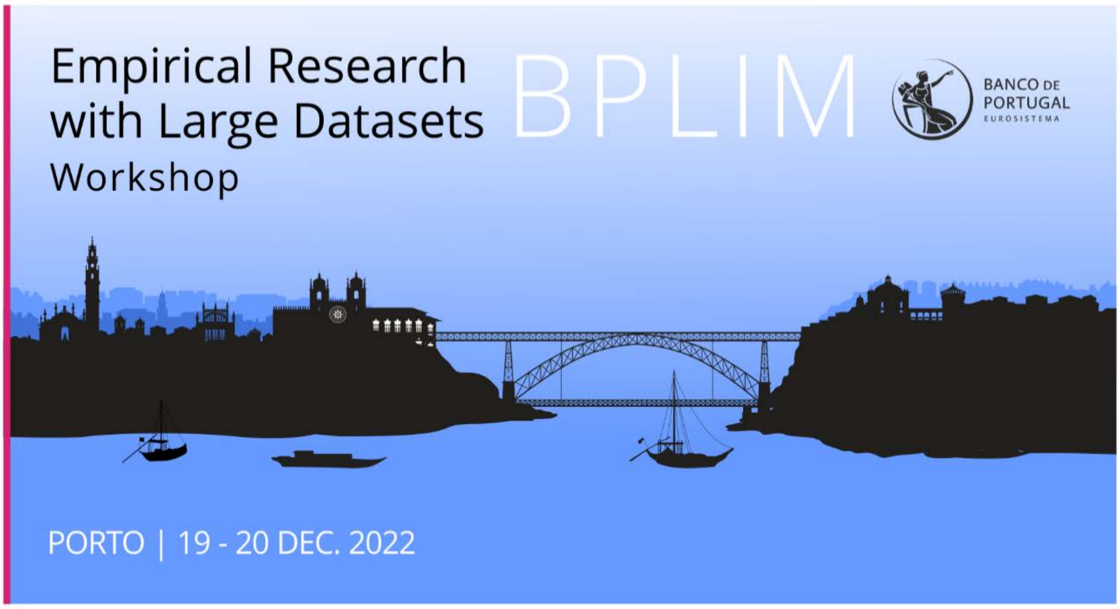
BANCO DE PORTUGAL
EUROSYSTEM



Program

(GMT times)

All materials available here.

Monday, December 19th, 2022

14h00 – 14h15 **Welcome**
Paulo Guimarães, BPLIM - Banco de Portugal

14h15 – 15h00 **“Reproducibility and collaboration when your data is really large or confidential”**
Lars Vilhuber, Cornell University and American Economic Association

The current paradigm in economics requires that results presented in research papers and articles be reproducible and transparent. This can pose challenges when data are large, confidential, and the processing complex and reliant on more than a personal laptop, in other words, when Big Data are involved. In this session, I will discuss considerations, both in terms of documentation and in terms of computational setup, that allow for reproducible and accessible research in such situations.

15h00 – 15h45 **“Big Data Analytics: A guide to economists making the transition to Big Data”**
Ulrich Matter, University of St.Gallen

Successfully navigating the data-driven economy presupposes a certain understanding of the technologies and methods to gain insights from Big Data. There is a to date unmet high demand in government and industry for professionals with an economics or social science background who are able to bridge the gap between data analytics/modelling and data engineering. This talk aims to help economists with such a background to successfully manage the transition to Big Data (a topic traditionally mainly in the focus of data engineering). Building on familiar content from applied econometrics and business analytics, the talk introduces basic concepts of Big Data

Analytics. Thereby, the aim is to give a brief overview over how to productively apply econometric and machine learning techniques with large/complex data sets, as well as on all the steps involved before actually analysing the data (data storage, data import, data preparation). I will point to additional conceptual/theoretical material as well as examples of the practical application of the concepts using R and SQL. Together, we will look at a few R code examples and tutorials, focused on empirical economic and business research, illustrating practical techniques to handle and analyse Big Data.

15h45 – 16h15 ***Coffee Break***

16h15 – 17h00 **“Large surveys and other continuous data streams in statistics production”**
Frauke Kreuter, LMU Munich and University of Maryland

This presentation will highlight two case studies and discusses challenges faced and lessons learned. The first example comes from surveying over hundred million people during the COVID-19 pandemic. The second example from comes from a data collection in the labor market context, where we applied passive measurements via an app to minute by minute actions of survey participants.

17h00 – 17h45 **“Thoughts on Working with Corporate Data”**
John Horton, MIT Sloan School of Management and NBER ***(Online)***

The talk will focus on the key considerations for researchers when it comes to getting research data, picking good research questions, managing collaborations, writing a research agreement, and dealing with disputes. The first step in conducting research is obtaining access to high-quality data, which involves making contact with potential data sources and negotiating access. The focus of the talk will be on how to make a compelling "pitch" the company is likely to be responsive to. Collaborating with companies can be mutually beneficial, but sometimes problems do arise. A well-drafted research agreement can help ensure that all parties understand their rights and obligations, and can provide a clear roadmap for resolving disputes if they arise. This talk will provide practical advice and guidance for researchers at all stages of the research process.

17h45 – 18h30 **Roundtable: “*The challenges of working with large empirical datasets in RDCs*”**
Lars Vilhuber, Cornell University and American Economic Association
Paulo Guimarães, BPLIM - Banco de Portugal
Pedro Campos, Instituto Nacional Estatística (INE)
Stefan Bender, Deutsche Bundesbank

Tuesday, December 20th, 2022

09h30 – 10h00 *Breakfast*

10h00 – 10h45 **“How to work efficiently with large datasets”**
Mauricio Caceres, Brown University

Methods and tools that work well with small datasets do not always scale to large (of even modestly-sized) datasets. We must take into consideration the increased memory usage of each operation and the often prohibitively slower runtime of even straightforward tasks. First, this presentation will give practical advice on how to deal with some common issues that arise with large datasets. We will then show how to speed-up various operations, focusing on Stata and the user-written package gtools, which provides a suite of fast commands. Last, we will discuss some examples where the largest speed-up results from using different (or custom) algorithms, rather than making your current algorithm run faster (i.e. working smarter, not harder).

10h45 – 11h30 **“Open source tools for really large data”**
Miguel Portela and Nelson Areal, Universidade do Minho

The purpose of this session is to present different ways of managing large data through the Arrow, DuckDB, and Spark environments. Using the information stored in the `parquet` format, we will discuss data management, both in memory, and out of memory when the size exceeds the RAM available. The illustration will be done in R using the `arrow`, `duckdb`, and `sparklyr` libraries. Data manipulation will rely on the `dplyr` and `dbplyr` packages.

11h30 – 12h15 **“Why is my computer so slow? How distributed computing can help you with data intensive workloads”**
Jannic Cutura, European Central Bank ***(Online)***

Big data --- datasets that are difficult to handle on standalone retail grade computers --- are rapidly becoming the norm in social science research. This is true both in academia as well as for policy oriented research in central banks and similar bodies. Yet traditional econometrics (and econometrics training) tells us little about how to efficiently work with large datasets. In practice, any data set larger than the researchers computer memory (~20-30GB) is very challenging to handle as, once that barrier is crossed, most data manipulation tasks becomes painfully slow and prone to failure. The goal of this presentation is to (i) explain what happens under the hood when your computer gets slow and (ii) show how distributed computing (in particular Hadoop/Spark) can help to mitigate those issues. By the end, participants will understand the power of distributed computing and how they can use it to both tackle existing data handling challenges and as well as new ones that were previously prohibitively expensive to evaluate on retail grade computers.

12h15 – 12h30 **Closing Remarks**

BIOGRAPHIES:

Frauke Kreuter, LMU Munich and University of Maryland

Frauke Kreuter is professor of statistics and data science at LMU Munich, professor of survey methodology at the University of Maryland and co-director of the (social) data science centers at the University of Maryland and the University of Mannheim.

Jannic Cutura, European Central Bank

Jannic Cutura is an economist turned data engineer turned software engineer who works as a python developer at the European Central Bank's Stress Test team. Prior to his current position he worked as research analyst/data engineer in the financial stability and monetary policy divisions of the ECB. He holds a masters and Ph.D. in quantitative economics from Goethe University Frankfurt and conducted research projects at the BIS, the IMF and Columbia University.

John Horton, MIT Sloan School of Management and NBER

John Horton the Richard S. Leghorn (1939) Career Development Professor and an Associate Professor of Information Technologies at the MIT Sloan School of Management. His research focuses on the intersection of labor economics, market design, and information systems. He is particularly interested in improving the efficiency and equity of matching markets. After completing his PhD and prior to joining NYU Stern School of Business in 2013, he served for two years as the staff economist for oDesk, an online labor market. He received a BS in mathematics from the United States Military Academy at West Point and a PhD in public policy from Harvard University.

Lars Vilhuber, Cornell University and American Economic Association

Lars Vilhuber is Executive Director of Cornell University's Labor Dynamics Institute. He is the Data Editor of the American Economic Association, Managing Editor of the Journal of Privacy and Confidentiality, and on advisory boards of restricted-access data centers in France, Canada, and the United States.

Mauricio Caceres, Brown University

Mauricio Cáceres Bravo is a PhD candidate from Brown University; his research fields of interests are labor and health economics. Mauricio has bachelor degrees in mathematics and economics from the University of Utah, and a master's degree in economics from Columbia University. Before joining the doctoral program at Brown, he worked for 4 years as a Research Assistant for professors at MIT and Yale. During this time he worked extensively with large datasets and developed the gtools package for Stata when he found speed (or a lack thereof) had become the primary bottleneck for his day-to-day tasks.

Miguel Portela, Universidade do Minho

Miguel Portela holds a PhD in Economics from the University of Amsterdam. He is Associate Professor with Habilitation at Universidade do Minho and the Vice Director of the Doctoral Program in Economics. He is also an affiliate of NIPE / U Minho, CIPES and IZA, Bonn. He has an ongoing collaboration with Banco de Portugal. He is a member of the Portuguese Council for Productivity. His research interests are labour economics, economics of education and applied econometrics. He published a set of articles, books and book chapters, with emphasis on publications in Econometrica, Scandinavian Journal of Economics, Regional Studies and Studies in Higher Education. He has research collaborations in different countries, leads and integrates research teams whose work has been funded by private and public entities. One highlights his recent FCT project "It's All About Productivity: contributions to the understanding of the sluggish performance of the Portuguese economy ". He has also written reports to define public policies on the Minimum Wage, Education and Employment in the Portuguese labour market. He has experience in consulting, both for private and public institutions.

Nelson Areal, Universidade do Minho

Nelson Areal is an Associate Professor of Finance at the School of Economics and Management, University of Minho. His research interests include risk measurement and forecasting, option valuation using numerical methods, performance measurement, socially responsible investments, and management education. He has a Ph.D. in Accounting and Finance, from Lancaster University, in 2006. He has provided data science consultancy for INE – Statistics Portugal. His career also includes two years (1992-1994) as an Information Systems Auditor at Ernst & Young. Co.

Pedro Campos, Instituto Nacional Estatística (INE)

Pedro Campos is currently Director of the Methodology Unit at Statistics Portugal in the Department of Methodology and Information Systems, in charge of Sampling frames, Methods and algorithms, Statistical Confidentiality, Small Area Estimation, among others.

Pedro Campos holds a Phd from University of Porto, where he is Assistant Professor at Faculty of Economics. His main courses are Statistics and Data Analisys, Data Mining, Marketing Analytics, and Agent-Based Modeling, connecting with his research interests: Data Science, Network Mining, Official Statistics, and Artificial Intelligence.

Pedro is also researcher at LIAAD (Lab. of Artificial Intelligence and Decision Support) at INESCT TEC.

Stefan Bender, Deutsche Bundesbank

Stefan Bender is Head of the Research Data and Service Center of the Deutsche Bundesbank and honorary professor at the School of Social Science at the University of Mannheim. Before joining the Deutsche Bundesbank Bender was head of the Research Data and Service Center of the Deutsche Bundesbank Federal Employment Agency, Germany at the Institute for Employment Research (IAB), where he developed one of the worldwide leading research data centers. His research interests are data access, data quality, merging administrative, survey data and/or big data, record linkage, unemployment, management quality. He has published over 100 articles in journals like the American Economic Review or the Quarterly Journal of Economics.

Ulrich Matter, University of St.Gallen

Ulrich Matter is an Assistant Professor of Economics at the University of St.Gallen. His primary research interests lie at the intersection of data science, quantitative political economics, and online media economics. In addition to his research activities, he has designed several new courses at the intersection of data science and applied econometrics (Big Data Analytics, Introduction to Web Mining, and Data Handling) has been honoured for excellence in teaching at the school and university level. Before joining the University of St. Gallen, he was a Visiting Researcher at the Berkman Klein Center for Internet & Society at Harvard University and a postdoctoral researcher and lecturer at the Faculty for Business and Economics, University of Basel as well as a postdoctoral guest researcher at the CABDyN Complexity Centre, Saïd Business School, University of Oxford.

BANCO DE PORTUGAL
EUROSYSTEM