14D007
Data Visualization

Winter Term - 3 ECTS
Elective Course
Prof. Michael Greenacre
and Prof. Ioannis Arapakis

Prerequisites to Enrol

Students should be familiar with R programming and have some knowledge of basic statistical concepts.

Overview and Objectives

The course deals with the visualization of large data sets as a means of communicating relevant data patterns in the form of graphical displays, in order to interpret and understand the data.

The first five weeks deal with "exact" data visualization and visualization of simple data summaries. We begin by covering topics on visual literacy, exploring the graphical elements used to create effective data visualizations, and rooting the choice of graphical display in the purpose for which it is intended. Next, we explore methods for visualizing univariate, bivariate and multivariate data types and conclude with advanced techniques for visualizing temporal and spatial data.

The last five weeks deal with "approximate" data visualization using multivariate techniques such as distance scaling and dimension reduction, and also different approaches to cluster analysis. In these multivariate approaches, some of the data variance is sacrificed, assuming it is "noise", in favour of making a simpler result that captures the main features in the data, assuming them to be "signal". The methodology is explained in a series of actual case studies, where each one tells its own data visualization story.

Course Outline

Week 1

• Introduction to visual literacy • Identifying key factors and learning about data • Basic principles of data presentation and colour theory • Taxonomy of data visualization methods • Tufte's principles of scientific graphics

Week 2

• Data preparation • R base graphics • Introduction to ggplot2 • Basic plot types • Grammar of Graphics (ggplot2)

Week 3

• Grammar of Graphics (ggplot2) • Mappings • Layers • Scales • Facets • Themes • Visualization of multivariate data • Revealing uncertainty • Dealing with overplotting

14D007
Data Visualization

Winter Term - 3 ECTS
Elective Course
Prof. Michael Greenacre and Prof. Ioannis Arapakis

Week 4

• Techniques and tools for visual representation of temporal data

Week 5

• Techniques and tools for visual representation of spatial data

Week 6

• Introduction to multidimensional data visualization • Case study 1: Climate indices in the Arctic (illustrates data standardization and principal component analysis)

Week 7

• Case study 2: Antibiotic use in Europe (illustrates the logarithmic transformation, principal component analysis and temporal trend) • Case study 3: Attitudes to women working and the effect on the family, across the world (illustrates simple and multiple correspondence analysis, and trends in attitude over time)

Week 8

• Case study 4: Visualizing "proximities" between districts based on demographic indicators (illustrates concatenating of data tables, weighted multivariate distance and focusing on group differences) • Case study 5: Fish species abundances in the Barents Sea (illustrates fuzzy coding, correspondence analysis, and constrained canonical correspondence analysis)

Week 9

• Case study 6: Identifying genes that classify four different types of child cancers (illustrates individual-level and group-level differences in dimension reduction and a novel tuning parameter for group prediction) • Case study 7:The microbiome challenge (illustrates compositional data analysis on a wide matrix with hundreds or thousands of variables)

Week 10

• Case study 8: Classification of lower back pain sufferers (illustrates data on mixed measurement scales, k-means clustering, deciding on number of clusters) • Case study 9: My "reconstruction" of the book "Mathematics and Archaeology" (illustrates analysis of text, preparing frequency tables, correspondence analysis, clustering and heat maps)

Required Activities

Attendance at classes, and submission of homework.

14D007
Data Visualization

Winter Term - 3 ECTS
Elective Course
Prof. Michael Greenacre
and Prof. Ioannis Arapakis

Evaluation

• Homework during the course (40%) • Two short practical projects, done individually, one at weeks 4-5 and another at weeks 9-10 (50%). • Class attendance/participation (10%)

Competences

☒ Solve the real problems that arise in the fields of study through the accurate analysis of the data.

☒ Visualize and interact with high-dimensional data in order to contextualize the information and facilitate subsequent decision-making.

☒ That students know how to apply the acquired knowledge and their ability to solve problems in new or unfamiliar environments within broader (or multidisciplinary) contexts related to their area of study.

☒ That the students know to communicate their conclusions and the knowledge and last reasons that sustain them to specialized and non-specialized publics in a clear and unambiguous way.

☒ That students have the learning skills that allow them to continue studying in a way that will be largely self-directed or autonomous.

Learning Outcomes

☒ Apply supervised and semi-supervised learning algorithms.

☒ Apply mathematical theory and statistics on data sets from disparate disciplines.

Pre-course reading

Read the short extract “Aesthetics and Technique in Data Graphical Design” from Edward Tufte’s highly recommended book *The Visual Display of Quantitative Information* (Cheshire Press) at this link:

`www.econ.upf.edu/~michael/DataViz/tufte-aesthetics_and_technique.pdf`

Materials

Online links to relevant material on the web (visualization examples, readings, videos) will be given as well as the class material and homework.

14D007
Data Visualization

Winter Term - 3 ECTS
Elective Course
Prof. Michael Greenacre
and Prof. Ioannis Arapakis

For an introduction to the visualization of multivariate data, you can consult the following books, all available for free download:

Greenacre, M. (2010) *Biplots in Practice*. BBVA Foundation, Madrid. Download from

www.multivariatestatistics.org

James, G., Witten, D, Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning, with Applications in R*. Download from

http://www-bcf.usc.edu/~gareth/ISL/

Supporting reading

Blasius, J. and Greenacre, M. (2014). *Visualization and Verbalization of Data.* Chapman & Hall / CRC.

Cook, D. and Swayne, D.F. (2007). *Interactive and Dynamic Graphics for Data Analysis.* Springer UseR! Series

Greenacre, M. (2016). *Correspondence Analysis in Practice*, *3rd edition.* Chapman & Hall / CRC.

Maindonald J. and Braun J. (2003). *Data Analysis and Graphics Using R. Third Edition.* Cambridge University Press.

Wickham H. and Grolemund G. (2016). *R for Data Science - Import, Tidy, Transform, Visualize, and Model Data. First Edition.* O'Reilly Media,

Wickham H. (2016). *ggplot2. Elegant Graphics for Data Analysis. Second Edition.* Springer UseR! Series.

Zettl, H. (2011). *Sight Sound Motion. Applied Media Aesthetics. Sixth Edition*. Wadsworth.