MINING SEQUENCES DATA

STQD6414 PERLOMBONGAN DATA

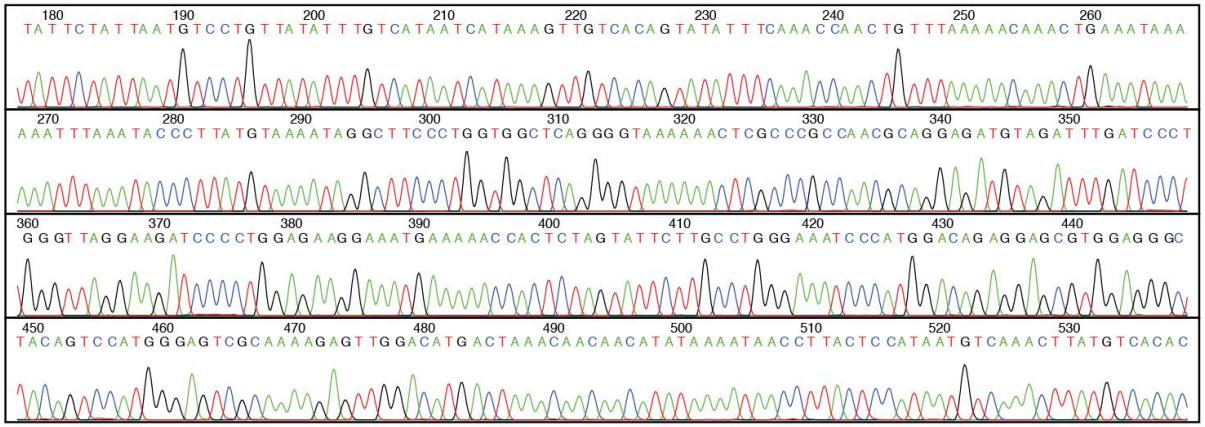
Assoc. Prof. Dr. Nurulkamal Masseran

Department of Mathematical Sciences

Universiti Kebangsaan Malaysia

INTRODUCTION:

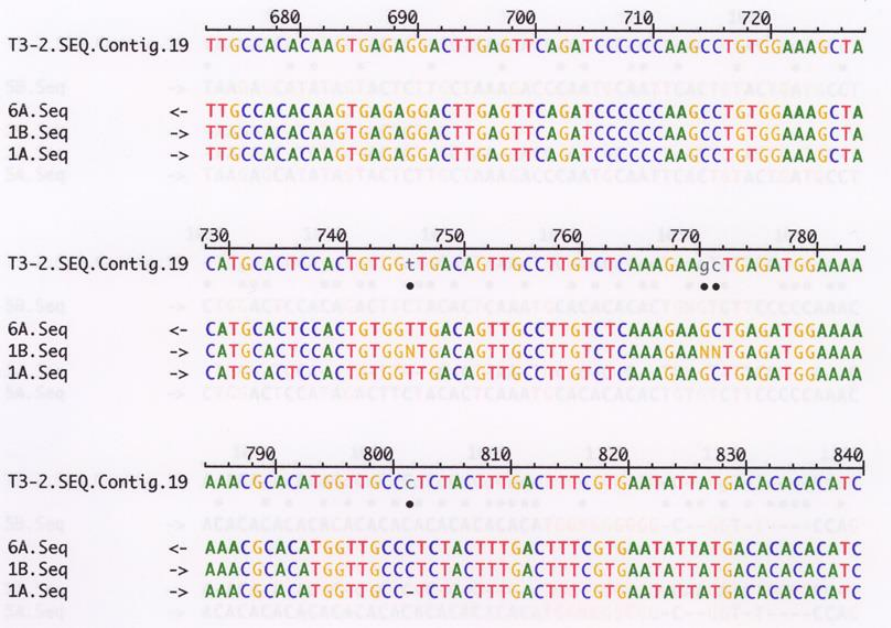
- This topic will discuss about categorical sequence data analysis.
- In the sequence data, the position of each consecutive states gives an interpretation in term of age, date, elapsed time or distance from the beginning of the sequence.
- Generally, this type of data refers to observations of a particular individuals or entities over a some period of time.
- The main objective is to analyze the behavior of the sequence of states for a particular entities.

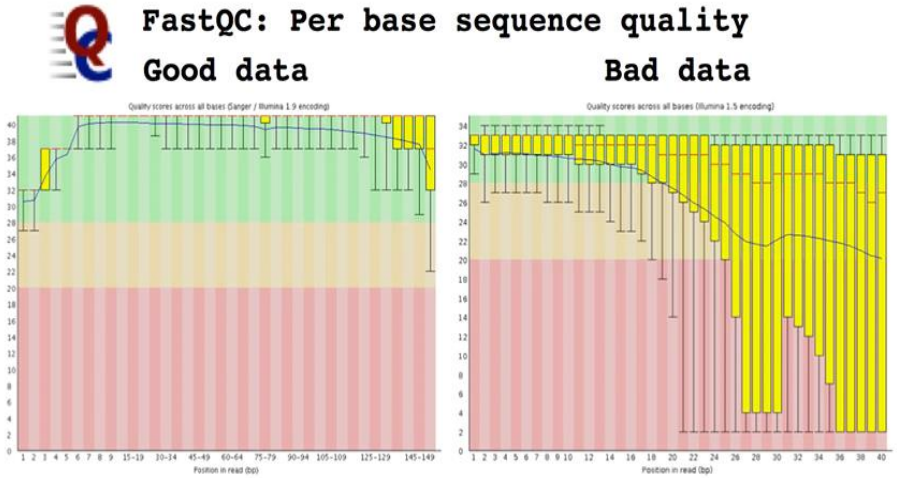


DNA sequence data from an automated sequencing machine

INTRODUCTION:

- For this topic, our discussion will focus on the sequence analysis for life trajectory data.
- However, most of the concepts and techniques for sequential analysis can be applied in various domain areas such as; biology, quality control, text data, log-web data, and etc.





SEQUENCES DATA:

- Sequences are complex objects, and it require specialized data mining techniques to analyze this kind of data.

- Among the interesting questions related to the sequence type data:

i) What are the characteristics of a sequence data?.

ii) What are the indicators that can be used to measure sequence data?.

iii) What are the appropriate plots to visualize a sequence data?.

iv) How can we compare the similarity between several sequences data?.

STATE SEQUENCES:

- Sequence of state is an important concept used to analyze the trajectory of life.
- Example: occupational histories, patient level history, cohabitation life courses and etc.

- Based on state sequence data of cohabitation life courses, we can determine:

i) The characteristics of social norm and standard trajectories of a life courses.

ii) The departures behaviors from the standards trajectories.

iii) The evolution patterns of a life course over time.

iv) The characteristics of cohabitation life correspond to a factor of sex, social origin, cultural, and etc.

STATE SEQUENCES:

- The analysis of state sequence will summarize and categorizing the sequential patterns into some particular groups that having similar properties.

- The sequential analysis techniques:

i) Statistical summary indicators.

ii) Visualization.

iii) Grouping.

iv) Comparing sequences.

- The obtained groups and summary indicators provide an information for further analysis involving various statistical methods.

TYPES OF SEQUENCES DATA:

- Several types of sequence data:

i) States-sequence (STS) format.

ii) State-permanence-sequence (SPS) format.

iii) Time-stamped-event (TSE) format.

iv) SPELLformat.

v) And many more.

Code	Example
STS	**Id** **18** **19** **20** **21** **22** **23** **24** **25** **26** **27** 101 S S S M M MC MC MC MC D 102 S S S MC MC MC MC MC MC MC
SPS (1)	**Id** **State 1** **State 2** **State 3** **State 4** **State 5** 101 (S,3) (M,2) (MC,4) (D,1) 102 (S,3) (MC,7)
SPS (2)	**Id** **State 1** **State 2** **State 3** **State 4** **State 5** 101 S/3 M/2 MC/4 D/1 102 S/3 MC/7
DSS	**Id** **State 1** **State 2** **State 3** **State 4** **State 5** 101 S M MC D 102 S MC
TSE	**id** **time** **event** 101 21 Marriage 101 23 Child 101 27 Divorce 102 21 Marriage 102 21 Child
SPELL	**id** **index** **from** **to** **status** 101 1 18 20 Single 101 2 21 22 Married 101 3 23 26 Married w Children 101 4 27 .. Divorced 102 1 18 20 Single 102 2 21 27 Married w Children

STATISTICAL SUMMARY INDICATORS:

- Among the important statistical summary indicators are:

i) Mean time spent in each state.

ii) Mean time spent in each state by groups.

iii) Number of transitions.

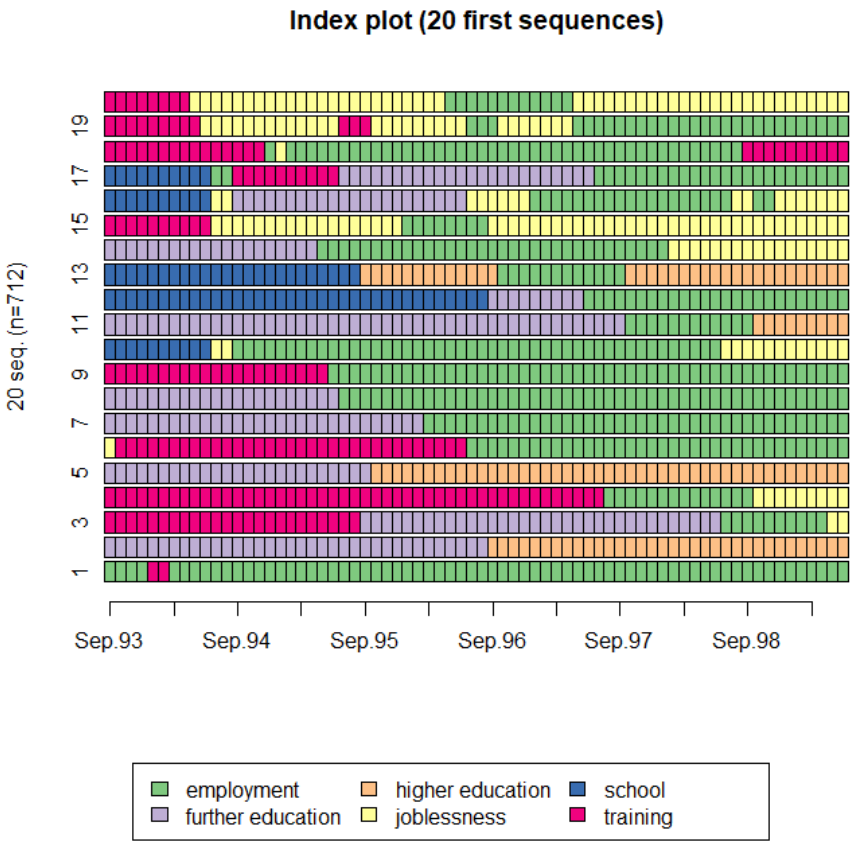
iv) Transition rates.

v) Time varying transition states.

vi) And many more.

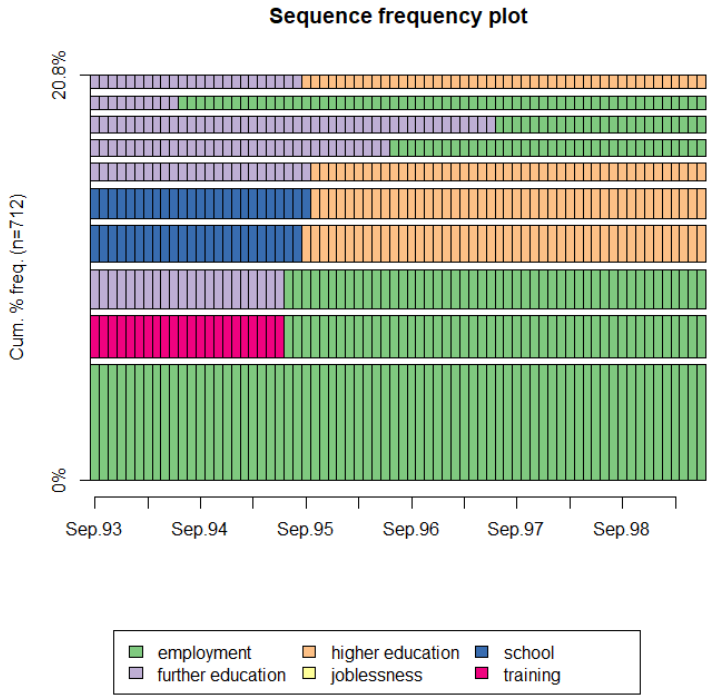
VISUALIZATION: SEQUENCE INDEX PLOT

- A sequence index plot can be used to visualize behaviors of state sequences.
- The plot represented by horizontally stacked boxes which are colored according to the state.
- The horizontal bar width represents a proportional of each frequency.
- Each bar with a different color and length displays information about individual longitudinal changes from one state to another.



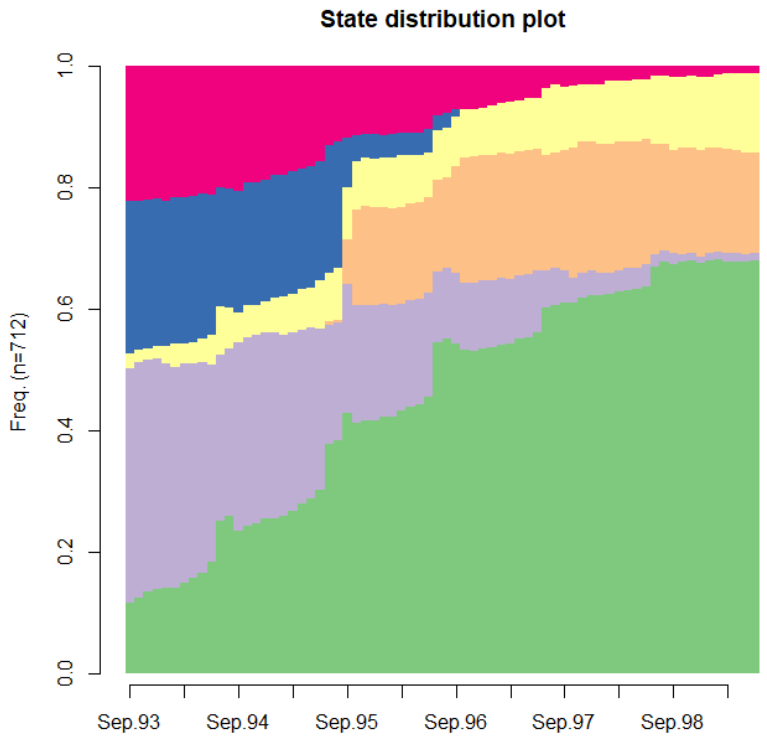
VISUALIZATION: SEQUENCE FREQUENCY PLOT

- Sequence frequency refers to the number and percentage of frequencies arranged in descending order.
- A sequence frequency plot provides a graphical display of the frequency of a sequence with the width of the bar proportional to its frequency.



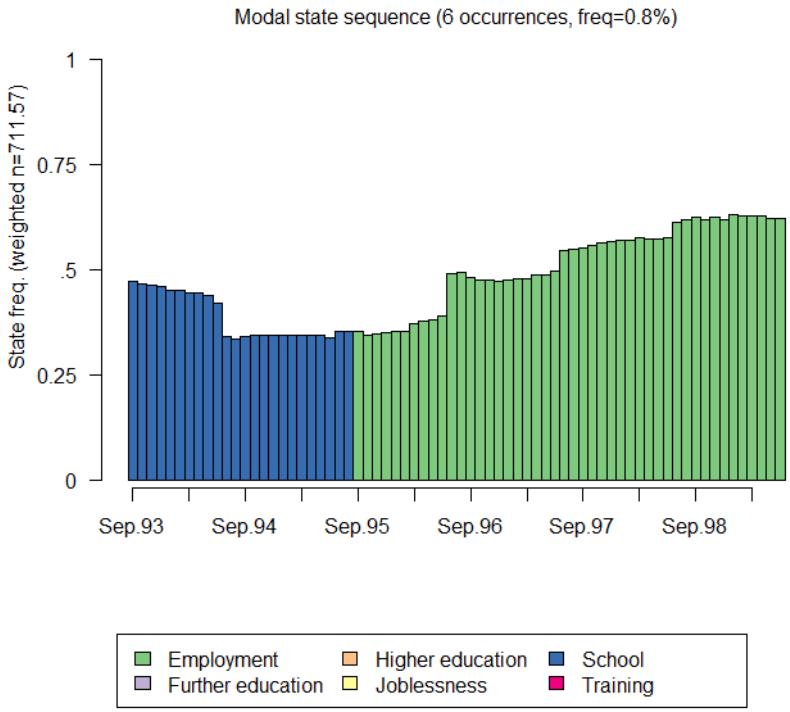
VISUALIZATION: STATE DISTRIBUTION PLOT

- This plot displays the general pattern of the whole set of trajectories in sequence data.
- It provides aggregated views for transversal characteristics of sequences data.



VISUALIZATION: MODAL STATE PLOT

- This plot provides information about the sequence made by the most frequent state at each position.
- It also shows a number of occurrences of the modal state sequence.



SEQUENCE CHARACTERISTICS BY ENTROPY INDEX:

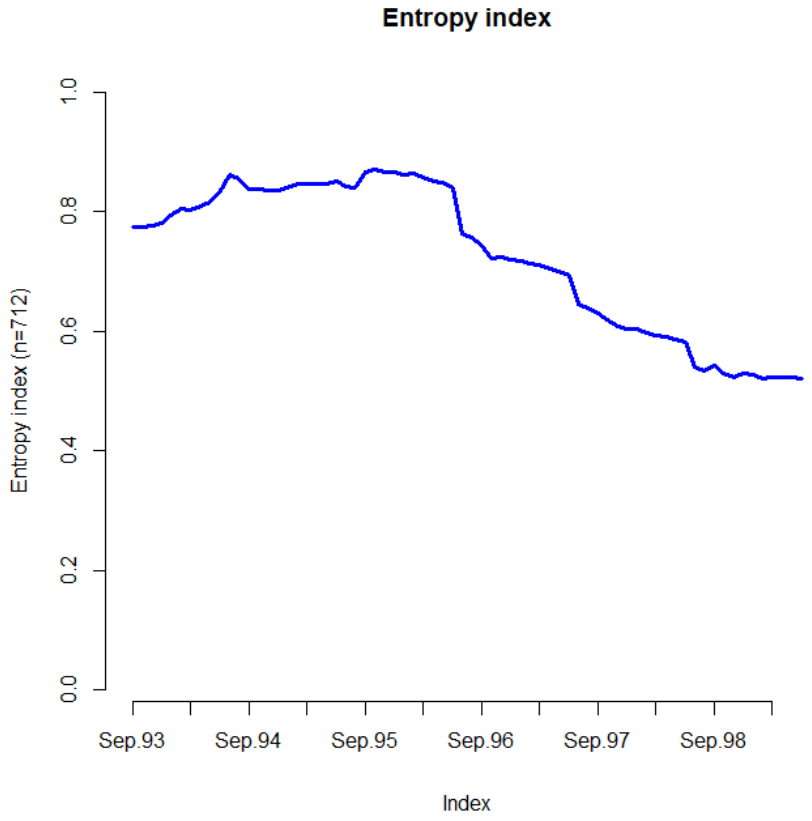
- The entropy provides a measure of the diversity of states.
- Entropy index for sequences data can be determine as follow:

$$h(p_1, \ldots, p_a) = -\sum_{i=1}^{a} p_i \log(p_i)$$

- where p_i is the proportion of cases/entities in state-*i*, *a* is the size of a sequence data.
- If the value of entropy=0, indicates that all cases are in the same state (variation is 0).
- If the value of entropy is high, indicates that the same proportion of cases are found in each state (variation is high).

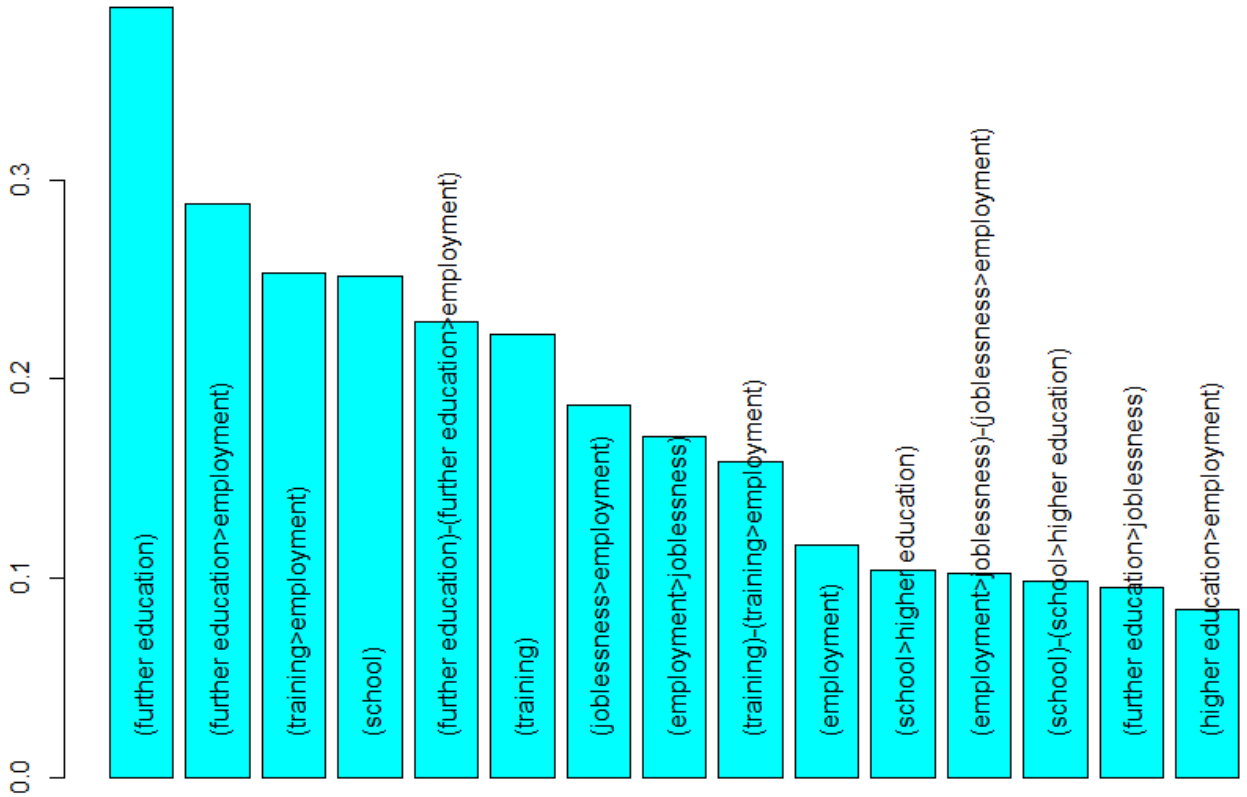
VISUALIZATION: TRANSVERSAL ENTROPIES

- The plot of transversal entropies displays information on the variation of states in the sequence data shown against the time factor.



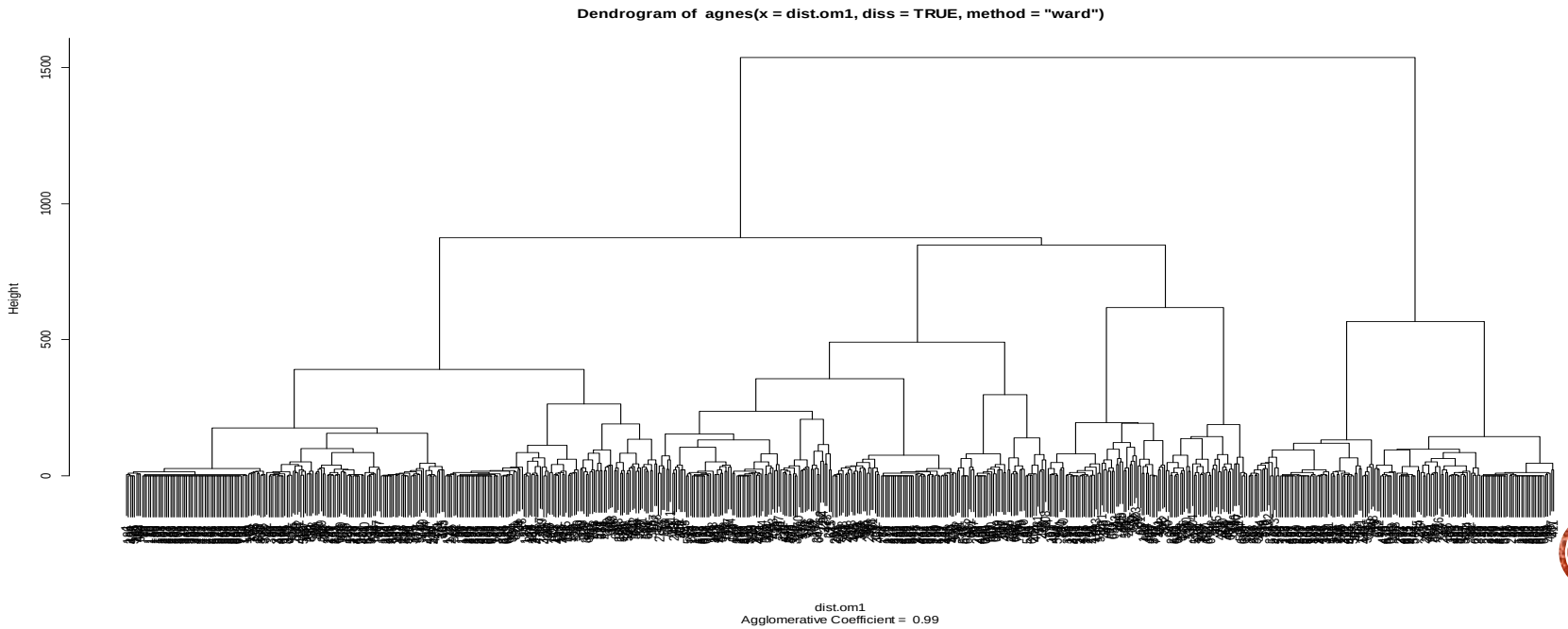
EVENT SEQUENCE:

- Instead of focusing on sequences of states, we can look at sequences of transitions or events.

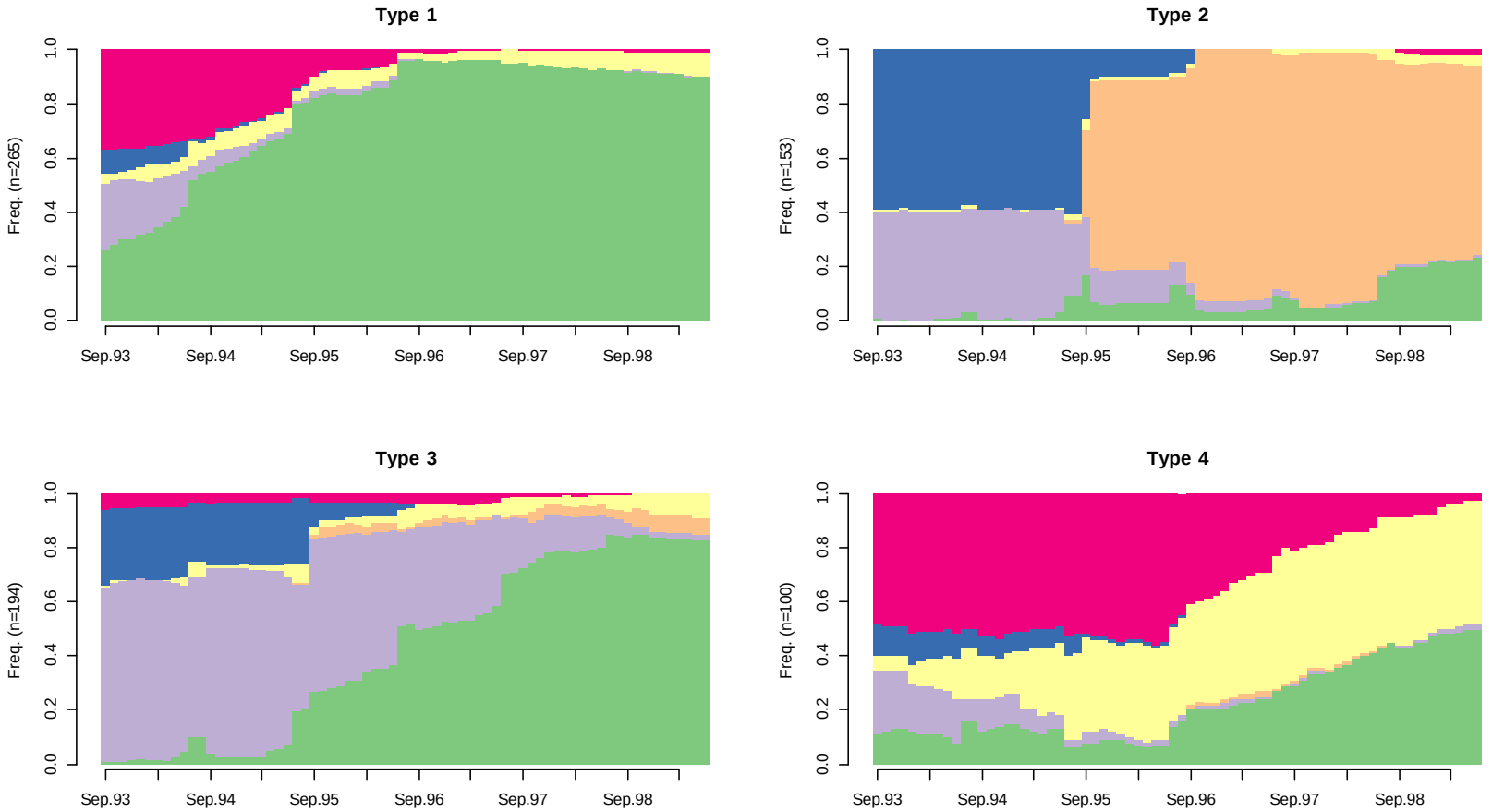


CATEGORIZING PATTERNS:

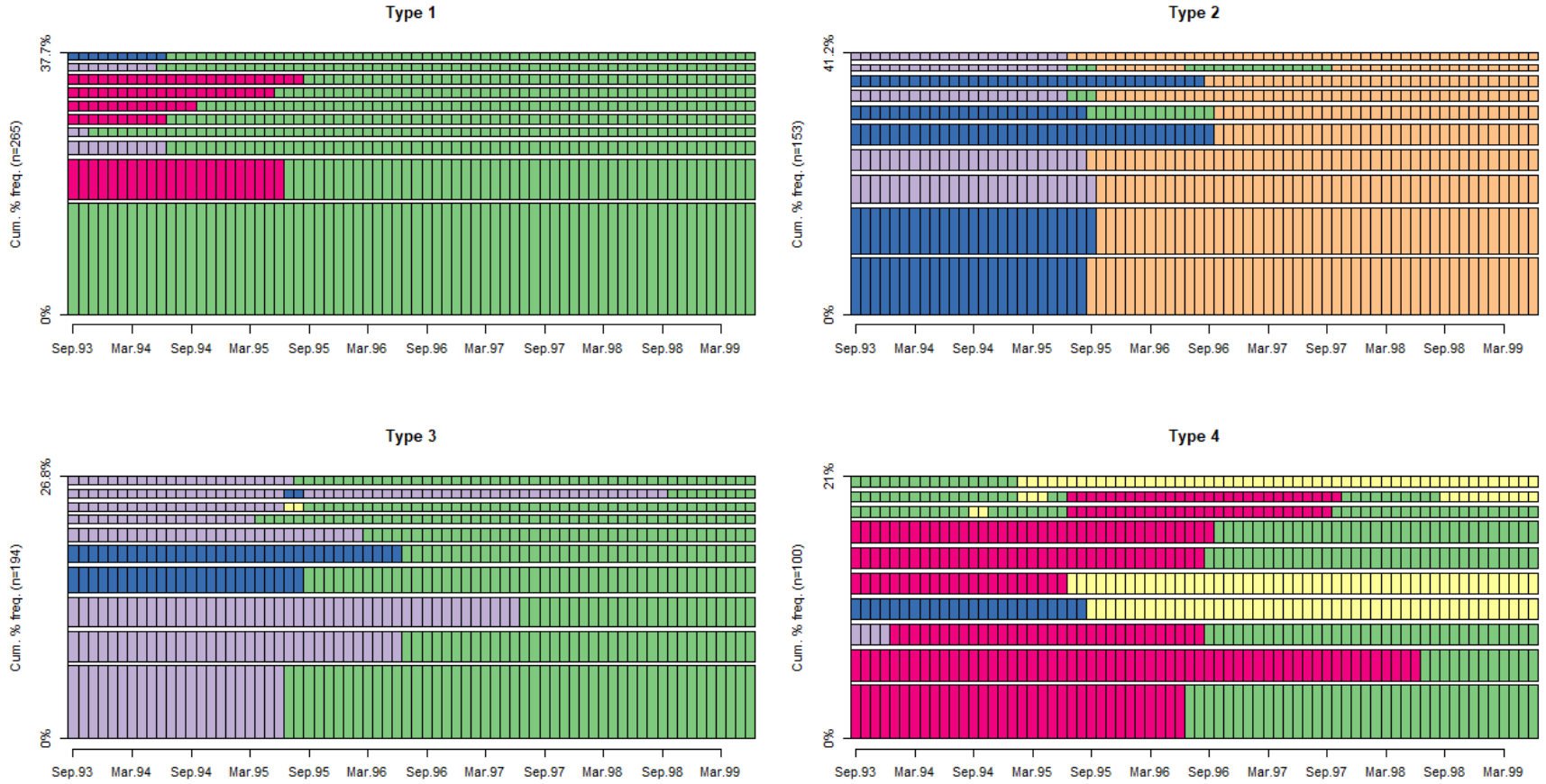
- Categorizing patterns provide information about a typology of a sequences.
- It can be done by measuring similarity between a pairwise distances between a sequences.
- This techniques are based on the algorithm of optimal matching.
- Each cluster of a groups entities indicate similar trajectories characteristics.



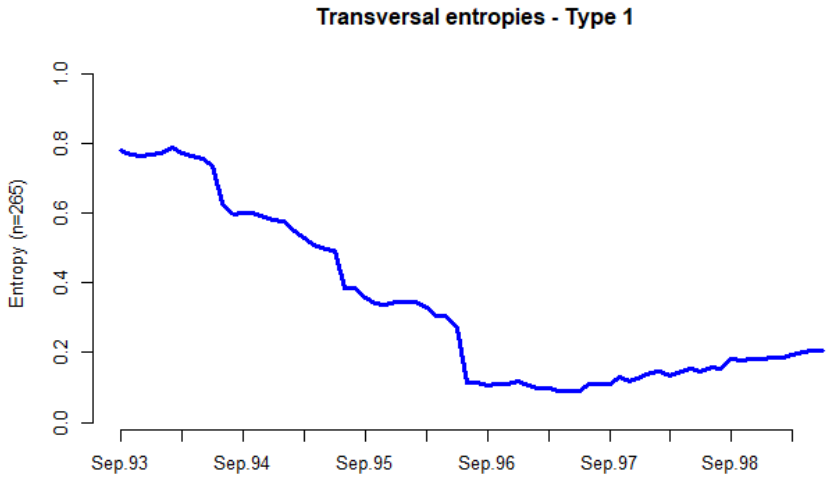
CATEGORIZING PATTERNS: STATE DISTRIBUTION

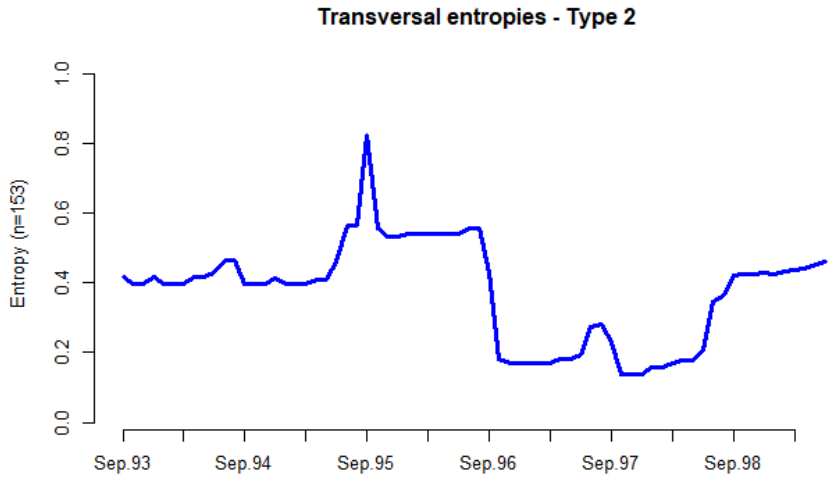


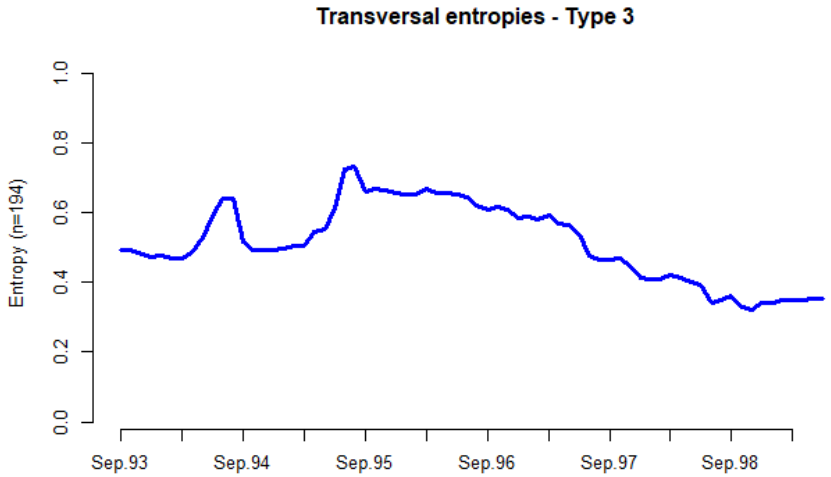
CATEGORIZING PATTERNS: SEQUENCE FREQUENCIES

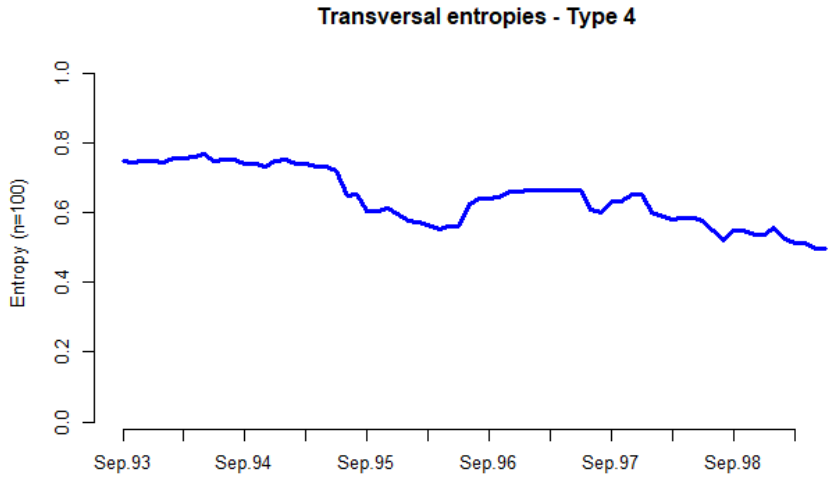


CATEGORIZING PATTERNS: MODAL STATE

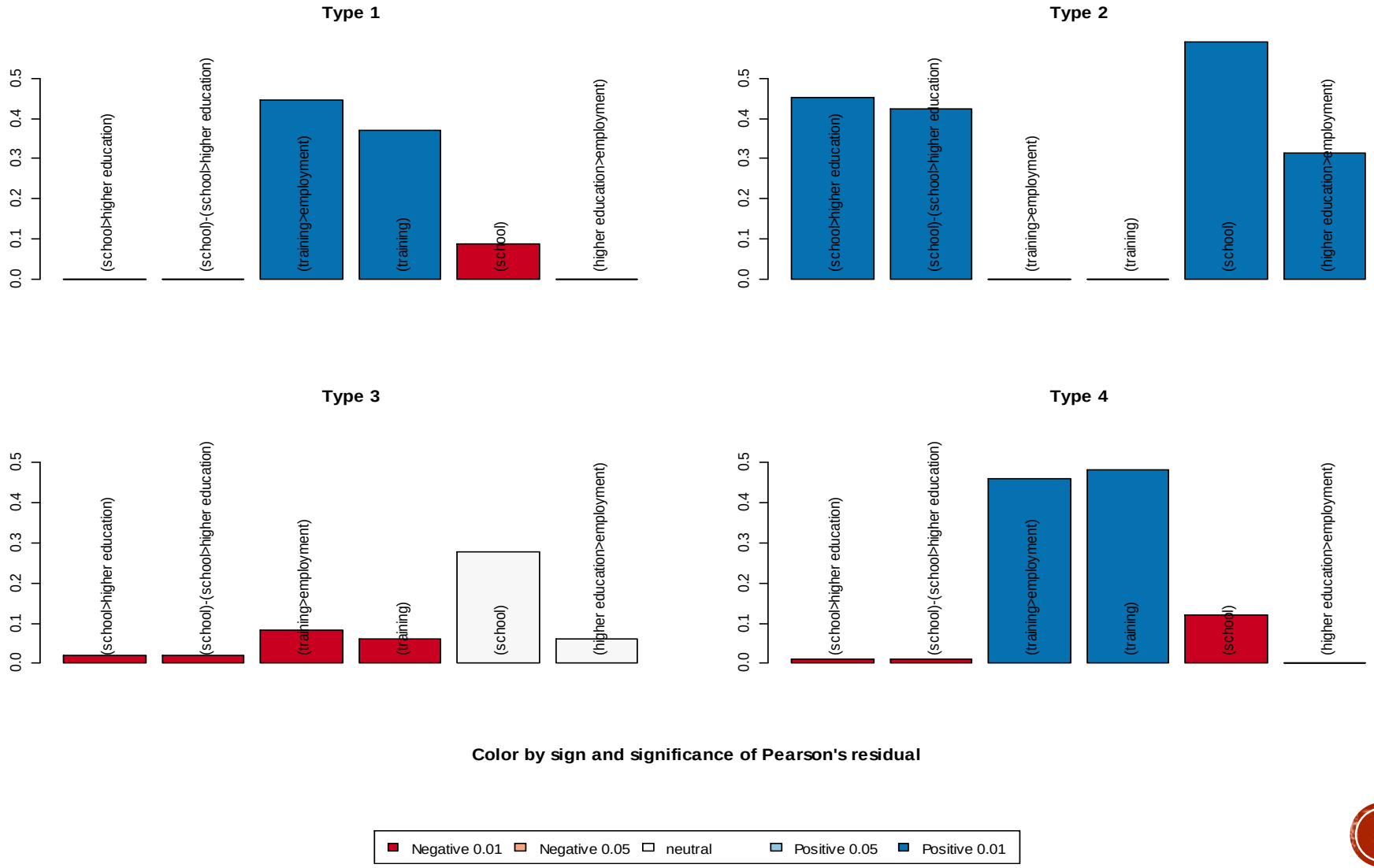








CATEGORIZING PATTERNS:
DISCRIMINATING TRANSITIONS



SEQUENCES ANALYSIS: OTHER APPROACHES

- There are a lot of advance approaches that can be used to deal with state sequences data.

- Some of them are:

i) Correspondence analysis of the states.

ii) Markov modeling.

iii) Event sequences analysis.

iv) Survival analysis.

v) Longitudinal analysis.

vi) Discrete panel data analysis.

vii) And etc.

REFERENCES:

- Curry, E. (2021). *Introduction to Bioinformatics with R: A Practical Guide for Biologists*. Boca Raton, Taylor & Francis.
- Gabadinho, A., Ritschard, G., Müller, N.S., Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Gabadinho, A., Ritschard, G. (2016). Analyzing State Sequences with Probabilistic Suffix Trees: The PST R Package. *Journal of Statistical Software*, 72(3), 1–39.
- Melnykov, V. (2016). ClickClust: An R Package for Model-Based Clustering of Categorical Sequences. *Journal of Statistical Software*, 74(9), 1–34.
- Raab, M., Struffolino, E. (2022). *Sequence Analysis*. SAGE Publications.

NEXT TOPIC:

Mining Text Data