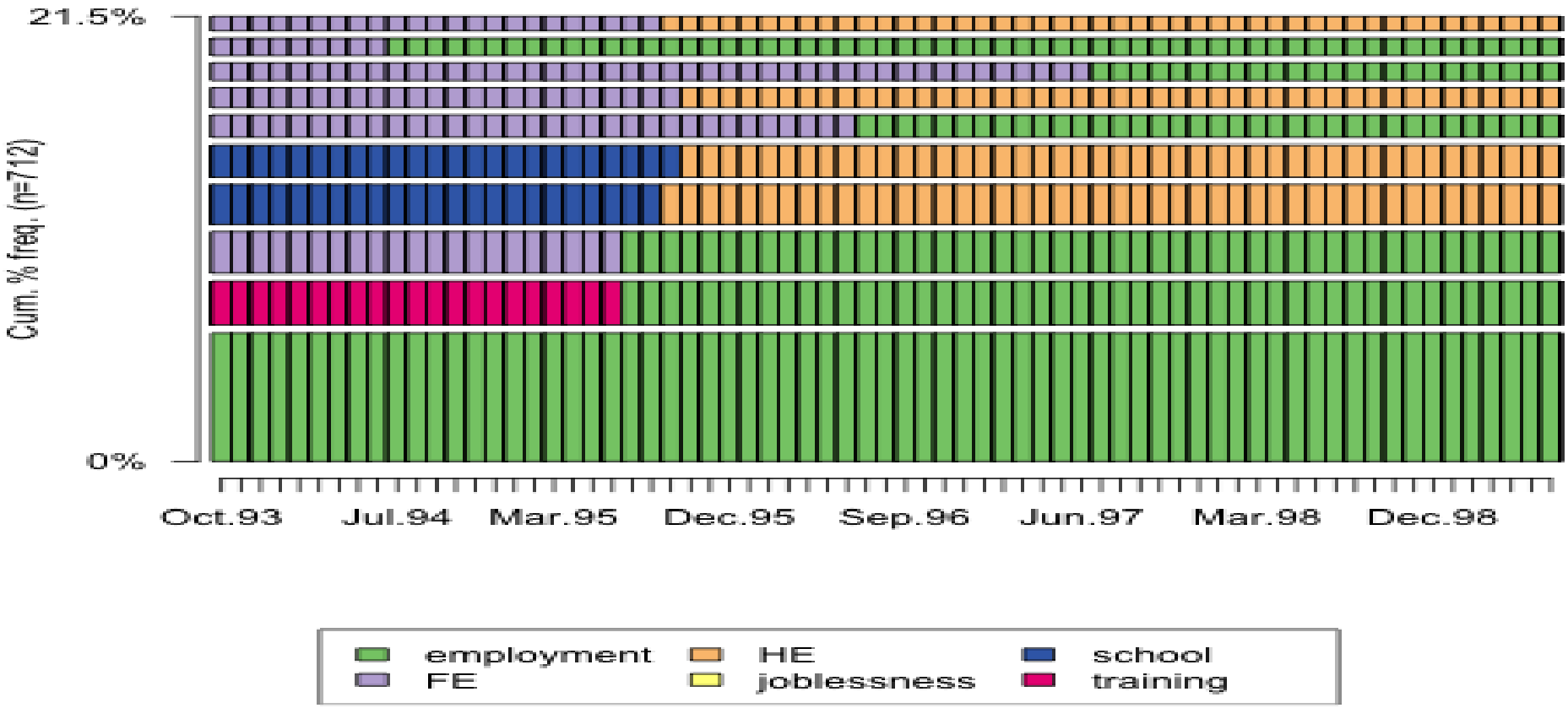
Frequencies

Often difficult to read when we have many trajectories



Frequencies

Better to look at the most common seq



Freq. table

	Freq	Percent
employment/69	52	7.30
training/21-employment/48	18	2.53
FE/21-employment/48	17	2.39
school/23-HE/46	16	2.25
school/24-HE/45	13	1.83
FE/33-employment/36	9	1.26
FE/24-HE/45	8	1.12
FE/45-employment/24	7	0.98
FE/9-employment/60	7	0.98
FE/23-HE/46	6	0.84

First ten common patterns

If we don't consider the time spent on each state the frequency of trajectories can change!

	Freq	Percent
training-employment	53	7.4
employment	40	5.6
employment-FE-employment	27	3.8
school-HE	27	3.8
FE-employment	25	3.5
joblessness-FE-employment	18	2.5
joblessness-training-employment	16	2.2
school-FE-employment	15	2.1
joblessness-FE-joblessness-employment	11	1.5
school-employment	11	1.5

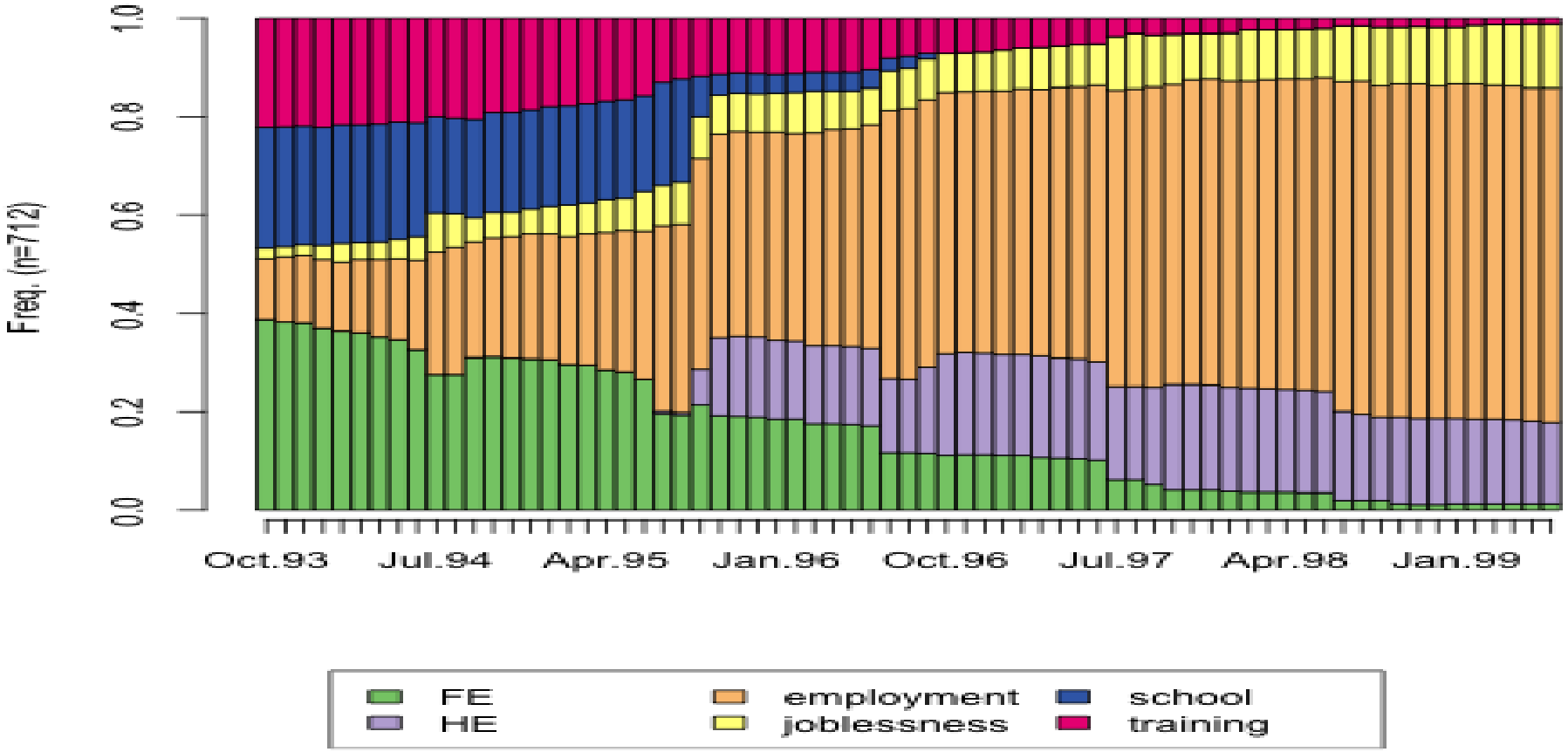
Trasversal

We can calculate the distribution on each state for the entire traj.

	Oct.93	Nov.93	Dec.93	Jan.94	Feb.94
employment	0.12359551	0.13342697	0.13764045	0.14044944	0.14044944
FE	0.38764045	0.38202247	0.38061798	0.36938202	0.36376404
HE	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
joblessness	0.02106742	0.01966292	0.02106742	0.02808989	0.03792135
school	0.24578652	0.24438202	0.24157303	0.24016854	0.24157303
training	0.22191011	0.22050562	0.21910112	0.22191011	0.21629213

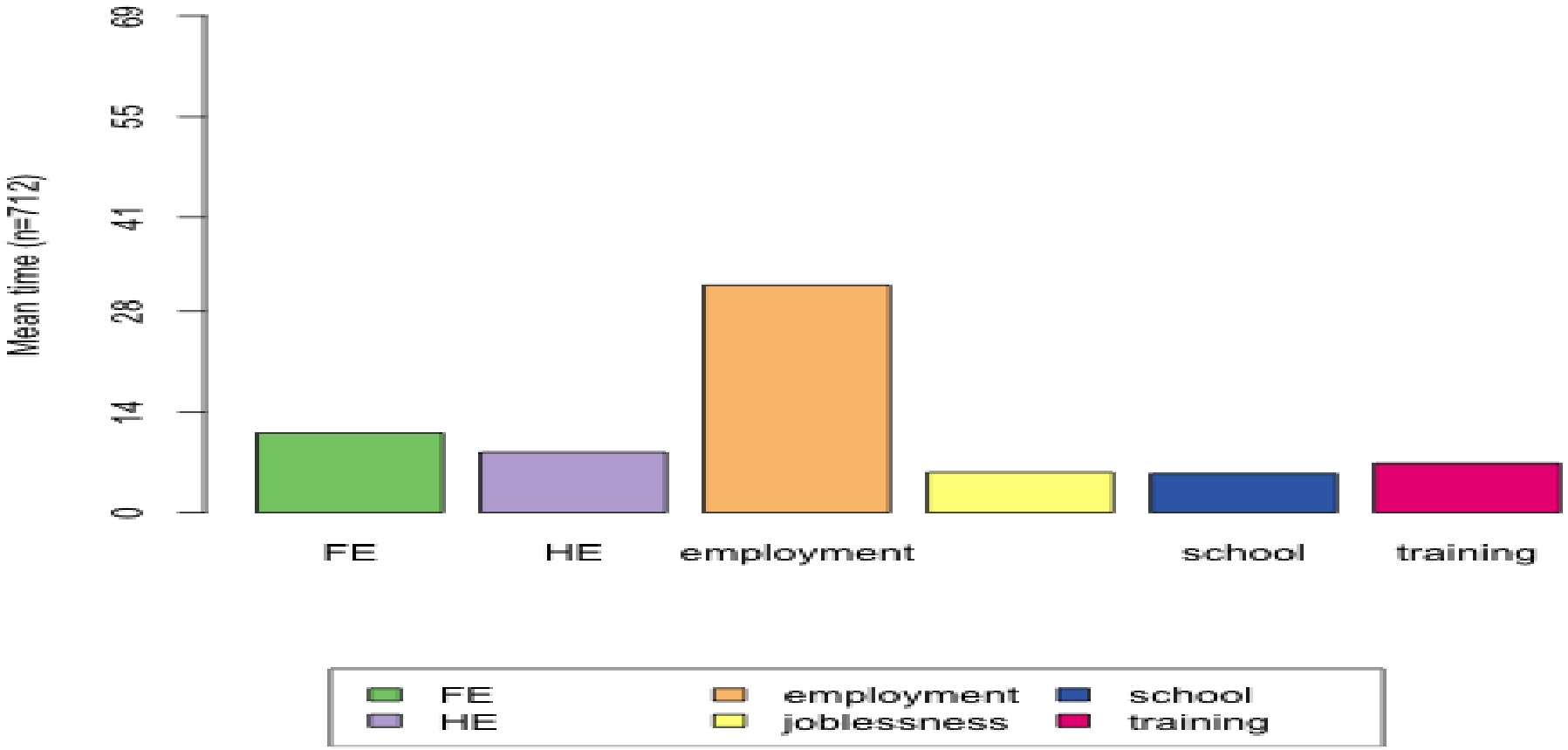
Sequence distribution

Or plot it!



Sequence time spent

Time spent in each state

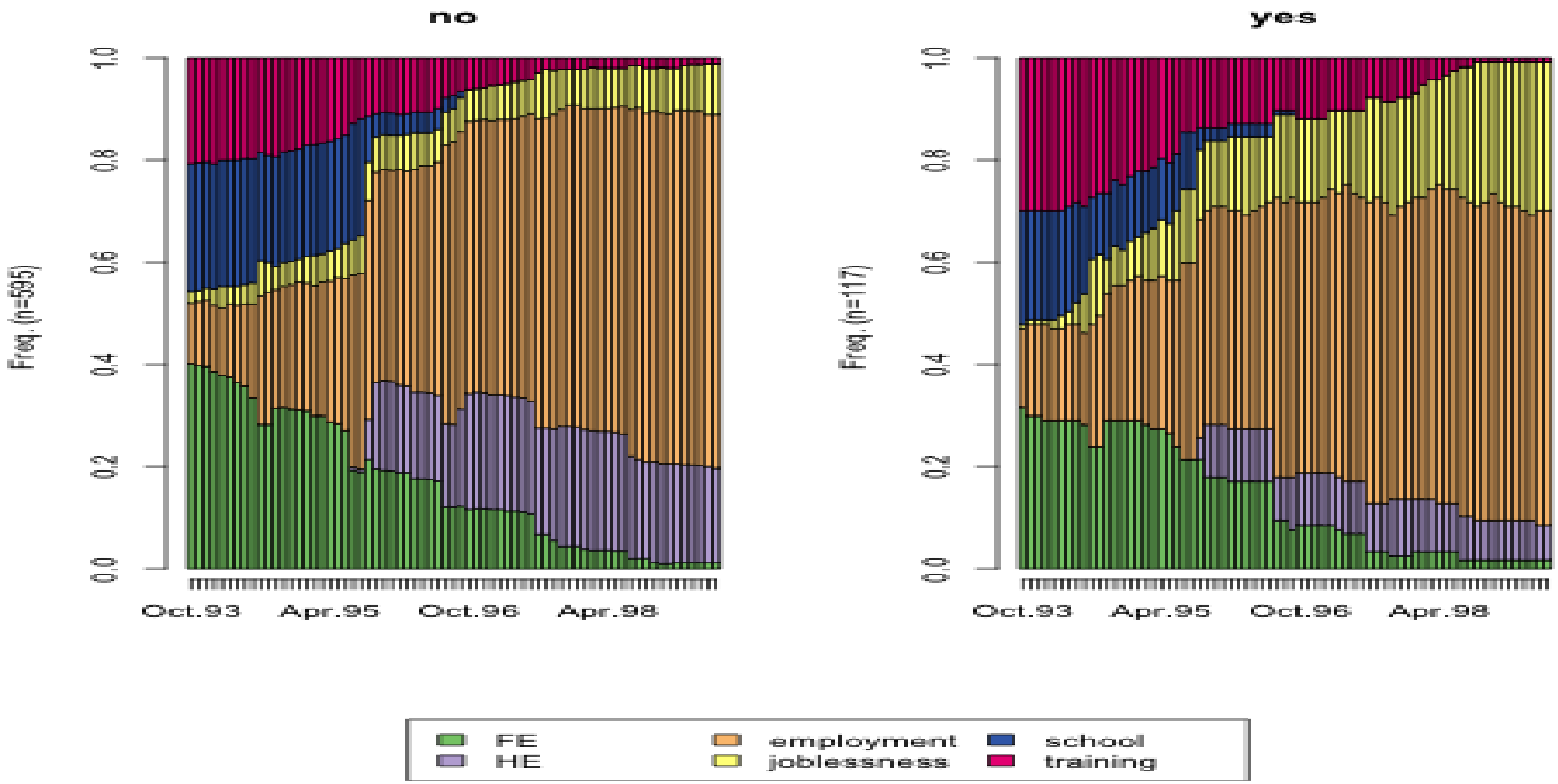


Bivariate graph

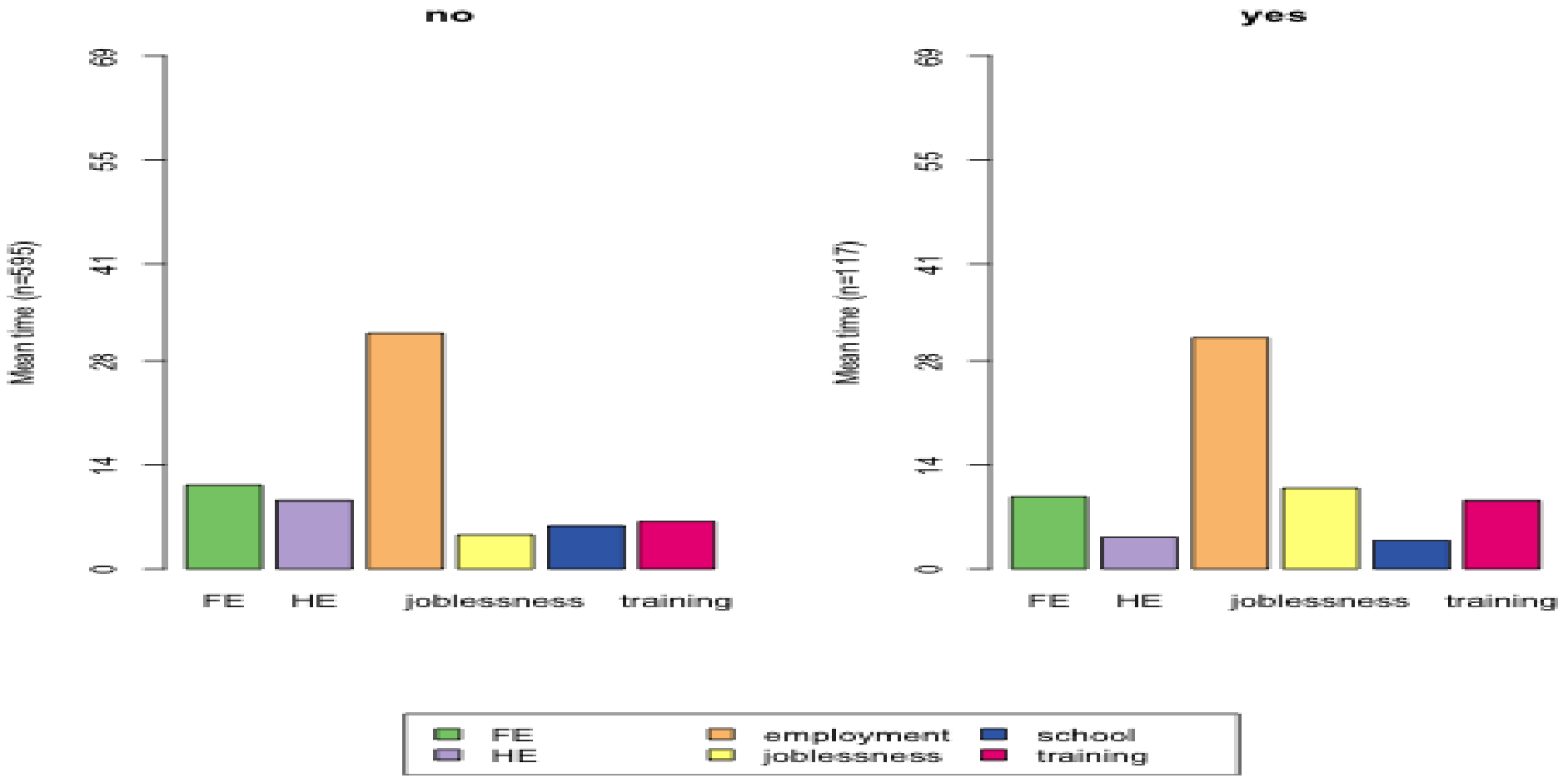
- We can analyze the previous indicators by group:
- Sex
- Cohort
- Education
- Treatment

Sequence distribution

variable `funemp` Father unemployment status



Sequence time spent

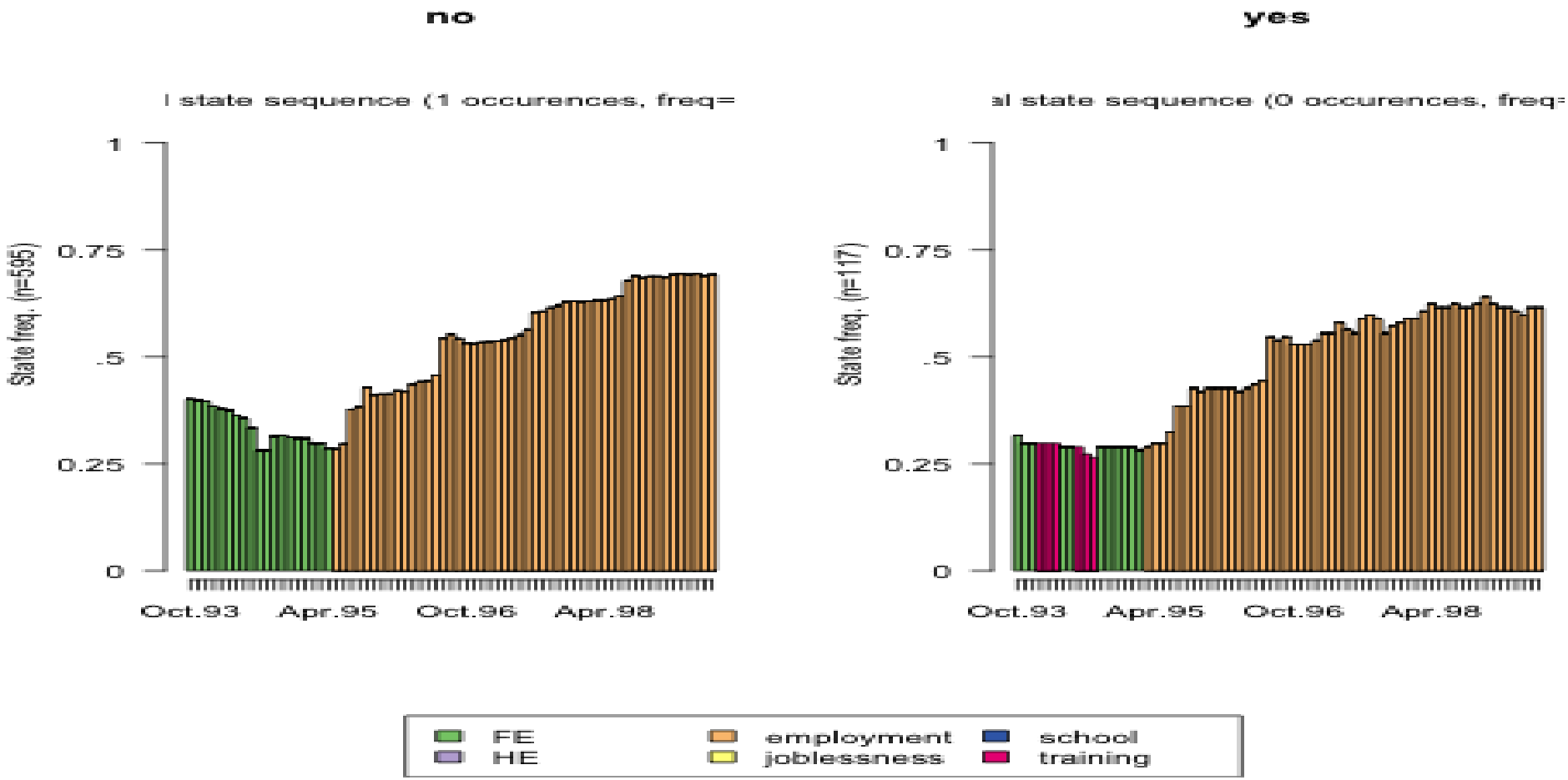


Transition rates

	employment	FE	HE	joblessness	school	training
employment	0.98637540	0.0019528589	0.002543258	0.006494391	0.0004087379	0.0022253508
FE	0.02878248	0.9502037697	0.006877229	0.009042282	0.0010188487	0.0040753948
HE	0.01023541	0.0001705902	0.987205732	0.001876493	0.0000000000	0.0005117707
joblessness	0.04147583	0.0081424936	0.002290076	0.939694656	0.0005089059	0.0078880407
school	0.01463039	0.0079568789	0.018993840	0.005646817	0.9496919918	0.0030800821
training	0.03911880	0.0035001029	0.000000000	0.014000412	0.0004117768	0.9429689109

Modal state

What is the most frequent state (by period)

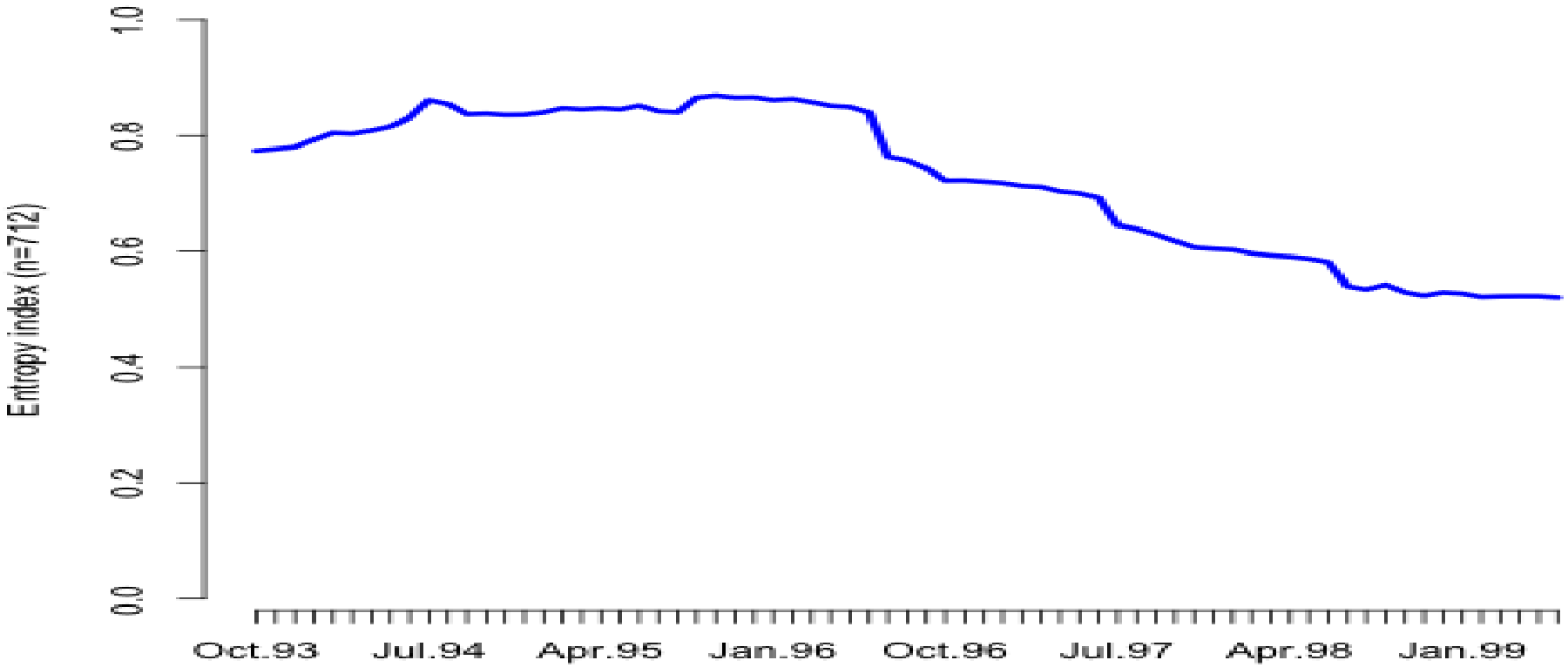


Entropy

- The Shannon entropy of the state distribution at each position i:
  $h(p_1; \ldots; p_s) = -\sum_{i=1}^{s} p_i log_2(p_i)$
- where p_i is the frequency of the i-th state and s is the size of the alphabet
- This indicator is called the entropy index
- It equals 0 when all cases are in the same state (it is thus easy to predict in which state an individual is)
- It is maximum when the cases are equally distributed between the states of the alphabet (it is thus hard to predict in which state an individual is)

Entropy graph

The series of of transversal entropies can be plotted



Longitudinal Statistics

- Number of transitions (*quantum*)
- Time before first transitions (*timing*)
- Number of specific transitions (*sequencing*)

Number of transition

We can compute the average number of transitions for each individuals

	yes	*no*
Catholics	2.063953	2.081522
Father Unemployed	2.145299	2.058824
Live with parents	2.117517	1.996169

Longitudinal entropy

- The entropy of the state distribution within a sequence is a measure of the diversity of its states.
- vector of the longitudinal Shannon entropies of the sequences, i.e. for each sequence the entropy $h(\pi_1; \ldots; \pi_a) = -\sum_{i=1}^{a} \pi_i log_2(\pi_i)$
- where a is the size of the alphabet and π_i the proportion of occurrences of the i-th state in the considered sequence

Other indicators

- **Turbolence**. Proposed by Elzinga and Liefbroer. Takes into account number of different subsequences and variance of time spent in each state.
- **Complexity**. Standardized indicator (0,1). Value 0 when sequence has no transitions and only one state. Value 1 when:
  - Each of the state in the alphabet is present in the sequence and the same time is spent in each of them
  - the number of transitions is equal to the length of the sequence -1

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

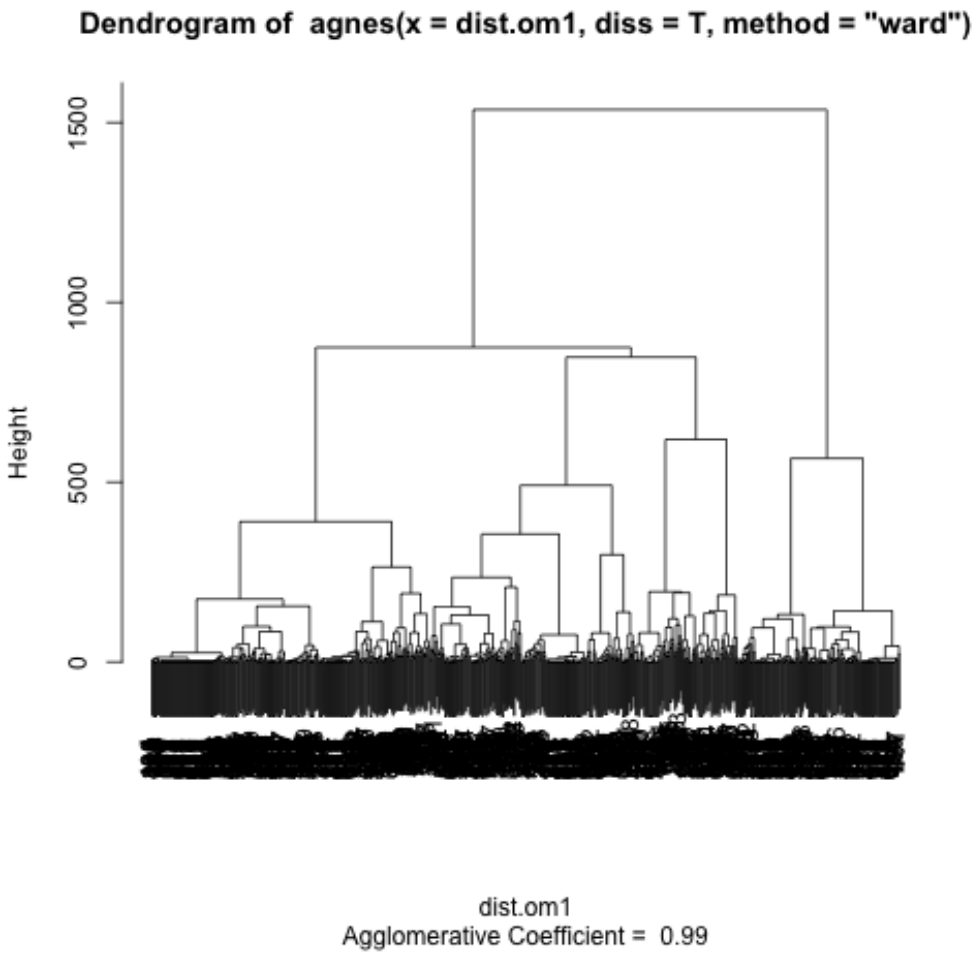
Ward clustering

- 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.

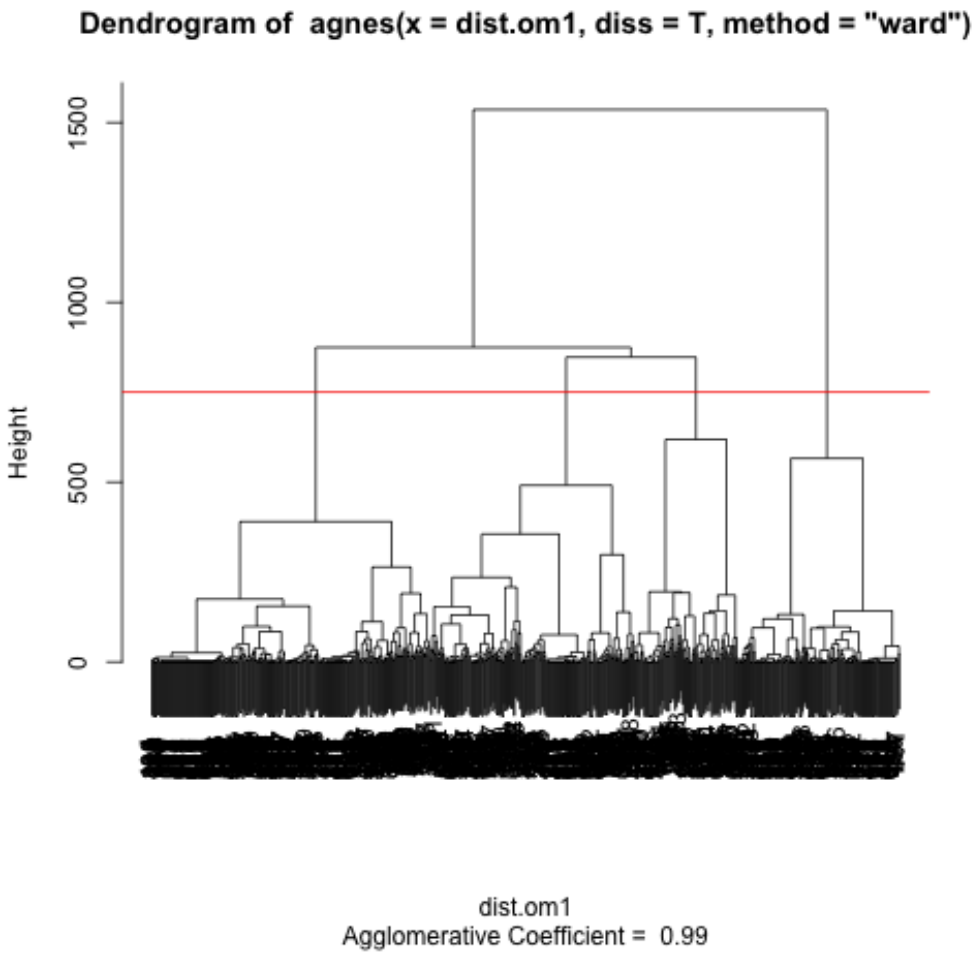
Number of clusters

- 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)
```