This sequence analysis course (6 ECTS) is given during the spring term in 4 hour weekly sessions. The course covers conceptual and theoretical aspects, but includes also an introduction to the practice of sequence data analysis in R with the TraMineR package.

Focus is on methods for exploring and analyzing categorical longitudinal data describing life courses such as family trajectory or professional careers. The aim is (i) to explain the whole process of sequence analysis from the preparation of longitudinal data and the exploration of sequences to the use of more advanced explanatory methods, and (ii) to train participants to the practice of sequence analysis.

Covered topics include, for state sequences: the visual rendering of sequence data, cross-sectional and longitudinal sequence descriptive statistics, optimal matching and other ways of measuring the dissimilarity between sequences, clustering individual sequences, identifying representative trajectories, discrepancy analysis and regression trees for sequence data; and for event sequences: rendering the sequencings, mining typical subsequences and associations between those subsequences, finding the subsequences that best discriminate between groups such as between women and men for instance, measuring the dissimilarity between event sequences and dissimilarity based analysis of event sequences.

The course is user oriented and includes an introduction to R where participants will acquire the basic knowledge required for using TraMineR. The scope of sequence analysis will be illustrated with real data from the Swiss Household Panel http://www.swisspanel.ch and other datasets that ship with the TraMineR package. Participants are encouraged to train the methods with their own data.

Course organization

Course/Seminar Gilbert Ritschard Wednesday 10h15 - 14h M-5383
Anne-Laure Bertrand (Ass)

First class: Wednesday February 19, 2014

Evaluation

- Each participant will have to realize a case study. (Possibly by groups of two.)
- Oral exams about the realized case study.

Reception hours

<table>
<thead>
<tr>
<th></th>
<th>Office</th>
<th>Phone</th>
<th>e-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gilbert Ritschard</td>
<td>on appointment</td>
<td>5232</td>
<td>022 37 98233</td>
</tr>
<tr>
<td>Anne-Laure Bertrand</td>
<td>on appointment</td>
<td>5203</td>
<td>022 37 99592</td>
</tr>
</tbody>
</table>

Web page: http://mephisto.unige.ch
Course outline (4311012)  
Sequential Data Analysis

1 Introduction
1.1 About longitudinal data analysis
1.2 What is sequence analysis (SA)?
   – How does SA compare with other longitudinal methods?
   – Chronological and non chronological sequences; states, events, transitions.
1.3 What kind of questions may SA answer to? Sequencing, timing and duration.
1.4 Preview of what you will learn.
1.5 TraMineR: an R package for sequence analysis
   – About TraMineR and other softwares for sequence analysis
   – A first run: creating a state sequence object and rendering the sequences

2 Starting with R and TraMineR
2.1 About the R statistical and graphical environment.
2.2 A short introduction to R.
2.3 TraMineR and other useful packages: installing a library and exploring its content and documentation
2.4 Importing data from other softwares and checking the content of data sets
2.5 Basic statistical analysis in R (tabulating data, linear and logistic regression, ANOVA, ...)

3 Rendering and describing state sequences
3.1 The seqdef() function and its options
3.2 Cross-sectional and individual longitudinal characteristics
3.3 Rendering sequences: three basic plots
3.4 Comparing groups and controlling the plots
3.5 Aggregated views of a set of sequences
   – Sequence of cross-sectional indicators (modal state, entropy, ...)
   – Mean time spent in each state, transition rates.
3.6 Longitudinal characteristics
   – Basic attributes: sequence length, number of transitions, state duration
   – Composite characteristics: within entropy, complexity, turbulence.
   – Studying the relationship between sequence characteristics and covariates

4 Handling sequence data
4.1 Formal representations of sequences
4.2 Converting between SPS and STS
4.3 Retrieving spell and person-period data
4.4 Building state sequences from panel data
5 Issues with sequential data
5.1 Missing data, time alignment, unequal sequence lengths
5.2 Time alignment and time granularity
5.3 State codings
5.4 Weights
5.5 What are the main limitations of sequence analysis?

6 Measuring pairwise dissimilarities
6.1 Dissimilarity measures
   – Why measuring dissimilarity?
   – A typology of dissimilarity measures
6.2 Which dissimilarity measure?
   – General guidelines: What matters in sequence differences?
   – Choosing costs in optimal matching
   – Other measures
6.3 OM in TraMineR
6.4 Distance in presence of missing states
6.5 Normalized distances
6.6 Distance to a reference sequence
6.7 Multichannel dissimilarities

7 Dissimilarity-based analysis of state sequences
7.1 Cluster analysis of sequences
   – What is clustering and which method should we use?
   – Plotting sequences by clusters
   – Cluster interpretation and relationship with covariates
   – Validation: How do we determine the number of groups?
   – Scope and limits of cluster-based analysis
7.2 Representation on principal coordinates (multidimensional scaling MSD).
7.3 Discrepancy, Neighborhood and Coverage
   – Measuring the discrepancy of a set of sequences
   – Neighborhood and coverage of a sequence
7.4 Extracting representative sequences
   – Aim and principle
   – Finding single representative
   – Looking for sets of representatives
   – Quality measures of the representatives

8 Further dissimilarity-based analysis: Discrepancy analysis
8.1 Discrepancy analysis: Motivation
8.2 ANOVA-like discrepancy analysis of sequences
   – Univariate analysis
   – Multifactor analysis
8.3 Regression trees for sequence data
8.4 More in depth analysis of discrepancy
  – Checking homogeneity of discrepancy across groups
  – Differences over time

9 Mining event sequences

9.1 Event sequences
  – Sequences of time stamped events: definition and representation.
  – Converting to and from state sequences
  – How does the analysis of event sequences compare with that of state sequences?
9.2 Rendering the sequencing
9.3 Seeking for frequent subsequences
  – Counting algorithm: Sequence mining versus itemset mining
  – Counting methods
  – Time constraints
9.4 Determining the most discriminating subsequences
9.5 Sequence association rules
9.6 Measuring pairwise dissimilarities among event sequences
9.7 Dissimilarity-based analysis of event sequences
Recommended readings


Guidelines for the Homework

The Home Work can be done alone or by group of two.


The report length should be between 12 and 20 pages, plus possibly appendices with extended outputs. Avoid detailed outputs within the main text. Results should as much as possible be synthesized in easily readable tables and graphics. Your capacity to synthesize results is part of the evaluation.

A tentative structure could look out as :

1. Introduction (1-2 pages). Description of the considered issue (which trajectory would you like to analyze?), of the hypotheses that you intend to test (bibliographical references would be welcome), and short enumeration of the data and methods that you will use. It is also good practice to announce in the introduction the main findings of your study.

2. Data (1 page). You should here give detailed information about your data : sources, concerned population, number of cases, precise definition of the states/events as well as of the covariates and their values and, when applicable, applied recoding and filters. If applicable, you should also explain the handling of missing values and whether sequences should be weighted. A third person should be able to reproduce your analysis and results.

3. Exploratory analysis of state sequences (2-3 pages). Descriptive cross-sectional and longitudinal statistics and plots, possibly by selected covariates of interest. Dissimilarity-based analysis (finding clusters, representative sequences, MDS, ...).


5. Exploring sequencing with methods for event sequences (2-3 pages). Identifying frequent sequences and discriminant subsequences for groups of interest.

6. Interpretation and discussion (1-2 pages) of the results in regard of the objective of your study.

7. Conclusion (1 page). Summary of the approach followed and of the main findings.

8. Bibliography. Alphabetical list of consulted references and used software and packages.

This structure is indicative, and it may be judicious in some cases to group some sections such as 4 and 6 for example into a same section.

Do not forget to give all necessary labels in tables and plots to avoid any ambiguity. When necessary, you may shortly explain how the plot or table should be read.