Advanced Statistical Methods

Visualizing and analysing categorical and textual data

Professor: Michael Greenacre
Office hours: By appointment
Course Type: Elective
Credits: 5
Term: 3rd

Course Description

In the present data deluge, tools are required to make sense out of the vast quantities of available information for purposes of decision-making in management and marketing. The statistical methods that are most often taught, such as regression modelling, analysis of variance and factor analysis, tend to concentrate on data that are already quantified on continuous scales, whereas in reality most of the data being collected are of a categorical nature (e.g. response categories in a survey, products purchased, preferences, lifestyle choices,...) or in the form of text (open-ended responses, social media posts, company reports, product descriptions,...). This course fills the gap by showing how to visualize, quantify and analyse such data.

Objectives

Students in this course will

- learn the importance of visual display in the understanding and interpretation of empirical data
- expand their knowledge and skills to the analysis of categorical data and data in the form of text
- be introduced to novel methods of categorical data analysis such as correspondence analysis, latest class modelling, market basket analysis, the analytic hierarchy process as well as the analysis of mixed-scale data that are typically found in practice, which involve a mix of both continuous, categorical and textual scales
- learn about various approaches to text mining and text analytics, from pre-processing and feature extraction to segmentation, pattern recognition and sentiment analysis

Methodology

The teaching approach is application-oriented and case-based. Many different examples of categorical and textual data will be studied. Students will have plenty of practical experience during the course and at least two classes will be in the form of a computing workshop. The software used in the course is the powerful programming environment of R, but students are welcome to use any other software that they are familiar with, e.g. Python or the package Stata.

The competences, the learning outcomes, the assessment elements and the quality of the learning process included in this Teaching Plan will not be affected if during the academic trimester the teaching model has to switch either to a hybrid model (combination of face-to-face and on-line sessions) or to a complete on-line model.

Evaluation criteria

- **Continuous assessment: 40%**
- **Participation (includes class presence): 10%**
- **Course project: 50%**

Students are required to attend **80% of the classes**. Failing to do so without justified reason will imply a zero grade in the participation/attendance evaluation item and may lead to suspension from the program.

Obtaining either less than 6/10 for the course project or less than 6/10 for the course as a whole will imply failure of the course. Students who fail the course during the regular evaluation are allowed ONE re-take of the evaluation. According to the previous history of this course, it is highly unlikely that the course project is the main reason for failure, since the projects are done under close supervision of the course leader. Hence, failure would more likely be due to a poor participation or continuous assessment record, in which case the retake would consist of either an additional set of home assignments or the chance to revise the course project to increase its component of the grade. If the course is again failed after the retake, the student will have to register again for the course the following year.

Plagiarism is to use another's work and to present it as one's own without acknowledging the sources in the correct way. All essays, reports or projects handed in by a student must be original work completed by the student. By enrolling at any UPF BSM Master of Science and signing the "Honour Code," students acknowledge that they understand the schools' policy on plagiarism and certify that all course assignments will be their own work, except where indicated by correct referencing. Failing to do so may result in automatic expulsion from the program. The use of artificial intelligence platforms should be acknowledged in any homework or in the final project, otherwise its detection in any submitted document will be regarded as plagiarism.

Calendar and Contents

Week 1	Introduction
	Examples of categorical and textual data
	Methods of data collection
	Graphical elements, styles and examples of visualization
Week 2	Frequency tables
	Chi-square tests, measures and interpretation of variable association
	Heat maps and tree maps
Week 3	Market basket analysis
	Loglinear modelling to study categorical variable interactions
	Market share and brand switching
Week 4	Preferences, pairwise comparisons
	Analytic hierarchical process for resource allocation and decision-making
Week 5	Survey data analysis
	Dimension reduction: introduction to correspondence analysis
	Multiple correspondence analysis
Week 6	Clustering of categorical data
	Latent class modelling
Week 7	Text analytics: basic concepts and pre-processing
Week 8	Text analytics: classification and visualization
Week 9	Text analytics: sentiment analysis
Week 10	Mixed-scale data analysis: the final challenge

The deadline for the project submission will be negotiated by the class according to their examination timetable, taking into account the grades deadline at the end of the trimester.

Reading Materials/ Bibliography/Resources

Anandarajan M, Hill C & Nolan T (2018) *Practical Text Analytics*. Springer.

Greenacre M (2016) *Correspondence Analysis in Practice*, 3rd edition. Chapman & Hall / CRC. The Spanish translation of the 2nd edition is available for free download at **`www.multivariatestatistics.org`**.

Greenacre M & Primicerio R (2010) *Multivariate Analysis of Ecological Data*. Free download at **`www.multivariatestatistics.org`**.

Szabo G, Polatkan G, Boykin O & Chalkiopoulos A (2019) *Social Media Data Mining and Analytics*. Wiley

In addition, several PDF files of relevant literature will be supplied to students, as well as web links.

Specific competences

The student can expect to acquire the following competences in the master program and specifically in this course.

SC3. Solve managerial problems through the use of analytical and research techniques.

SC4. Acquire the skills for the design and implementation of problem-solving models, based on insights from the social sciences.

SC7. To integrate relevant and current scientific research to generate insights in support of business practice.

SC8. Apply the techniques and theories acquired in the Master's Degree to solve problems relevant to the business world.

Bio of Professor

Michael Greenacre is Professor of Statistics at the UPF's Department of Economics and Business since 1994. He has a research record of over 100 peer-reviewed papers and 10 books on applied multivariate statistics in the areas of social science and ecology, specializing in correspondence analysis and compositional data analysis (several of his books are open-access – see the link below). He has participated as a statistician in many ecological projects in the Arctic region of Norway and has taught regularly to marine biologists in several countries. He has previously given a course on methods of marketing research for several years in the BSM and also taught data visualization in the Master of Data Science of the Barcelona School of Economics. He is also a keen musician, both pianist and guitarist, has published two albums of his own original music (hear the theme songs on his personal page indicated below), wrote 'The Millennium Song', which has been recorded in several languages, and has written several satirical songs about statistics, posted on his YouTube channel **StatisticalSongs**.

www.econ.upf.edu/~michael	(updated personal page)
www.multivariatestatistics.org	(open-access books)
www.youtube.com/StatisticalSongs	(YouTube channel of statistical songs)
www.globalsong.net	(Millennium Song project)

('The Millennium Song' was recorded originally in English and the four languages of Spain, and published as a CD that formed the official UPF gift for its 10th anniversary).