Extracting knowledge from life courses: clustering and visualization[1]

Nicolas S. Müller, Alexis Gabadinho, Gilbert Ritschard, Matthias Studer

Department of Econometrics, University of Geneva

10th International Conference on Data Warehousing and Knowledge Discovery, Torino 2008

[1] This study has been realized within the Swiss National Science Foundation project SNSF 100012-113998/1.

Outline

- Individual life course paradigm.
  - Following macro quantities (e.g. #divorces, fertility rate, mean education level, ...) over time
    insufficient for understanding social behavior.
  - Need to follow individual life courses.
  - The life course must be seen as a "whole", not only separate events
- Data availability for familial life courses
  - Large panel surveys in many countries
    (SHP, CHER, SILC, GGP, ...)
  - Biographical retrospective surveys (FFS, ...).
  - Statistical matching of censuses, population registers and other administrative data.

An example : my academic life

- My academic life as an example of life course
- In 2006, I receive a master in sociology
- In 2006, I begin working as a research assistant at the Department of Econometrics
- In 2007, I begin working as a teaching assistant at the Department of Econometrics (statistics for social sciences)
- In 2008, I receive a master in information systems
- This is why I'm here today, presenting you a study that is a mix of algorithms, statistics and sociology

What are we looking for

- We wanted to see how typical life courses evolved through the 20th century.
- We created a typology of familial life courses in order to verify some sociological hypotheses.
- We decided to use sequence analysis in order to be consistent with the life course paradigm.

How can we represent a life course?

Alternative views of Individual Longitudinal Data

Table: Time stamped events sequence

leaving home in 1970 | marriage in 1971 | first child in 1973

Table: State sequence view

year	1969	1970	1971	1972	1973
left home	no	yes	yes	yes	yes
is married	no	no	yes	yes	yes
has child	no	no	no	no	yes

To create a single sequence per individual, we define one state per combination of events that have occured or not

	LHome	marriage	childbirth	divorce
0	no	no	no	no
1	yes	no	no	no
2	no	yes	yes/no	no
3	yes	yes	no	no
4	no	no	yes	no
5	yes	no	yes	no
6	yes	yes	yes	no
7	yes/no	yes	yes/no	yes

The previous example can then be translated into a single sequence

Table: State sequence view

individual	1969	1970	1971	1972	1973
id1	0	1	3	3	6

Analysis of sequences

- Frequencies of given subsequences
  - Essentially event sequences.
  - Subsequences considered as categories $\Rightarrow$ Methods for categorical data apply (Frequencies, cross tables, log-linear models, logistic regression, ...).
- Markov chain models
  - State sequences.
  - Focuses on transition rates between states.
    Does the rate also depend on previous states?
    How many previous states are significant?
- **Optimal Matching**
  - Based on the Levenshtein distance (Edit distance between pairs of sequences)
  - State sequences
  - Allows the clustering of sequences.

Distances between sequences

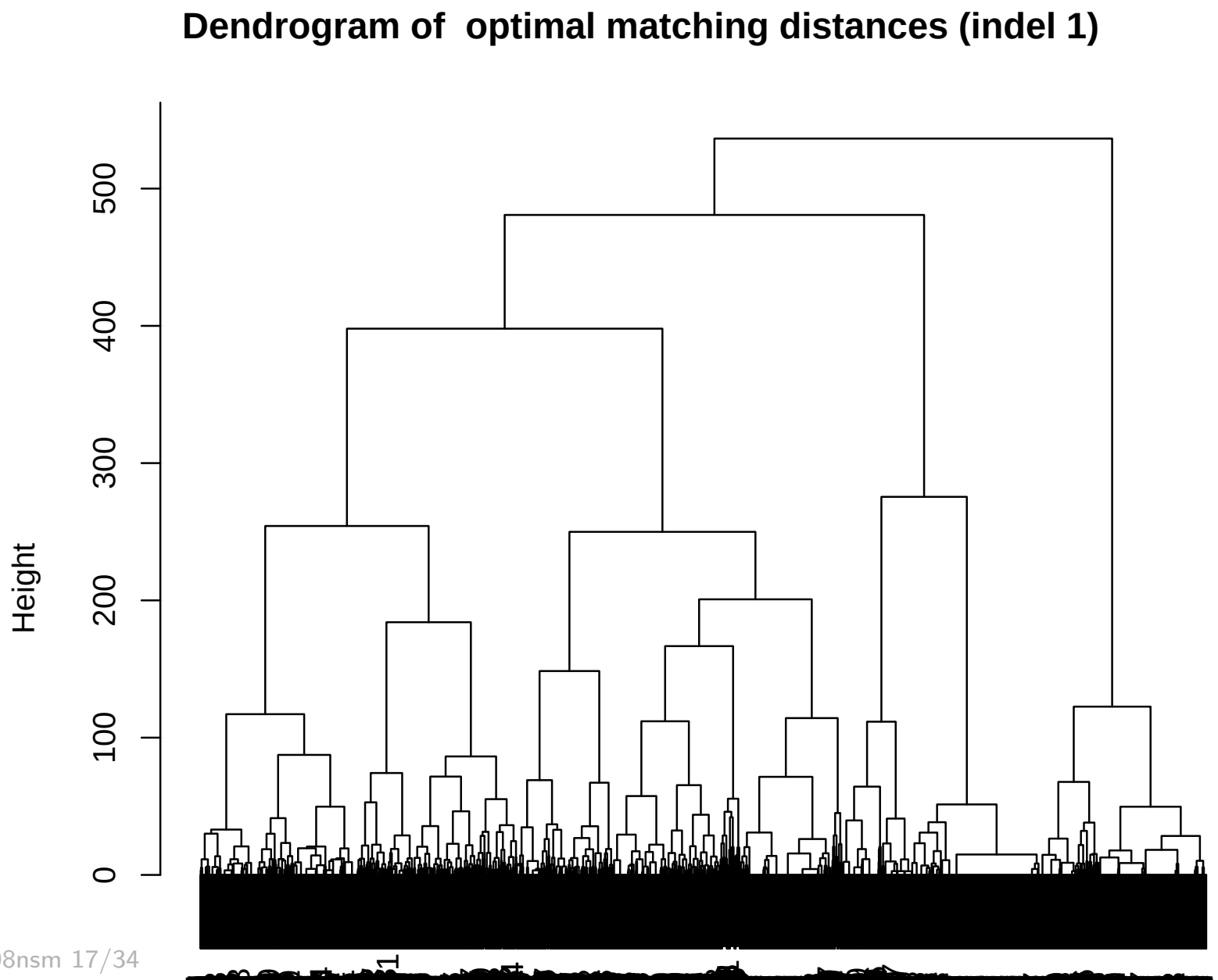
- Levenshtein distance (known as Optimal matching in Social sciences)
  - $d(x, y)$ Total cost of insert, deletion and substitution changes required to transform sequence x into y.
  - For example :
    - sequence x is "0-0-0-1-3" and sequence y is "0-0-1-1"
    - If a substitution op. costs 2 and an insertion costs 1, $d(x, y) = 3$ (inserts "3", substitute "0" by "1")
- Different solutions depending on indel and substitution costs.
- We can attribute specific substitution costs
- Details of the algorithm are in the paper (Needleman-Wunsch algorithm)

Presentation of the "BioFam" data

- Data from the retrospective survey conducted in 2002 by the Swiss Household Panel (SHP) (with support of Federal Statistical Office, Swiss National Fund for Scientific Research, University of Neuchatel.)
- Retrospective survey: 5560 individuals
- Retained familial life events: Leaving Home, First childbirth, First marriage and First divorce.
- Age 15 to 30 $\longrightarrow$ 4318 remaining individuals, born between 1909 et 1972.

Application to the familial life courses data

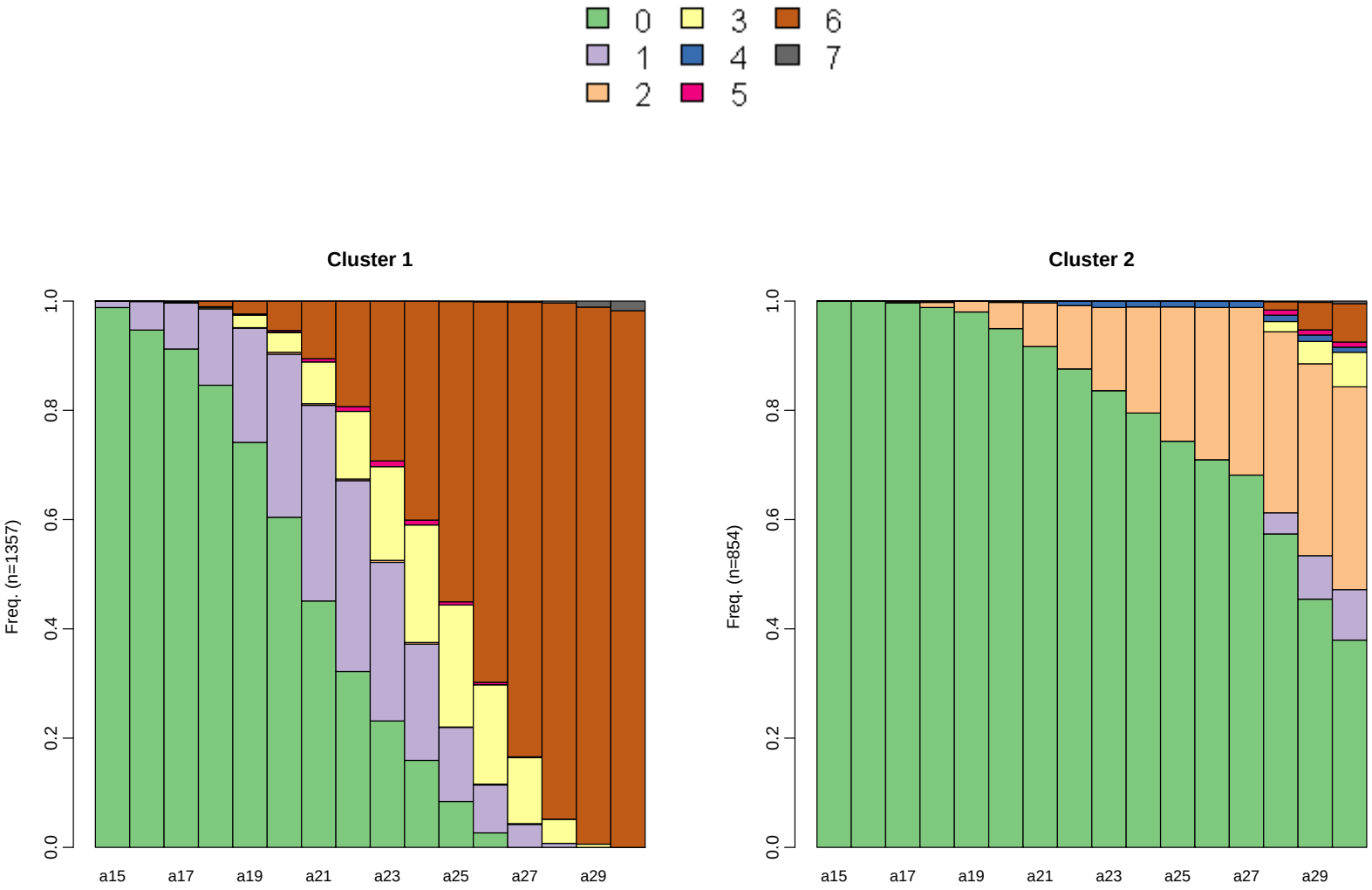
1. Creation of sequences of states
2. Optimal matching analysis
   - Indel were fixed at 1
   - Substitution costs were based on the rate of transition
     - $c[w(i,j)] = c[w(j,i)] = 2 - p(i_t|j_{t-1}) - p(j_t|i_{t-1})$
   - We compute the distance between each pair of sequences
3. Resulting distances matrix used in an agglomerative cluster analysis (Ward method)
4. Vizualisation and interpretation of the results with specific plots


Dendrogram of optimal matching distances (indel 1)
Height
0
100
200
300
400
500

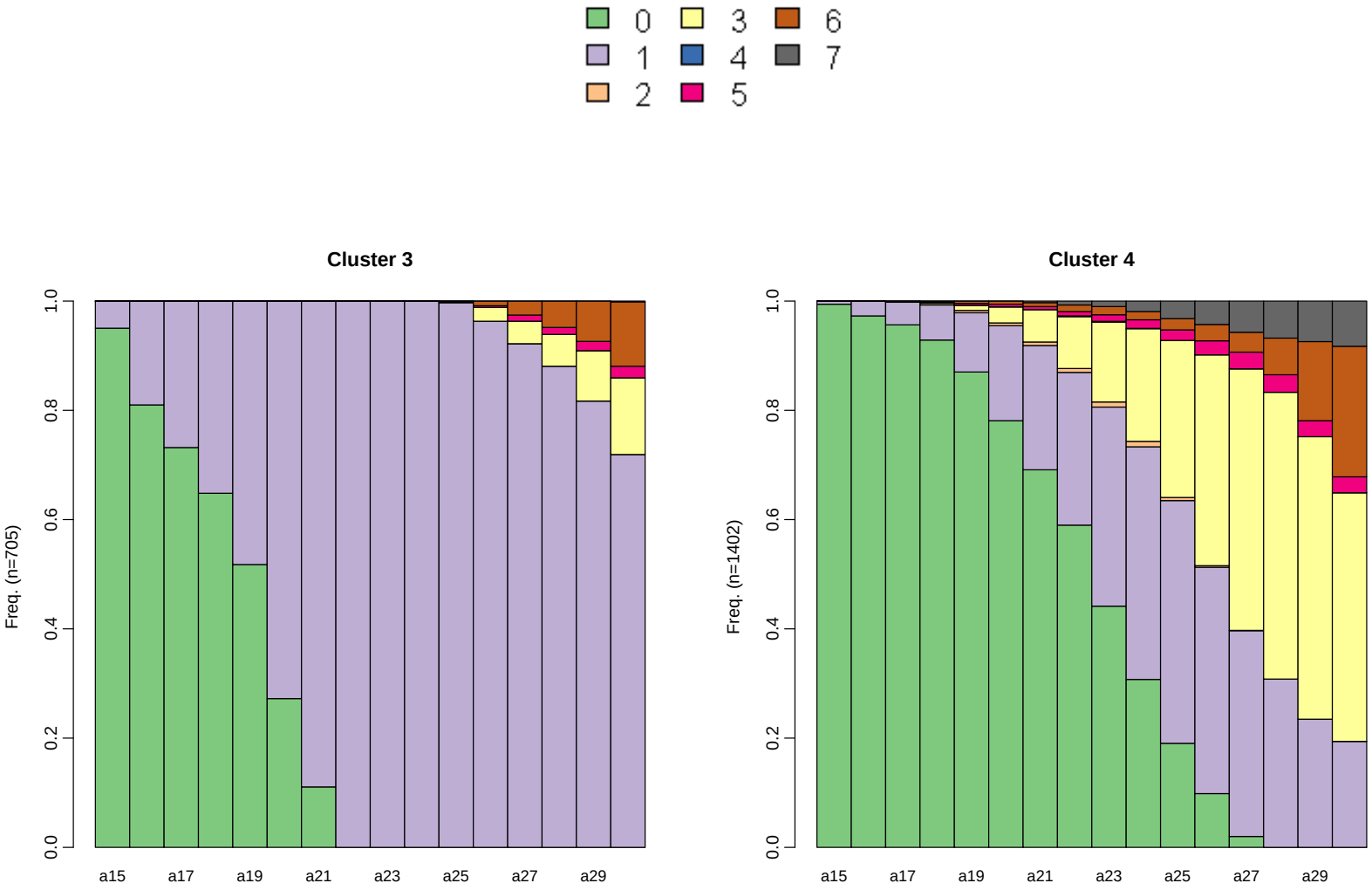
Description

- A density plot shows the proportion of individual in each state for each age
- It presents aggregated data, it is not really suitable for a life course interpretation

Density plots (1/2)



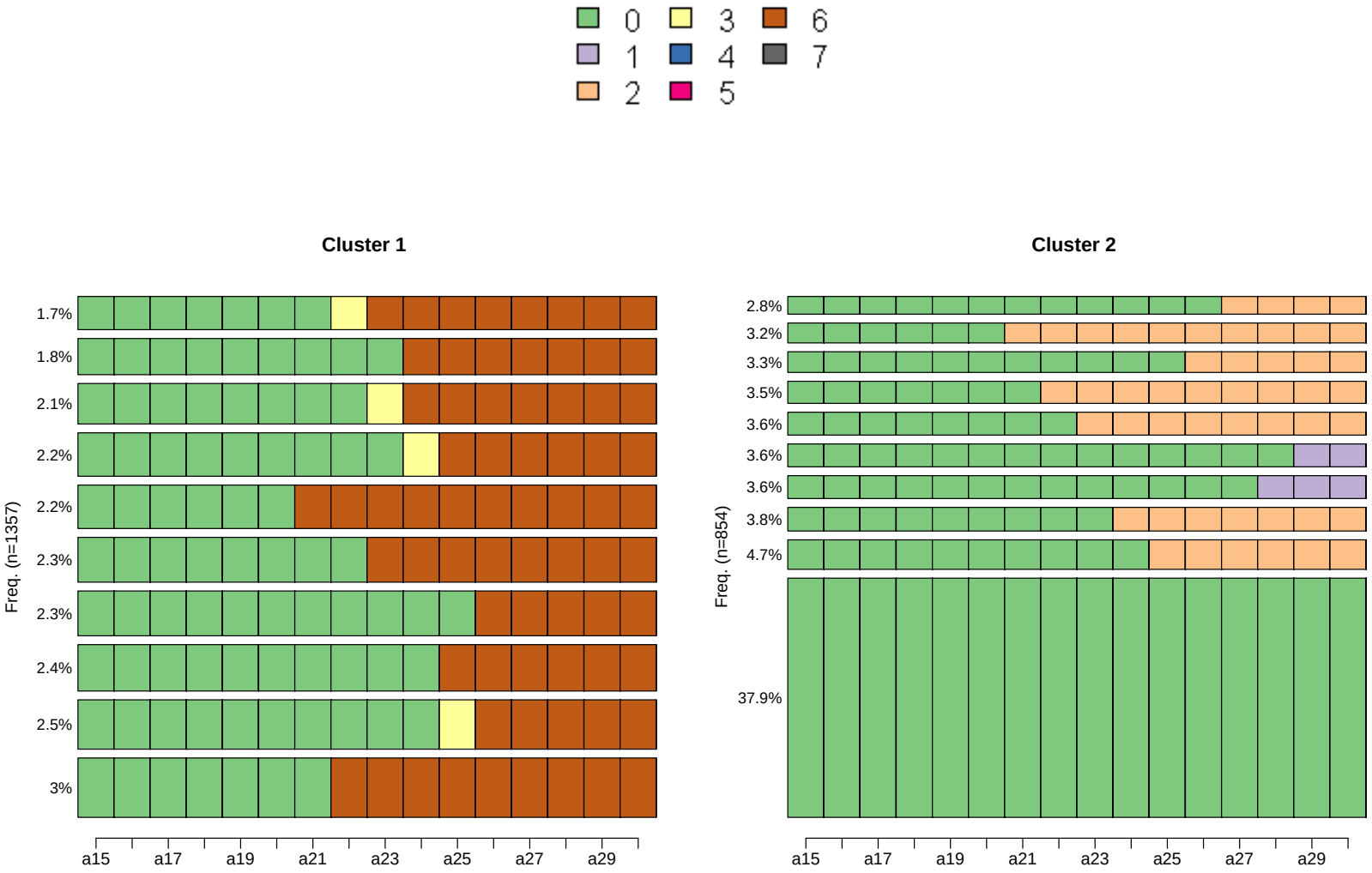
Density plots (2/2)



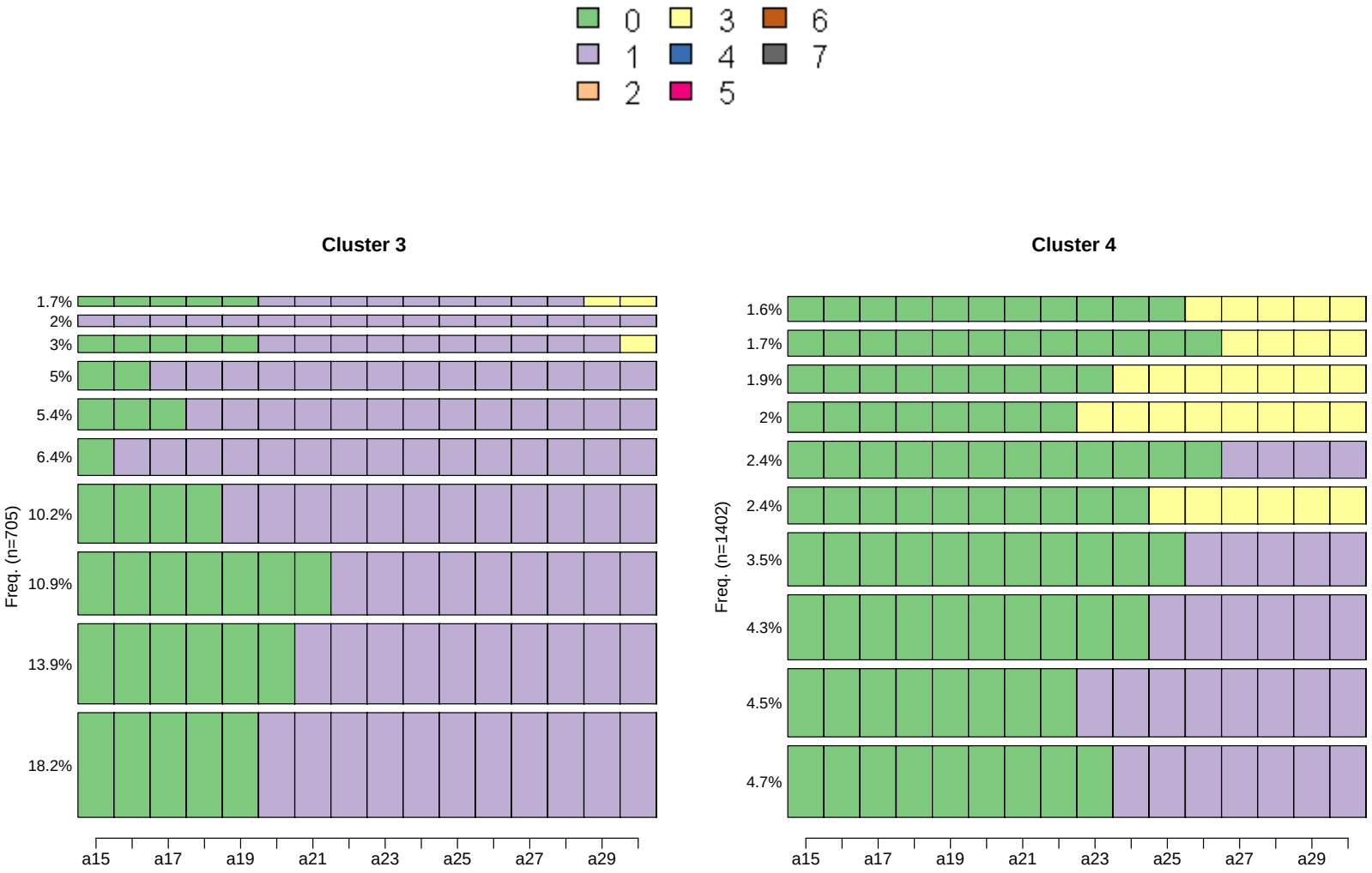
Frequency plots

- Plot of the n most frequent sequences.
- Individual life sequences are plotted
- The wider the bar representing the sequence, the more frequent it is

Frequency plots (1/2)



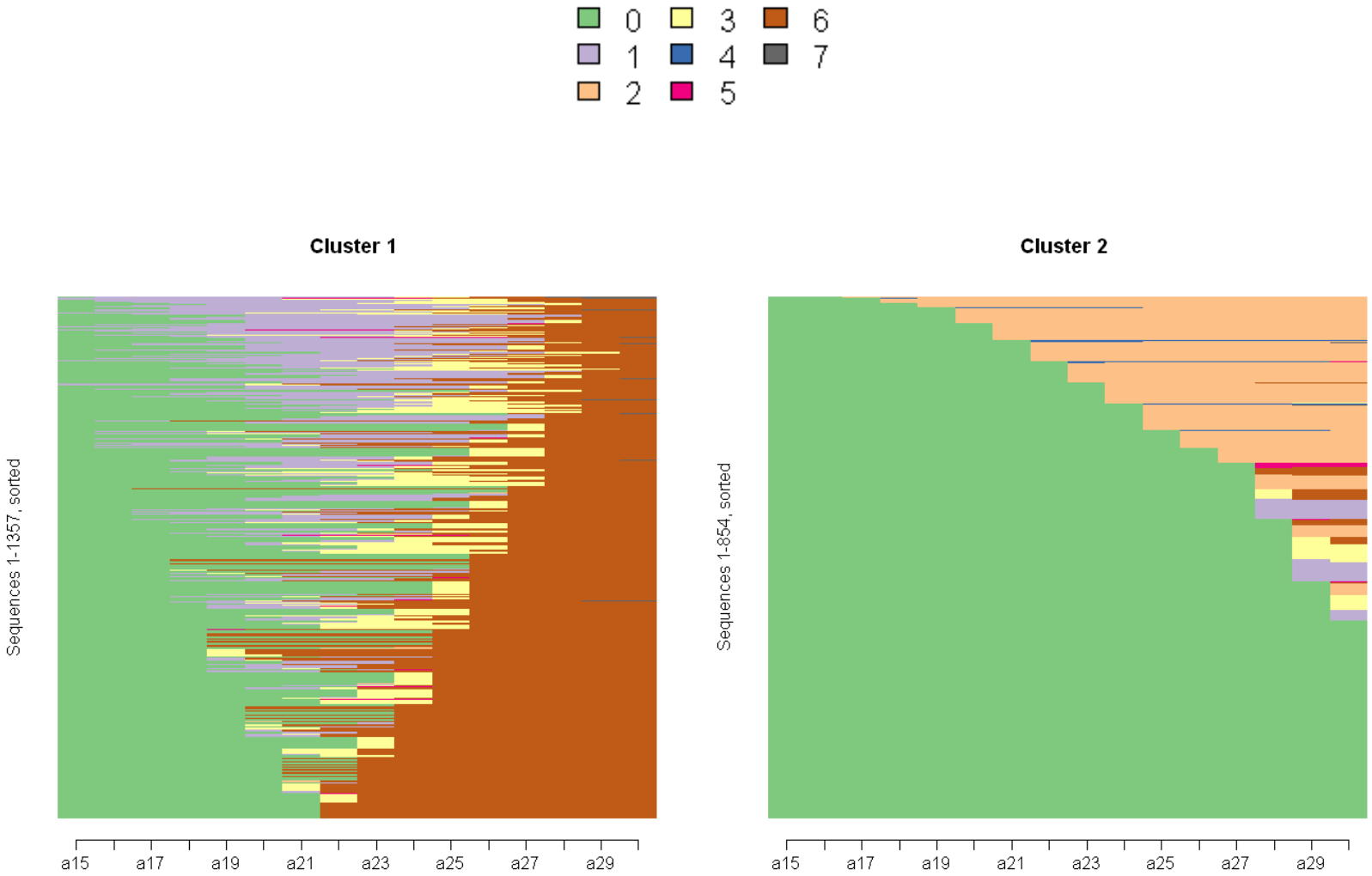
Frequency plots (2/2)



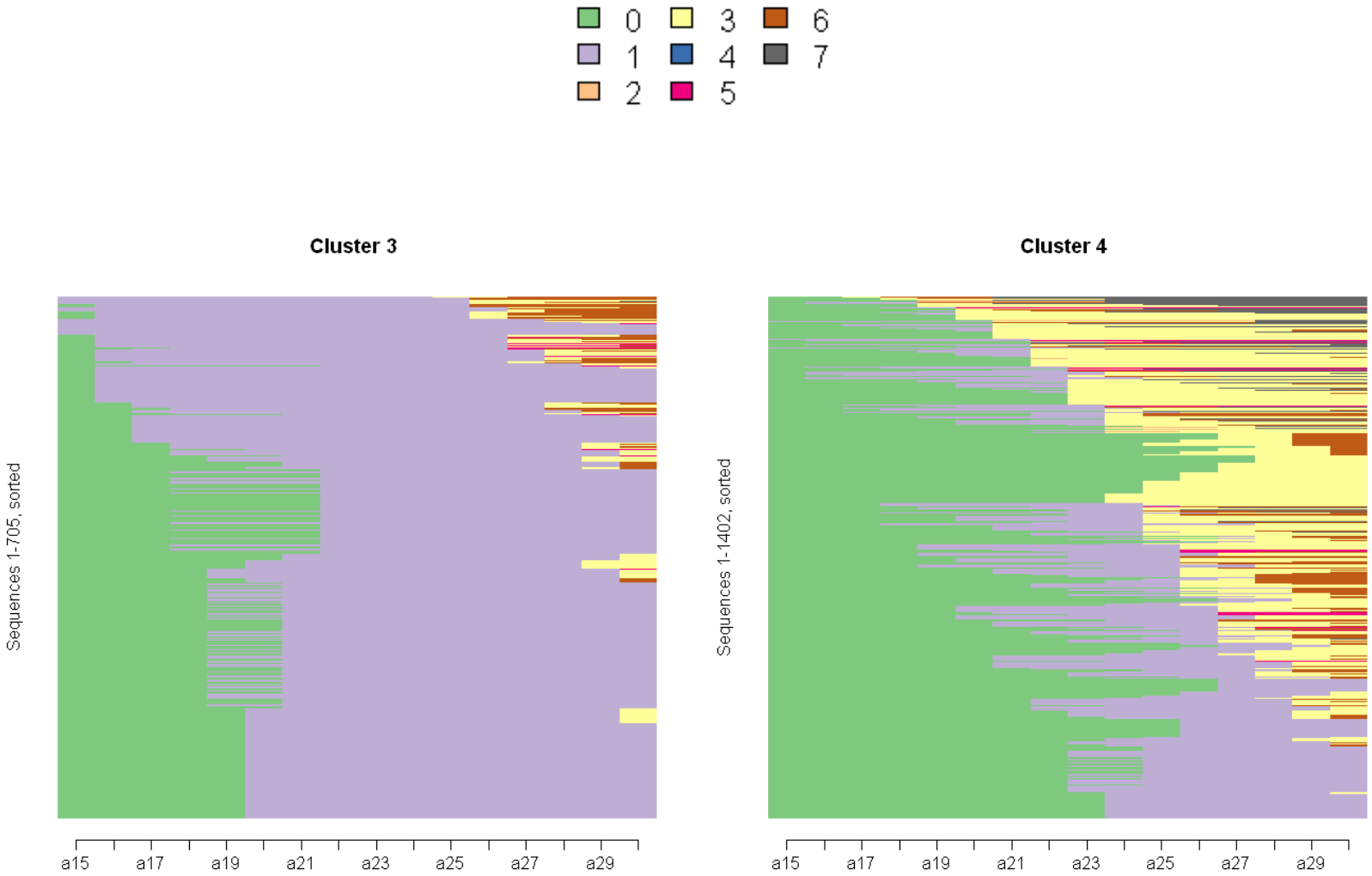
Index plots

- Each sequence represented by a stacked bar (or line)
- Plot n first sequences (not necessarily the most frequent)
- Sequences are sorted by their edit distance to the most frequent sequence
- Index plots of all sequences show diversity of the sequences.

Indexplots (1/2)



Indexplots (2/2)



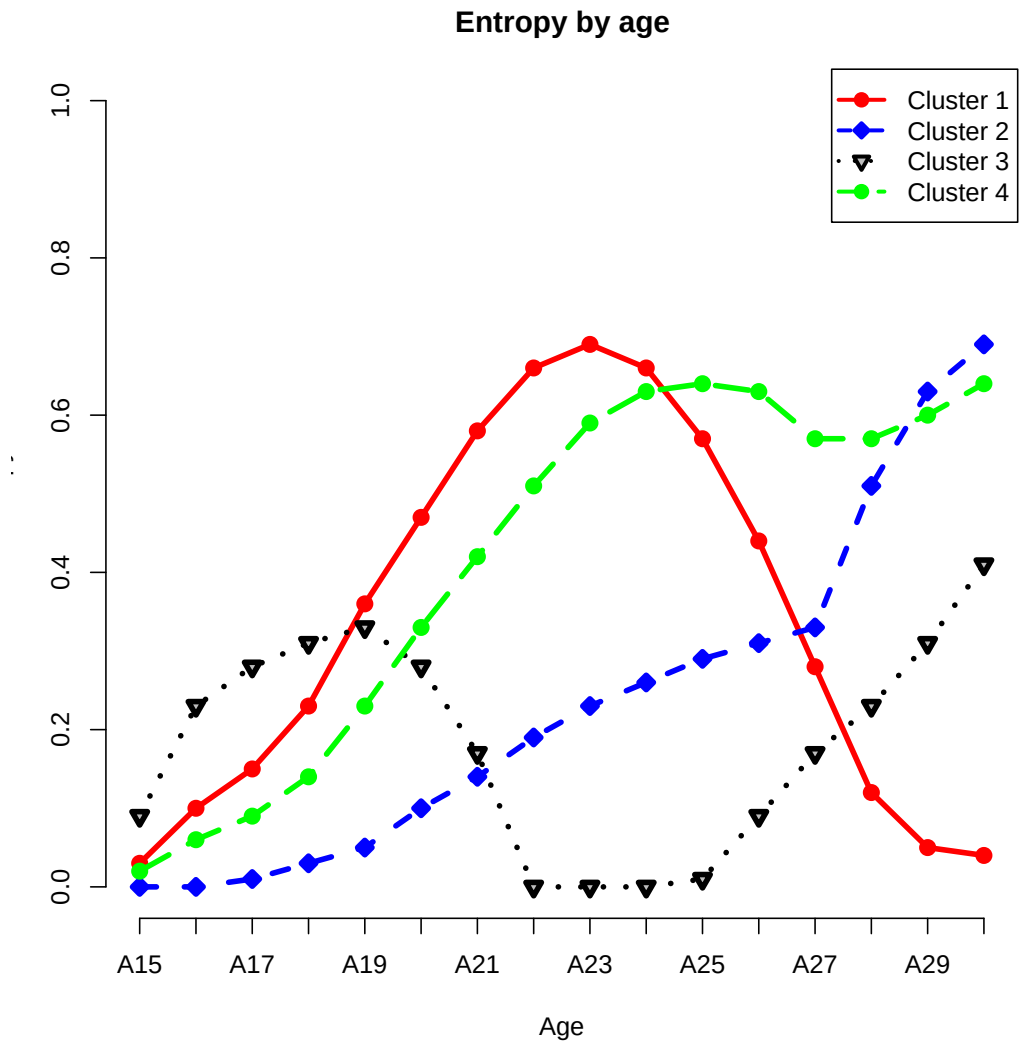
What can we learn from these clusters?

- Using logistic regression modelling, we can identify cohort and gender effects in the cluster membership.
  - For example, a woman has an odd ratio of almost 2 to be in cluster 1, meaning they have 2 times more chances to be in this cluster than a man
  - The same can be said about the birth year, the older the individual, the more chances he has to be in the cluster 1 ("classical" familial life courses)

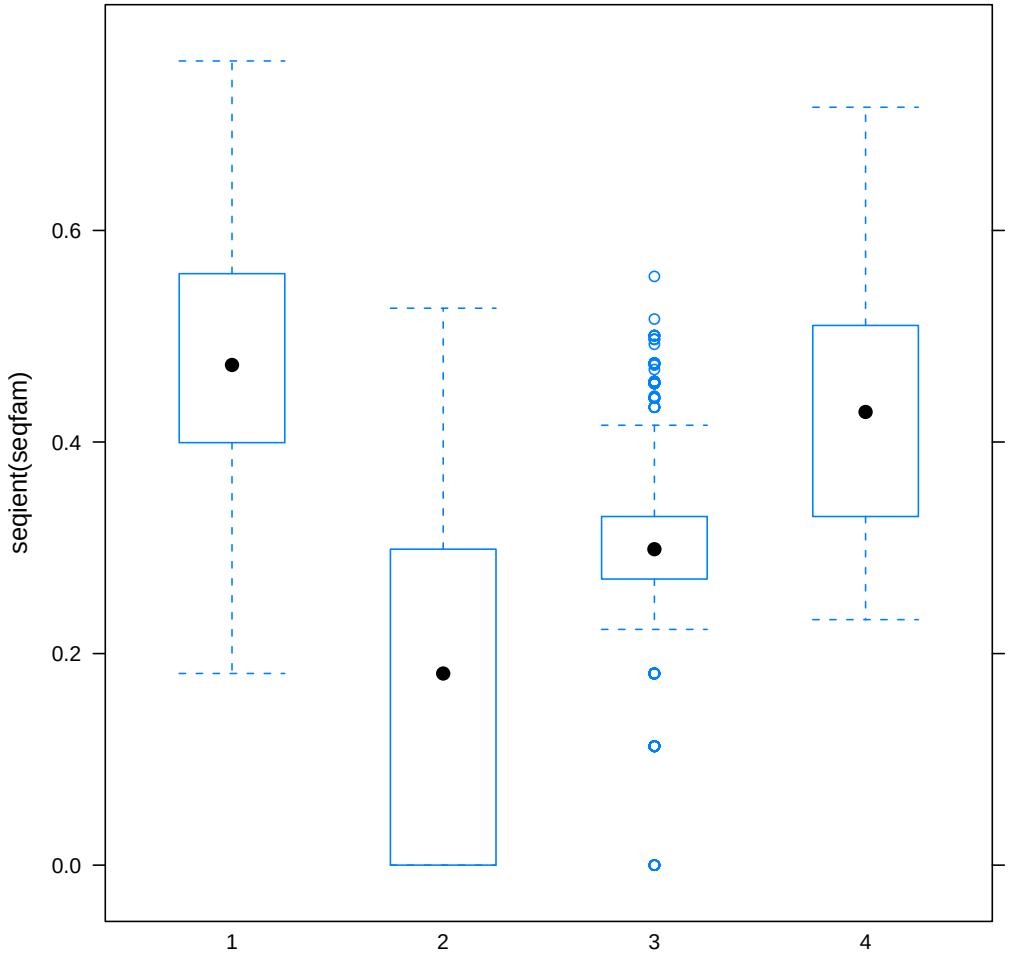
Definition

- Entropy: measure of uncertainty regarding sequence predictability.
  - p_i, proportion of cases (or time points) in state i.
  - Shannon $h(p) = \sum_i -p_i \log_2(p_i)$
  - Other type of entropies: Quadratic (Gini), Daroczy, ...
- Two ways of using entropies.
  - Entropy of the state at each time (age) point: Entropy increases with diversity of states observed at each time point (age).
  - Entropy of each individual sequences: Entropy increases with diversity of states during the observed life course and varies with the time spend in each state.

Entropy of the state at each time (age) point



Entropy - boxplots



TraMineR

The TraMineR R module provides methods to analyze life courses :

- Distance between sequences computation (optimal matching, LCS, LCP)
- Descriptive measures of sequences (entropy, turbulence)
- Sequence visualization tools (density/index/frequency plots)
- Frequent sub-sequence mining

The End

Thank you!