UNIVERSITÉ
DE GENÈVE
GENEVA SCHOOL OF
SOCIAL SCIENCES

Institute for Demographic and Life Course Studies

Sequential Data Analysis Sequential, Spring 2014, Info
4311012 Course by Gilbert Ritschard
Master level

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 Anne-Laure Bertrand (Ass)	Wednesday	10h15 - 14h	M-5383

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

		Office	Phone	e-mail
Gilbert Ritschard	on appointment	5232	022 37 98233	*gilbert.ritschard@unige.ch*
Anne-Laure Bertrand	on appointment	5203	022 37 99592	*anne-laure.bertrand@unige.ch*

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

Abbott, A. and A. Tsay (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research 29*(1), 3–33. (With discussion, pp 34-76).

Aisenbrey, S. and A. E. Fasang (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods and Research 38*(3), 430–462.

Billari, F. C. (2001). Sequence analysis in demographic research. *Canadian Studies in Population 28*(2), 439–458. Special Issue on Longitudinal Methodology.

Billari, F. C. (2005). Life course analysis: Two (complementary) cultures? Some reflections with examples from the analysis of transition to adulthood. In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, and E. Widmer (Eds.), *Towards an Interdisciplinary Perspective on the Life Course*, Advances in Life Course Research, Vol. 10, pp. 267–288. Amsterdam: Elsevier.

Blanchard, P., F. Buhlmann, and J.-A. Gauthier (2012). Sequence analysis in 2012. In *Lausanne Conference on Sequence Analysis (LaCOSA), June 6-8*.

Bürgin, R. and G. Ritschard (2014). A decorated parallel coordinate plot for categorical longitudinal data. *The American Statistician*. forthcoming.

Elzinga, C. H. (2010). Complexity of categorical time series. *Sociological Methods & Research 38*(3), 463–481.

Elzinga, C. H. and A. C. Liefbroer (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population 23*, 225–250.

Gabadinho, A. and G. Ritschard (2013). Searching for typical life trajectories applied to childbirth histories. In R. Levy and E. Widmer (Eds.), *Gendered life courses - Between individualization and standardization. A European approach applied to Switzerland*, pp. 287–312. Vienna: LIT.

Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1–37.

Gabadinho, A., G. Ritschard, M. Studer, et N. S. Müller (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des nouvelles technologies de l'information RNTI E-19*, 61–66.

Maindonald, J. H. (2008). Using R for data analysis and graphics : Introduction, code and commentary. Manual, Centre for Mathematics and Its Applications, Austrialian National University.

Piccarreta, R. and F. C. Billari (2007). Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*(4), 1061–1078.

Piccarreta, R. and O. Lior (2010). Exploring sequences: a graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 173*(1), 165–184.

Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society A 170*(1), 167–183.

Ritschard, G., R. Bürgin, and M. Studer (2013). Exploratory mining of life event histories. In J. J. McArdle and G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Quantitative Methodology, pp. 221–253. New York: Routledge.

Ritschard, G., A. Gabadinho, N. S. Müller, and M. Studer (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

Ritschard, G., A. Gabadinho, M. Studer, and N. S. Müller (2009). Converting between various sequence representations. In Z. Ras and A. Dardzinska (Eds.), *Advances in Data Management*, Volume 223 of *Studies in Computational Intelligence*, pp. 155–175. Berlin: Springer-Verlag.

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. LIVES Working Papers 24, NCCR LIVES, Switzerland.

Studer, M., N. S. Müller, G. Ritschard, et A. Gabadinho (2010). Classer, discriminer et visualiser des séquences d'événements. *Revue des nouvelles technologies de l'information RNTI E-19*, 37–48.

Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research 40*(3), 471–510.

Widmer, E. and G. Ritschard (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life Course Research 14*(1-2), 28–39.

Guidelines for the Homework

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

Deadline : The report must be returned before Sunday, May 11, 2014.

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, ...).
4. Causal analysis (4-5 pages). Studying relationship of complexity and/or cluster membership with covariates by means of regressions. Discrepancy analysis and regression trees.
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.