Ljubljana Doctoral Summer School
8 – 12 July 2024
9:00 – 13:00 (CET, Ljubljana)

DATA PRODUCTION, PROCESSING AND ANALYSIS IN APPLIED RESEARCH: METHODS AND TOOLS (ECTS: 4)

Jesus GONZALEZ-FELIU
Excelia Business School, France

Aims of the course

The purpose of this intensive course is to present the basics of data production and analysis, i.e. the main concepts and methods for the production, processing and analysis of data (as an input and an output of a model) to understand and describe phenomena and subsequently forecasting or estimation procedures. This course will focus on three main categories of notions: first, the theory and practice of decision making, decision processes and decision modelling; second, the methodological foundations of data retrieval, data production, processing and preparation; finally, on the main mathematical, computational and statistical methods to prepare, analyze and visualize data for economics and business research and decision making. More precisely, this couse will allow the students to define and design the global process of data production and analysis (defining objectives, identifying the sample/population to enquire, collecting or estimating data, processing those data for the given objectives, analysing them and interpreting the results, to finally relating the given outputs to the searched reality), applicable to different research and scientific fields, as well as to choose and use the most suitable methods related to their objectives and needs.

Syllabus

In this course, students will address the following content:

1. The different ways of the scientific path: deduction, induction and abduction.
2. Basics of data production and collection for both quantitative and qualitative data.
3. The different stages of modelling and estimation (from objective definition to model validation, including data retrieval).
4. Main elements of data collection, processing and analysis: most common assumptions, validity of uses and groups of techniques.
5. Small set statistics: which methods and estimates can be used with very small sets of data.
6. Hypothesis testing and choice of the most suitable tests: main tests, hypotheses and elements of test choice related to the aim of the analysis.
7. Classification and clustering methods. Forecasting methods.
8. Interpretation, replicability and research and practical considerations

TENTATIVE SCHEDULE

Day 1: Research, data production and decision making

1. An introduction to scientific path and its three ways of knowledge-seeking: deduction, induction and abduction.
2. Data production and collection: definitions and main methods (for both quantitative and qualitative data, based on observations, declarations, measurements or estimations, among others).
3. The decision process and its modelling issues: decision problem identification, problem structuring, problem solving, problem validation.
4. The main categories of statistical and computational methods for data production and analysis: how to choose, when to use them.

Reading assignments: Lecture notes and basics given before the course (Gonzalez-Feliu, 2023; Gonzalez-Feliu, 2024 chapter 1).
Work to be done before class: Read the notes.

Day 2: The data production process

1. Definition of the main stages of the data production and analysis process.
2. Objectives, variables and choice of the suitable data production method.
3. The basics of sampling: types of samples, choice of the sample size, reliability considerations.
4. Collecting data: practical considerations in deploying observations, measures and surveys.
5. Data completion, modelling and simulation: main considerations for producing data by estimations.
6. Adjusting and processing collected or produced data.

Reading assignments: Course lecture notes (Gonzalez-Feliu, 2024 chapter 2), book chapter on questionnaire design and sampling (Parfitt, 2013).
Work to be done before class: Read the assignments.

Day 3: Processing and analyzing data

1. Characteristics of dispersion, asymmetry and flattening: how to choose the suitable indicators and interpret them. Types of means and dispersion indicators, statistical distribution analysis, descriptive statistics plots. Exercices with Excel or R.
2. Considerations on using normality, statistical distributions and probability theories in analyzing and interpreting data. Normality hypotheses, pseudo-normality issues. Discrete and continuous statistical distribution approximations.
3. An introduction to small set statistics: methods, conditions of applications and validity.

4. Considerations on processing big sets and big data: the role of sampling, computing and clustering.

Reading assignments: Lecture notes (Gonzalez-Feliu, 2024, chapter 3), two blog issues on small set statistics (Sauro, 2013; Lewis and Sauro, 2022).
Work to be done before class: Read the assignments.

Day 4: Classification and clustering issues

1. Role and interest of classification and clustering.
2. Introduction to principal component analysis and factorial analysis.
3. Non-hierarchic clustering methods: the example of k-means.
4. Hierarchic clustering methods: the example of agglomerative clustering.
5. Using hypotheses tests for classification and clustering: T-Tests and F-Tests.

Main exercises will be done in Excel or R.

Reading assignments: Course lecture notes (Gonzalez-Feliu, 2024, chapter 4).
Work to be done before class: Read the assignments.

Day 5: Forecasting, interpretation, research and practical considerations

1. Introduction to forecasting methods and statistical applications: linear, quadratic and generalized linear regressions.
2. Computational forecasting methods: introduction to learning algorithms (neural networks or ant colonies).
3. From data to theories: how concluding and discussing obtained data in a scientific way.
4. Replicability and transferability of data production and analysis methods.

Reading assignments: Lecture notes (Gonzalez-Feliu, 2024, chapter 5), research paper on the impacts on collection/production choices in data quality and interpretation keys (Gonzalez-Feliu, 2019).
Work to be done before class: Finish the two case studies.

The course will be evaluated on the basis of continuous assessment on the exercises proposed during the course and an individual assignment.

Course materials / List of readings

Gonzalez-Feliu, J. (2024). *Introduction to data production, processing and analysis. Preliminary and basic notions.* Notes specially prepared for this course that will be available online before the course.

Gonzalez-Feliu, J. (2023). *Data-driven decisión making*. Lecture Notes. Excelia Business School. Available online.

Gonzalez-Feliu, J., & Sánchez-Díaz, I. (2019). The influence of aggregation level and category construction on estimation quality for freight trip generation models. *Transportation Research Part E: Logistics and Transportation Review*, *121*, 134-148.

Lewis, J., & Sauro, J. (2022). Five Styles of Statistical Rhetoric. Retrieved from MeasuringU webpage, https://measuringu.com/five-styles-of-statistical-rhetoric/, on Jan 17th 2023.

Parfitt, J. (2013). Questionnaire design and sampling. In *Methods in human geography* (pp. 78-109). Routledge.

Sauro, J. (2013). Best Practices for Using Statistics on Small Sample Sizes. Retrieved from MeasuringU webpage, https://measuringu.com/small-n/, on Jan 17th 2023.

Course credit

Students needing course credit for their PhD studies will have to successfully pass an oral presentation on the last day of the course and submit a small report of a case study.

Course leader

Jesus Gonzalez-Feliu is Full Professor in Supply Chain Management at Excelia Business School since 2020. Previously, he was associate professor at Mines Saint-Etienne and research engineer at CNRS. He has a PhD. in computer and systems sciences from Politecnico di Torino, Italy and a Habilitation to Direct Researches from Université Paris Est, France. Before, he was associate professor at Ecole des Mines de Saint-Etienne and Research Engineer in Data Production at the National Center of Scientific Research, France, among others. His main teaching and research subjects are the design of logistics and transport systems, the estimation of urban freight and shopping demand, the processes of decision and management driven by data and information, the evaluation of supply chains including sustainability, resilience and maturity, and collaborative decision making. He is author or editor of five books, co-author of more than 50 peer-reviewed journal papers, guest editor of ten special issues in international journals and has coordinated various courses on the field in international weeks in different countries, at Bachelor, Master or Doctoral level. He has also worked with leading companies and institutions on applied research projects about data production, analysis and decision support, and contributed to the development of city policies based on data-driven approaches.

Contact: gonzalezfeliuj@excelia-group.com