BANCO DE PORTUGAL
EUROSISTEMA


# AUTOMATION OF THE RESEARCH PROCESS

## 18-19 DEC. 2023

All materials are available here.

## Programme

(GMT times)

### Monday, December 18th, 2023

**14:15 – 14:30** **Welcome**
Clara Raposo, Banco de Portugal

**14:30 – 15:10** **Producing automated tables using Stata**
Roxanne Connelly, University of Edinburgh

There are great benefits to automating the production of tables of statistical results. Automation will promote efficiency and reproducibility and reduce the opportunity for error. Cutting and pasting results is a recipe for disaster and should always be avoided. This session will introduce options for the automated production of tables in Stata. We will focus in particular on Stata's *–etable-* and *–collect-* commands, to create tables of descriptive statistics and regression results. By the end of this session, participants should have sufficient understanding of these commands to develop their own template to automatically produce tables with their required content and in their preferred style.

**15:10 – 15:50** **Dynamic documents in Stata**
Luiza Andrade, Development Innovation Lab

Dynamic documents combine code and text to automate the creation of outputs such as presentations, papers, and reports. This session will discuss when and why dynamic documents should be used, review some of the Stata tools for creating them, and demo one of these tools, called *Markstat*. Participants who wish to follow along with the demo should install *Pandoc* and *TeX/LaTeX* in advance of the session.

**15:50 – 16:30** **Coffee Break**

**16:30 – 17:10** **Visualizing complex and hierarchical data in Stata**
Asjad Naqvi, Austrian Institute for Economic Research

This session aims to introduce the audience to how to recognize, structure, and visualize multi-layer and hierarchical data. This includes creating more advanced visualizations such as treemaps, circle packing, sunburst graphs, sankeys, alluvial plots, and bump charts. We will also briefly showcase upcoming community-written packages.

**17:10 – 17:50** **Research workflow: Stata, Github, Overleaf and R**
***(Online)***
Alex Hollingsworth, Ohio State University

An overview of one possible research workflow. We will walkthrough how to integrate analyses and writing with the goal of ensuring reproducibility, minimizing errors, and lowering the cost of starting new projects/editing old ones. We will set-up an identical analysis in both Stata and R; go over how to share code, output, and data with co-authors and across machines using tools like GitHub, and Dropbox; and start a simple collaborative writing document in Overleaf. Example files will be provided, but no work prior to the workshop is required.

**17:50 – 18:30** **Research workflow with confidential data: The experience of BPLIM**
BPLIM Team

Since 2016, BPLIM has been facilitating researchers' access to confidential data. Our approach involves providing researchers with the necessary information for their analysis, which is then executed by BPLIM staff on the original data. Over time, we've implemented several adjustments to streamline the research process and improve the reproducibility of the workflow. In this session, we will share our vision on how to implement research with confidential data while satisfying the requirements for research reproducibility.

## Tuesday, December 19th, 2023

**09:00 – 09:40** **Reproducible research in R: A talk on how to do the same thing more than once**
Aaron Peikert, Max Planck Institute for Human Development

Computational reproducibility is the ability to obtain identical results from the same data with the same computer code. The high rate of irreproducible research limits the reach of results and decreases the efficiency of researchers. Reproducible research is a building block for transparent and cumulative science because it enables the originator and other researchers, on other computers and later in time, to reproduce and thus understand how results came about. In this session, we present an approach to automate the whole process from raw data to publishable manuscripts. This automation is possible by combining dynamic document generation (via *R Markdown*), version control (via *Git*), workflow orchestration (via *Make*), and software management (via *Docker*). The resulting workflow is, hence, highly transferable across machines and time. These core properties of reproducibility are demonstrated for any recipient by continuously and automatically reproducing the manuscript online. We highlight that this level of reproducibility enables new ways of collaboration and approaches to transparency in science.

**09:40 – 10:20** **Friends don't let friends copy-paste: Computationally reproducible APA-style manuscripts with the R package papaja**
***(Online)***
Frederik Aust, University of Cologne

When reporting quantitative results, researchers routinely resort to copy-paste reporting - they copy statistical results from their analysis software and paste them into a word processor. Copy-paste reporting is tedious: When the analytic approach evolves during manuscript preparation or revision, copy-pasting starts anew. More importantly, copy-paste reporting is error-prone as evidenced by the non-trivial rate of journal articles reporting inconsistent statistics (e.g., Brown & Heathers, 2016; Nuijten et al., 2016; Petrocelli, Clarkson, Whitmire, & Moon, 2013). Even with access to the original data, reproducing reported results is often difficult, if not impossible (e.g., Artner et al., 2020; Eubank, 2016; Naudet et al., 2018; Stodden, Seiler, & Ma, 2018; Vilhuber, 2020). Enter dynamic documents, an alternative to copy-paste reporting that saves time, minimizes errors, and improves computational reproducibility. Dynamic documents fuse manuscript and analysis scripts, automating the reporting of results. The R package *papaja* allows researchers to create dynamic, submission-ready, APA-style manuscripts and revision letters.

**10:20 – 10:50** **Coffee Break**

**10:50 – 11:30** **Reproducible Data Analysis at the Speed of Thought**
Jannik Buhr, Heidelberg Institute for Theoretical Studies

This session is all about workflows and tools. We will learn about *Quarto*, the next iteration of *Rmarkdown*, and integrate it into our workflow. We will see the targets R package in action to help us keep track of complicated analysis pipelines and produce verifiably reproducible results. Lastly, we will venture into uncharted territory to find out firsthand, why it always pays off to learn the intricacies of our tools with a live demonstration of using *Quarto* and targets in *Neovim*.

**11:30 – 12:10** **Styling documents and building extensions with Quarto**
Nicola Rennie, Lancaster University

*Quarto* is an open-source scientific and technical publishing system that allows you to combine text with code to create fully reproducible documents in a variety of formats. The addition of custom styling to documents can make them look more professional and recognizable. This session will give an overview of ways to customize HTML outputs (including documents and *revealjs* slides) as well as PDF documents. We'll also discuss the use of *Quarto* extensions as a way of sharing customized templates with others, demonstrate how to install and use extensions, and show the process of building your own custom-style extension. The *PrettyPDF Quarto* extension will be used as an example.

**12:10 – 12:50** **WORCS: A Workflow for Open Reproducible Code in Science**
Caspar van Lissa, Tilburg University

The Workflow for Open Reproducible Code in Science (WORCS) is a step-by-step procedure to make a research project open and reproducible, compliant with the FAIR principles (Findable, Accessible, Interoperable, and Reproducible), and the TOP-guidelines (Transparency and Openness Promotion). WORCS is an easy workflow that can be used either in parallel to or in the absence of, existing institutional requirements - and yet can be extended to meet advanced users' needs. WORCS is based on universal reproducibility principles - which are relevant regardless of your preferred analysis software - but for R-users, the workflow is implemented in a

package that automates most steps. This session briefly introduces the principles of reproducibility, and then provides a live demonstration of creating a reproducible project using the WORCS R-package, with a focus on data sharing and alternative solutions when data cannot be shared; and auditing of analyses via continuous integration (i.e., analyses are automatically reproduced in the cloud).

**12:50 – 13:00** **Closing Remarks**

# BIOGRAPHIES:

## Luiza Andrade, Development Innovation Lab

Luiza Andrade is the Data Analytics Lead at UChicago's Development Innovation Lab. Her work focuses on incorporating non-traditional data sources into development research, promoting transparency and reproducibility in social sciences, and developing software tools to simplify research data work. Prior to joining DIL, she was a Junior Data Scientist at the World Bank's Development Impact Evaluation department. Luiza is a Brazilian national and holds a BA and an MSc in economics from the University of Sao Paulo.

## Roxanne Connelly, University of Edinburgh

Roxanne Connelly is a Senior Lecturer in Sociology and Quantitative Methods at the School of Social and Political Science at the University of Edinburgh. Her substantive research is focused on social inequalities with special emphasis on social stratification and the sociology of education. She is a specialist in the analysis of large and complex social survey and administrative data resources, and longitudinal data analysis techniques. She is also an advocate for the improvement of research transparency and reproducibility in the social sciences.

## Asjad Naqvi, Austrian Institute for Economic Research

Asjad Naqvi is currently working as a Senior Consultant to the World Bank on developing policy-relevant climate macro models for finance ministries, and as a Senior Climate Economist at the Austrian Institute for Economic Research (WIFO). From 2011-2013, he was the Research Director at the Center for Economic Research in Pakistan (CERP). He received his Ph.D. in Economics from the New School for Social Research (New York) (2007-2012) and his habilitation in Economics in 2020. His current research interests are exploring environment-economy linkages with focus on climate change, climate finance, climate policies, and their subsequent direct and indirect spillover impacts. He regularly develops data visualization packages for Stata and runs the Stata Guide blog on Medium.

## Alex Hollingsworth, Ohio State University

Alex Hollingsworth is an associate professor at The Ohio State University with joint appointments in the Department of Agricultural, Environmental, and Development Economics; the Department of Economics; and the John Glenn College of Public Affairs. He is a Research Associate at the National Bureau of Economic Research, a co-editor at the Journal of Policy Analysis and Management, and an associate editor at the Journal of Health Economics. Hollingsworth is an applied microeconomist who examines how regulations affect health with interests in environmental economics, population health, substance abuse, and access to care. His research has been published in outlets including American Economic Journal: Economic Policy, the Journal of Public Economics, and the Journal of Human Resources. His research has been covered by Scientific American, the Washington Post, CNBC, the Atlantic, VOX, and the Los Angeles Times. He also co-hosts a podcast, The Hidden Curriculum with Sebastian Tello-Trillo. The Hidden Curriculum aims to cover topics relevant to academic life with a focus on things that are not formally taught in graduate school.

## Miguel Portela, Universidade do Minho

Miguel Portela holds a PhD in Economics from the University of Amsterdam. He is Full Professor at Universidade do Minho, Director of the Centre for Research in Economics and Management (NIPE), and Co-Editor of the Portuguese Economic Journal. He is an affiliate of NIPE/UMinho, CIPES and IZA, Bonn. He collaborates with Banco de Portugal and is a member of the Portuguese Council for Productivity. His research interests include labor economics, the economics of education, and applied econometrics, with a particular focus on the Portuguese economy. He published articles, books and book chapters, emphasizing

publications in Econometrica, Labour Economics, Scandinavian Journal of Economics, Regional Studies and Studies in Higher Education. He has research collaborations in different countries and leads and integrates research teams whose work has been funded by private and public entities. One highlights his 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 labor market. In addition, he has experience in consulting, both for private and public institutions.

### Aaron Peikert, Max Planck Institute for Human Development

Aaron Peikert leads the "Formal Methods in Lifespan Psychology" research group at the Max Planck Institute for Human Development, drawing inspiration from the rich legacy of the Center for Lifespan Psychology founded by Paul B. Baltes. Aaron combines innovative tools and reevaluates traditional methods, harnessing interdisciplinary insights from software engineering and the philosophy of science. This dynamic approach is geared towards advancing the methodological foundations essential for psychological research. A graduate with a BA, MA, and Ph.D. in Psychology from Humboldt University and an honorary researcher at University College London, Aaron is a staunch advocate for transparent knowledge dissemination and rigorous research methodologies. He is not above employing LLMs for writing a bio sketch.

### Frederik Aust, University of Cologne

Frederik Aust is a postdoctoral researcher in the Research Methods and Experimental Psychology group at the University of Cologne. He develops and tests mathematical models of human cognition using Bayesian statistics. In addition, he tries to improve the computational reproducibility of psychological science.

### Jannik Buhr, Heidelberg Institute for Theoretical Studies

Jannik is a Ph.D. student in the Molecular Biomechanics group of Prof. Dr. Frauke Gräter at the Heidelberg Institute for Theoretical Studies (HITS), where studies the chemistry of collagen under physical stress using computer simulations and quantum mechanical calculations. He enjoys thinking about mental models and teaching concepts with a focus on data analysis with the programming language R. When he's not out climbing or slacklining, you can probably find him writing another plugin for Quarto or the code editor Neovim to accommodate a wordplay about otters.

### Nicola Rennie, Lancaster University

Nicola Rennie is a lecturer in Health Data Science, based in the Centre for Health Informatics, Computing, and Statistics in the medical school at Lancaster University. She has a background in statistics and operational research, with a focus on applications of time series analysis and machine learning to health data. Nicola is particularly interested in reproducible research and software development and is the author of several R packages. She is a regular contributor to the R community through speaking at meetups, mentoring within the R4DS community, and blogging about data science.

### Caspar van Lissa, Tilburg University

Caspar van Lissa is an associate professor of social data science at the Department of Methodology & Statistics, chair of the Open Science Community Tilburg, and member of the Tilburg Young Academy. His research addresses the epistemological implications of machine learning for theory formation in the social sciences, evidence synthesis (summarizing existing research quantitatively and qualitatively), and open reproducible science. He is an advocate for open-source research software and has (co-)authored ten R-packages.