

International Conference “Use of R in Official Statistics (uRos)”

– Book of Abstracts

uRos2026

Paris, France | 18–20 November 2026

Published by the Ecological University of Bucharest
Bucharest, 2026

Publication information

International Conference “Use of R in Official Statistics (uRos)” – Book of Abstracts

Published by the Ecological University of Bucharest
Bucharest, 2026

1G Vasile Milea Bvd.
Bucharest 061341
Romania

Tel. +40722743670
Email: urosconf@gmail.com
Website: https://uros.r-project.ro/

ISSN: 3153–5143

Placeholder, to be replaced with the official ISSN once assigned.

First issue of the serial: 2018

This volume contains the abstracts for uRos2026, held in Paris, France, on 18–20 November 2026. The editorial and publication location is Bucharest, Romania.

Editorial Board

Editor-in-Chief

Ciprian-Antoniade Alexandru
Ecological University of Bucharest, Romania

Managing Editor

Ana Maria Ciuhu
Statistics Romania and Institute of National Economy, Romanian Academy

Editorial Board Members

Anna Smyk
Statistics France

Alexander Kowarik
Statistics Austria

Mark van der Loo
Eurostat and Leiden University

Bogdan Oancea
University of Bucharest, Romania

Editorial information: https://uros.r-project.ro/editorial-board.html

Contents

Editorial Board 3

Keynotes 5

Tutorials 6

Regular Presentations 11

Lightning Talks 49

EMOS Presentations 61

Author Index 65

Practical Conference Information 68

Presenting authors are underlined. Superscript numbers identify the affiliations supplied with each submission.

Keynotes

Practical AI for Data Science

Simon Couch

Posit PBC

How do we build competent data analysis agents? Data analysis requires a willingness to pause, question conclusions, dig into subtleties, and sit with uncertainty. Frontier LLMs, however, are optimized to push tasks toward completion rather than refraining from answering unanswerable questions. Drawing on his experience building data analysis agents at Posit, Simon will share evaluations that expose where LLM-driven analysis goes wrong and design patterns that keep analyses correct, transparent, and reproducible.

“R and Open-Source: A love story”

Rebecca Killick

Clemson University

Different people have different motivations for starting and continuing to contribute to open-source software. During the last 20+ years, I have gone through various periods of love, hate, joy, and frustration with R and the wider open-source community. I have found myself, entering my mid-career phase and have been enjoying reflections on how I got to where I am currently, and the pivot points along the way. Open-source has been a constant throughout and that was no accident, but things could have been very different. In this talk, I will share some of these reflections with you and discuss the different ways that we can all make our own contributions, whether for the short-term or as a career path.

Tutorials

From Data to Models: Hands-On Machine Learning with R

Pérez Bote Adrián[1], Barragán Andrés Sandra[1]

1 - Spanish Statistical Office (Spain)

This tutorial provides a practical introduction to Machine Learning using the R programming language, combining essential theoretical concepts with hands-on implementation. Over the course, participants will explore the fundamentals of supervised learning, with a focus on both regression and classification tasks, while gaining direct experience building and evaluating models in R. The session covers key families of predictive models widely used in applied data science. Participants will begin with linear models, understanding their assumptions, interpretation, and use cases for continuous target variables with the aim of understanding key concepts of Machine Learning techniques. The tutorial then progresses to decision trees with categorical targets, highlighting their intuitive structure and flexibility. Then, the concepts and implementation of bagging techniques, including Random Forests, will be discussed, emphasizing variance reduction and robustness. Finally, boosting methods will be introduced, illustrating how sequential learning can enhance model accuracy by addressing previous errors. A strong emphasis is placed on practical application. Attendees will work through guided coding exercises using both widely adopted CRAN packages and custom-developed tools, including the RMLUtils object-oriented R package, which facilitates streamlined workflows with a common interface for model training, validation, and comparison. Through these exercises, participants will learn how to preprocess data, fit models, tune hyperparameters, and assess model performance using appropriate metrics. By the end of the tutorial, participants will have a solid understanding of core supervised learning techniques and the ability to implement them effectively in R. This course is designed for practitioners and analysts seeking to strengthen their machine learning skills in a reproducible and efficient programming environment, bridging the gap between theory and real-world application.

Getting Started with Positron: A Next-Generation IDE for Data Science

Cetinkaya-Rundel Mine[1,2], Dervieux Christophe[2], Michonneau François[2]

1 - Duke University, Department of Statistical Science (United States), 2 - Posit, PBC (United States)

Positron is a next-generation data science IDE that combines the best features of RStudio and Visual Studio Code. This tutorial will introduce R users to Positron's core capabilities, with a special emphasis on helping RStudio users while highlighting its seamless integration with the R ecosystem. For R programmers coming from RStudio, Positron delivers a familiar yet enhanced environment for data analysis and package development, while offering a path to Python when needed. Unlike traditional software-development-oriented IDEs, Positron provides first-class support for data science-specific workflows through native R support (via the Ark kernel), along with designated areas for Variables (Environment), Connections, Plots, Help, and more that RStudio users have come to rely on. During this hands-on tutorial, participants will learn how to: - Install, configure, and update Positron for an R-focused workflow - Navigate Positron's interface and understand how it compares to RStudio - Work with R scripts and Quarto documents in Positron for interactive R coding and data exploration - Try Posit Assistant for core data science workflows, including exploratory analysis, data cleaning, data app development, and modeling - Customize Positron with useful settings, extensions, and keyboard shortcuts We'll explore Positron's innovative features that enhance R productivity, such as the improved interactive Data Explorer and the ability to switch between different R installations. The tutorial will include practical demonstrations of key workflows, such as developing and publishing Shiny apps and Quarto documents. For R users curious about Python, we'll briefly demonstrate how Positron makes Python accessible within a familiar environment. By the end of this tutorial, participants should understand how to perform their most-used workflows in Positron and tailor the IDE to their specific needs.

Managing statistical disclosure control from microdata to protected linked tables with rtauargus

Ferrer-Pradines Nadège[1], Desclodure Julien[1]

1 - Institut national de la statistique et des études économiques (INSEE) (France)

The rtauargus package provides an R interface for τ-Argus, a reference tool to apply Statistical Disclosure Control to tabulated (aggregated) data. While τ-Argus is a tool designed to be used via a Java GUI or called using a batch-file (script-file), rtauargus package allows to create inputs (rda, arb, hst and tab files) from data in R format. It generates the sequence of instructions to be executed in batch mode (arb file) and launches a τ-Argus batch in command line, then retrieves the results in R. These different operations can be executed in one go, but also in a modular way. They allow to integrate the tasks performed by τ-Argus in a processing chain written in R. The package presents other additional functionalities, some of them unique to this package, such as managing the protection of linked tables (tables sharing common cells that must be protected consistently) or a tool for the analysis of demand (from a dissemination template, detecting clusters of tables to be treated together as linked tables). The package also allows to create a hierarchical variable from correspondence table or to build tabular data fit to be used with rtauargus from microdata. In this tutorial, we will follow a dissemination project, from microdata and dissemination template to protected linked tables, including non-nested hierarchies. Installation of τ-Argus on a Windows OS will be necessary to follow this tutorial step-by-step.

Mapping Official Statistics with the tmapverse and cols4all

Tennekes Martijn[1]

1 - Statistics Netherlands / CBS (Netherlands)

Maps are among the most effective ways to communicate official statistics, but producing them well requires both the right tools and careful design choices. This tutorial introduces the tmapverse: the tmap package together with its growing family of extensions. With tmap (version 4), participants learn to build thematic maps using a layered, grammar-of-graphics approach that scales from quick exploration to publication-ready output, in both static and interactive form. We then move beyond standard choropleths into the extension packages: tmap.cartogram for area-distorting cartograms, tmap.glyphs for small inset charts such as donuts and flowers, tmap.networks for network data, and tmap.mapgl for modern web-based interactive maps. A central theme is color. For official statistics, palette choice is not cosmetic: it affects readability, accessibility for people with color-vision deficiency, and the trustworthiness of the message. We introduce cols4all, an analysis tool for exploring and evaluating color palettes against criteria such as color-blindness friendliness, perceptual uniformity, and contrast, and show how to feed well-chosen palettes directly into tmap. The tutorial is hands-on. Participants work through realistic examples with regional and spatial statistical data, leaving with a reusable workflow for turning their own data into clear, accessible, and reproducible maps.

Opening the Black Box: Interpreting Machine Learning Models with SHAP Values in R

Panaiotis Thelma[1]

1 - Pyranhia (France)

Machine learning models are increasingly used in official statistics, but their lack of transparency is a real problem: when a model influences a policy decision or flags an anomaly in administrative data, people reasonably want to know why. This tutorial is about practical tools for answering that question: understanding what a model has learned, which variables it relies on, and how it arrives at individual predictions. There are several approaches to model interpretation. Variable importance tells you which features matter globally; partial dependence plots show how predictions change as a single variable varies. Both are useful, but neither explains individual predictions, and neither handles feature interactions well. SHAP values do both: rooted in cooperative game theory, they decompose each prediction into variable contributions in a consistent, theoretically grounded way. Another underrated advantage: unlike permutation-based methods, SHAP values do not require true labels, which means model behaviour can be audited on fresh, unlabelled data. This two-hour session focuses on SHAP in depth rather than trying to cover everything superficially. Each concept is introduced briefly, then immediately applied in R on a real dataset. Starting from a trained gradient boosting model on a structured socioeconomic dataset, we will build intuition for what SHAP values actually measure. The session covers two levels: individual predictions (waterfall plots) and global model behaviour (beeswarm summary plots). Pre-computed SHAP values for the full dataset will be provided for computational purposes, but we will also go through how to compute them for a single new observation. The tutorial closes with an honest look at SHAP’s limitations: what correlated features do to attribution, what SHAP can and cannot tell us about causality, and what to keep in mind when using these tools in a reporting or production context. No installation needed: participants will work on a shared RStudio Server with everything pre-configured. Prerequisites: intermediate R, comfortable with the tidyverse; basic ML concepts (train/test split, model fitting). No prior knowledge of SHAP required. Packages used: tidymodels, tidyverse, DALEX, DALEXtra, shapviz, ggplot2

R in the Age of LLMs: From AI-Assisted Development to Chat-Based Data Applications

Michonneau François[1]

1 - Posit, PBC (United States)

LLMs are transforming how we develop applications and interact with data. This 3-hour, fully hands-on workshop will teach you to (1) leverage LLMs to help you build, debug, and implement new features when coding in R; and (2) build a Shiny web application that lets users query and visualize data using natural language. No setup is required: we will use a pre-configured environment ready to use, so you can focus on learning. Using Positron and its AI integration, Posit Assistant, you will experience how LLMs can accelerate R application development. You will get to experience how Posit Assistant can help debug code and implement new features, and how it can be used to guide you in the initial exploration of a new dataset. In the second part of the workshop, you will build a Shiny application integrating the querychat package, giving users a chat-based interface to interact with data. Along the way, you will develop skills applicable to more complex, real-world applications. The workshop concludes with deploying your Shiny application to Connect Cloud, a hosting platform for R and Python applications with a free-tier plan, so your work is immediately shareable. For this workshop, we will use Positron, a free, source-available, IDE for R and Python designed to support the full spectrum of data science work. Prerequisites and intended audience Prerequisites: Basic familiarity with R (comfortable reading and writing simple R scripts) Some exposure with Shiny is helpful but not required No prior experience with AI tools, LLMs needed A laptop with web browser Intended audience: This workshop is aimed at R users who are curious about using LLMs for development and into their applications but have not done it before. This workshop is also a good opportunity to learn how to use Positron and its AI features Outline of the tutorial Introduction to the workshop + LLM overview (20 min) Part 1: AI-assisted coding and exploratory data analysis in Positron Tour of Positron, its layout, brief overview of its functionalities and extension system, and its data explorer (15 min) Using Posit Assistant for debugging: learn how Positron AI can fix errors in your code (15 min) Using Posit Assistant for implementing new features: create a new visualization or table from your data (20 min) Explore and learn from your data (30 min) Q&A for Part 1 (20 min) Part 2: Integrate a chat-based interface in your Shiny app Overview of a Shiny app (15 min) Building a Shiny app (15 min) Integrating querychat into your Shiny app (20 min) Use Posit Assistant to implement new features into your app (30 min) Part 3: Publish your application to Connect Cloud (20 min) Workshop wrap-up and Q&A (20 min)

Spatial Analysis and Cartography with R

Antunez Kim[1]

1 - Insee (France)

[2-hour tutorial] This tutorial offers an introduction to the main packages and tools available in the R ecosystem for manipulating spatial data and making thematic maps, with a focus on vector data. A large place will be devoted to the handling of the sf package and the realization of simple geoprocessing. We will also cover making static thematic maps with ggplot2 and mapsf as well as using mapview to easily make interactive data mining maps. Finally, a short introduction to more specific tools and questions will complete this introduction: raster data, spatial smoothing, geocoding, Open Street Maps data, etc. The majority of the workshop will be devoted to the application of the various concepts and tools on the annual database of road traffic injury accidents.

Tidy Survey Analysis in R: Working with Weights and Sampling Designs for Official Statistics

Zimmer Stephanie[1], Powell Rebecca[2], Velásquez Isabella[3]

1 - Research Triangle Institute International (United States), 2 - Fors Marsh (United States), 3 - Posit (United States)

Many official statistics come from surveys capturing attitudes, behaviors, and other information from the people and businesses in the country. As it is often impractical to field a census each time a country is interested in data, researchers turn to sampling a subset of the population and calculating weights for subsequent analysis. Using these official statistics data without incorporating the weights and sample design variables can result in biased estimates, so it is important to know how to properly apply weights to accurately analyze the data. This interactive tutorial will introduce a unifying workflow for analyzing survey microdata (such as official statistics data) with weights in R. We will begin with descriptive analysis, obtaining weighted proportions, means, and quantiles from survey data. Then, we will move into statistical testing, using t-tests for comparing means and χ-squared tests for comparing proportions. Finally, we will discuss common probability sampling designs and how to create survey design objects from existing survey data in R to accurately account for these designs. The tutorial will include time for exercises using real-world data, so you can get hands-on experience. We will also provide example code for creating survey design objects for various publicly available official statistics datasets. For the best learning experience, we recommend you have some prior experience with R and the tidyverse, including familiarity with mutate, summarize, count, and group_by. We will be using Posit Cloud, so you do not need to have R or RStudio preinstalled on your computer.

Working with categorical data and codelists in R

Van Der Laan Jan[1], De Jonge Edwin[1]

1 - Statistics Netherlands / CBS (Netherlands)

Categorical variables are very common in official statistics. Most of the output statistical offices produce are tables where the population of interest is summarised according to different categories. Examples are age categories, economical activities (NACE), educational level and many more. The possible values of these categories are often stored in code lists. These consist of a set of permissible codes with associated meta data such as labels, descriptions and hierarchies. In this workshop we will show the different ways of working with categorical variables and code lists in R. We start with the basics: the factor and different methods in base R for working with these. The forcats package that proves methods that make some common operations such as recoding easier. The labelled type used by the haven package which provides basis support for code lists. And, finally, the codelist package which offers more extensive support for working with code lists. The workshop is aimed at R-users of all levels and most of it will be hands-on in the form of demonstrations in combination with short exercises. No prior knowledge other than basic R skill is required. Participants need to install the forecast, haven and codelist packages. All are available on CRAN.

Regular Presentations

110 billion euros, one deadline, zero regression: migrating a critical calculation service to R with AI

Vincent Guyader[1], Matthieu de Canteloube[2]

1 - ThinkR (63 rue de la pigaciere 14000 CAEN France), 2 - Theodo (France)

Migrating legacy statistical systems to R is a growing challenge for public institutions facing the end of life of proprietary software. The ATIH (French Technical Agency for Information on Hospital Care) faces this challenge in a critical context: its calculation service, which orchestrates the distribution of 110 billion euros of the national health insurance budget allocated to hospitals, relies on around sixty legacy procedures that must be migrated to R by the end of 2026. Theodo and ThinkR co-designed a three-step migration method, built on two open-source tools developed for the occasion. DATALineage statically analyzes the legacy code to map data flows and extract business rules. This structured representation is passed to Claude Code, which rewrites each procedure in R within a constrained framework: target architecture (Docker containers on Kubernetes), maintainable coding conventions, and an automatic feedback loop driven by datadiff. This open-source R package, a layer on top of pointblank, compares the outputs of each original procedure with its R equivalent, giving the AI a precise signal to self-correct. The same validation is then applied across all French healthcare facilities to guarantee non-regression in production. We will present the complete method, the lessons learned from the first migrations, and the metrics obtained on real data from the French hospital system. We detail the three steps: DATALineage to map the existing code, since one must understand what the code actually does before rewriting it; Claude Code for constrained rewriting, where the model receives a precise specification (infrastructure constraints, expected conventions, output format rules) and iterates on datadiff's quantitative feedback until convergence without human intervention; and datadiff for two-level non-regression, using configurable YAML rules (numerical tolerance, text normalization, row-count validation, missing-value handling) both as a feedback loop during development and as a production safeguard on real data. At submission time, the procedures are still being migrated and non-regression metrics on real data are encouraging. uRos2026 will be the opportunity to present the full results of the at-scale deployment, the productivity gains compared to a manual migration, the edge cases encountered (ambiguous business rules, undocumented behaviors), and how the method handles them. Both tools, DATALineage and datadiff, are open source and reusable in any context of legacy statistical code migration, whatever the institution.

3D Thematic Mapping in R with tmap v4 and tmap.mapgl

Martijn Tennekes[1]

1 - Statistics Netherlands / CBS (Netherlands)

Thanks to a thriving R-spatial community, the tools available for mapping in R are now genuinely state-of-the-art, and they increasingly rival dedicated GIS and web-mapping platforms. This talk shows where that frontier currently sits for thematic mapping in official statistics, and introduces 3D mapping as a practical addition to the statistical cartographer's toolbox. We start with tmap version 4, a complete redesign of the package built around a flexible, layered mapping grammar that supports both static and interactive output. We then focus on tmap.mapgl, an extension that brings the MapLibre and Mapbox GL rendering engines into tmap, enabling fast, modern, web-based interactive maps directly from R. The centerpiece is 3D thematic mapping in tmap.mapgl's """"maplibre"""" mode, where statistical values are encoded as the height of extruded polygons. This works surprisingly well in practice and addresses a long-standing problem with choropleths: area bias. Population density, in particular, is notoriously hard to map as a choropleth, because large, sparsely populated regions visually dominate small, dense ones. Encoding density as height instead of (or alongside) color gives a far more faithful picture. We illustrate this with real statistical data and discuss when 3D maps clarify a message and when they risk obscuring it.

A Comparison of Model-Based and Unsupervised Learning Methods for Detecting Scaling Errors in R

Marcello D'orazio[1], Cristina Faricelli[1], Renato Magistro[1]

1 - Italian National Institute of Statistics (Italy)

Numeric data editing begins by detecting systematic errors, which often are due to misreported units of measurement or scaling errors (e.g., values in millions instead of thousands, or kilograms instead of metric tons, etc.). If uncorrected, these errors severely bias final statistics. They are generally detected by comparison with past data (panel surveys), external data, etc. The R package UnitMix (Faricelli and Magistro, 2026), recently developed at Istat, implements Gaussian mixture model–based clustering (Di Zio et al., 2005), where each cluster represents a distinct error pattern. Users should pre-specify the number of clusters based on the expected error structure. The method requires the availability of an error-free variable that is strongly correlated with the variables affected by errors. Within this framework, assigning a unit to a cluster amounts to identifying its error pattern and determining the corresponding correction. Since unit measurement errors often result in substantial over- or under-reporting, they can also be detected using outlier detection methods during the selective editing phase. At Istat, this is supported by in-house R packages such as univOutl (D'Orazio, 2026) and SeleMix (Guarnera et al., 2026). More recently, D'Orazio (2023a, 2023b) investigated the use of isolation forest (IF), an unsupervised machine learning method for anomaly detection (Liu, 2008), and its multivariate extension, Extended Isolation Forest (EIF; Hariri et al., 2018). IF is scale-invariant, does not rely on distributional assumptions, and is effective when anomalies are relatively rare; it is also reported to perform well under skewed or multimodal distributions. In R, the standard IF is implemented in the solitude package (Srikanth, 2021), while isotree (Cortes, 2026) also provides support for EIF. We compared the IF implemented in isotree, with the model-based method in UnitMix using some real-data. Preliminary results indicate that UnitMix is effective in detecting scaling errors and remains reasonably robust under deviations from Gaussian (or log-normal) assumptions, although its performance is strongly influenced by user-specified inputs, particularly the assumed number and structure of error patterns. In contrast, the standard EIF approach appears less effective: high anomaly scores do not directly identify the affected variables, requiring additional sensitivity analyses and robust data standardization. EIF also performs poorly with highly skewed distributions unless log transformations are applied, and ratio transformations are often needed to better exploit error-free variables. Further research is needed to assess whether the two approaches can be combined to leverage their complementary strengths while mitigating their limitation.

A wavelet protection method for publishing maps

Edwin de Jonge[1]

1 - Statistics Netherlands / CBS (Netherlands)

Cartographic maps often provide an insightful means for publishing regional and spatial data. Their essential feature is that is make spatial patterns visibile. While many ofifical statisitcs concern administrative regional data, more and more details spatial data is available. Because adminstrative regions may obscure detailed spatial distributions, policymakers often are interested in detailed maps which show spatial patterns. However a detailed map may introduce disclosure of individuals, so it needs to be protected for that. Several methods are available, e.g., R package sdcSpatial provides a kernel density smoothing and quadtree method. The presentation will focus on a new method based on wavelet decomposition, that addresses issues with those methods. We will present the method, compare it with the other methods, and show how to use it within R.

AccScript: an R-based domain-specific language for reproducible national accounts production

Iliass Taleb[1]

1 - Haut Commissariat au Plan (Morocco)

National accounts production requires the periodic balancing of large volumes of structured data under strict methodological constraints and short production deadlines. In many statistical offices, compilation workflows rely on expert-driven repeated actions in specialized national accounting software and graphical interfaces. These workflows can make reproducibility, systematic validation, audit trails, and comparison across years and iterations difficult. Tracking decisions often requires additional reports outside the production workflow or depends on individual memory, which adds workload and can weaken institutional traceability. AccScript is an R-based domain-specific language and execution engine designed to support reproducible national accounts production. It allows national accountants to express compilation and balancing logic in readable scripts, using concepts that are directly aligned with the production process: production, imports, taxes, margins, intermediate consumption, final consumption, gross fixed capital formation, changes in inventories, exports, values, volumes, and price indices. The language is designed to remain close to the terminology and logic used by national accountants, so that scripts can be understood, reviewed, and discussed by production experts without requiring advanced programming knowledge. Basic arithmetic, familiarity with the production concepts, and a general understanding of scripted instructions are sufficient to use AccScript. For example, for each product, or for each branch depending on the level of aggregation considered, the accountant writes explicitly the logic applied to the corresponding supply-use table. The R engine then parses and validates the AccScript script, applies the instructions to structured extraction files, produces controlled output tables, generates reintegration files, and records modified cells for traceability. AccScript is designed to operate around an existing national accounts production system. Its role is to add a reproducible scripting and control layer to environments where structured input and output data already exist and can be mapped consistently with System of National Accounts concepts. To showcase the approach, AccScript is applied to product-level Supply–Use Table balancing in an ERETES-style national accounts environment, corresponding to the tool currently used in the Moroccan official statistics production context. For each product, AccScript applies explicit balancing instructions to value and volume data, recalculates derived indicators, and produces outputs that can be reviewed, rerun, compared across years and iterations, and reintegrated into the production workflow. However, the design is not tied to a specific software package. It relies on configurable mapping files, making it adaptable to other national accounts production systems that expose structured data. This contribution presents the motivation, language design, R implementation, validation logic, and a worked Supply–Use Table balancing example. It shows how R can move beyond analysis and reporting in official statistics to provide a controlled, auditable, and reproducible language layer for statistical production in national accounting.

Aggregating incomplete data: accounting for hierarchical relationships

Viktoria Kis[1], Erika Lee[1], Simon Normandeau[1], Sara Otavo Peña[1]

1 - Directorate for Education and Skills (France)

Education finance statistics at the OECD, collected from over 40 countries once a year, share a recurring problem: countries cannot always report expenditure in the categories requested. A local breakdown may be folded into a regional total, or pre-primary spending grouped with primary. These inclusion relationships are signaled in the submitted Excel questionnaires, but handling them in aggregations has long relied on manual corrections and expert judgement – a process that is slow and hard to reproduce. Existing tools were not designed for this problem. Excel does not naturally handle hierarchical data across multiple dimensions. While existing R packages (e.g. data.tree, igraph) address related needs (e.g. visualise our data), none directly supports aggregation over incomplete, partially nested data. We developed an R package that treats inclusion relationships and missing values as first-class operands rather than annotations to be resolved before processing. The package reconstructs hierarchies from the bottom up, starting from the most granular data available and falling back to the smallest available branch when a breakdown is missing. Because these relationships travel upward as the tree is built layer by layer, the composition and quality of every aggregate is fully traceable. We will present the design principles behind the package, key decisions made to align its vocabulary with existing R packages, and lessons learned from developing alongside live data submissions. The approach generalises to any domain where hierarchical data is reported incompletely across multiple dimensions.

An AI-Powered Analysis Assistant for Large-Scale Official Statistics

Oliver Hauke[1]

1 - Destatis (Germany)

Official statistics face the constant challenge of processing significantly increasing data volumes while simultaneously meeting the demand for increasingly complex analyses from researchers across diverse disciplines. Advances in data engineering and artificial intelligence (AI) offer new opportunities to address these challenges and improve the efficiency of statistical production and research workflows. The Federal Statistical Office of Germany (Destatis) significantly increased its hardware resources and investigated suitable analytical technologies, including the R packages arrow, duckdb, and polars in combination with the Parquet file format, for highly efficient computing. The suitable combinations of tools provide significant performance improvements while maintaining a user-friendly analytical environment. Based on these findings, the Research Data Centre (RDC) at Destatis developed coding templates that enable researchers to efficiently analyze large datasets while complying with coding guidelines and output requirements. However, these templates are typically tailored to specific statistics and predefined analytical tasks. Recent advances in generative AI and the availability of secure public-sector AI infrastructure create new possibilities for more flexible analytical support. Destatis can leverage both KIPITZ, the Federal Government's sovereign AI platform developed by ITZBund, and local GPU resources for deploying AI models within a secure environment.This paper presents an AI Analysis Assistant that has access to tools and skills using efficient large-scale data processing methods. Through a conversational interface, the assistant can translate complex research questions into optimized analytical workflows, query large datasets within seconds, and generate export-ready results. Researchers can review, modify, and extend the generated analyses, ensuring transparency and reproducibility. Integrated into modern agent frameworks, the assistant also provides autonomous implementation of research projects, significantly reducing the effort required for large-scale statistical analyses.

An R-Based Nowcasting Dashboard for Industrial Production Index: Comparative Analysis of Classical and Machine Learning Methods

Duygu Dedeoglu[1], Osman Sert[2]

1 - Duygu Dedeoglu (Turkey), 2 - Osman Sert (Turkey)

Accurate nowcasting of the industrial production index (IPI) is essential for economic monitoring, policy analysis, and short-term decision-making. This study presents an interactive nowcasting dashboard, developed using the R-Shiny package, designed to analyze and predict the IPI by integrating both classical econometric and machine learning approaches within a unified framework. The proposed system incorporates various nowcasting approaches, including Autoregressive (AR), Autoregressive Integrated Moving Average (ARIMA), Autoregressive Distributed Lag (ARDL), two-variable Vector Autoregressive (VAR2), three-variable Vector Autoregressive (VAR3), and Extreme Gradient Boosting (XGBoost) models. For the VAR2 and VAR3 models, all possible pairwise and triplet combinations of the auxiliary variable set were utilized for nowcasting. To determine the most accurate approach for IPI nowcasting, the out-of-sample performances of these models were comparatively evaluated based on the Root Mean Square Error (RMSE) criterion. The dashboard was implemented using R-based analytical and visualization tools, enabling automated model estimation, nowcast generation, performance comparison, and interactive graphical reporting. Users can dynamically examine nowcast outputs, compare model performances, and monitor prediction behavior through a single integrated interface. This structure provides a reproducible and flexible environment for applied economic nowcasting/forecasting studies. Empirical results reveal that nowcasting performance varies according to model structure and information usage. While traditional time series models provide stable and interpretable nowcasts, machine learning approaches such as XGBoost exhibit competitive predictive accuracy, particularly in capturing nonlinear relationships within the data. Furthermore, VAR-based multivariate models show strong nowcasting performance when incorporating auxiliary economic indicators into the system. The study highlights the advantages of combining econometric and machine learning techniques within an interactive open-source framework. In addition to its empirical findings, the proposed dashboard demonstrates how reproducible R-based applications can support modern forecasting workflows, comparative model evaluation, and data-driven economic analysis. Keywords: Industrial Production Index, forecasting, nowcasting, ARIMA, VAR, ARDL, XGBoost, machine learning, dashboard, R

Assessing Direct and Indirect Seasonal Adjustment Approaches for Labour Cost Indices: A Serbian Case Study

Jelena Suzić[1], Melinda Tokai[1], Dunja Katić[1], Ana Markovic[1]

1 - Department for Time Series Analysis and Composite Indicators Development (SORS) (Serbia)

The increasing demand for timely and analytically relevant labour market indicators within the framework of evidence-based policymaking has highlighted the importance of high-quality time series analysis in official statistics. In Serbia, ongoing activities related to the development of the Decision-Making Support System (DMSS) under the EU4SORS project have further emphasized the need for reliable seasonally adjusted indicators that can support analytical and decision-making processes. This paper will present a practical case study of seasonal adjustment applied to Labour Cost Indices (LCI), focusing on the comparison of direct and indirect adjustment approaches. The analysis is being conducted using JDemetra+ and R, allowing for a comprehensive assessment of the impact of both approaches on seasonal patterns, coherence between components and aggregates, revisions, and the overall quality of adjusted series. A detailed comparative analysis will be performed for selected labour cost components and aggregated indices in order to identify the most appropriate approach for regular production. Particular attention will be given to methodological aspects related to aggregation consistency, revision behaviour, treatment of calendar effects, and the practical implementation of seasonal adjustment within an official statistics production environment. The paper will also discuss methodological considerations related to Eurostat recommendations and the use of R for supporting diagnostics, revision analysis, and quality assessment. The findings are expected to contribute to the broader objective of strengthening the analytical capacity of official statistics and improving the availability of high-quality labour market indicators for policy analysis and decision support. The analysis is currently ongoing; the full paper will present final results and practical recommendations based on the completed comparative assessment. The experience gained may be relevant for National Statistical Institutes and other organisations facing similar methodological challenges in the production of seasonally adjusted Labour Cost Indices.

Automated Public Science Communication: Building a Bluesky Bot for Real-Time Official Statistics Visualisation with R

Damien Dupre[1]

1 - Dublin City University [Dublin] (Ireland)

In an era of rapid digital communication, public institutions and researchers must find efficient ways to disseminate data insights directly to citizens. While traditional statistical reports often remain confined to static databases, social media platforms offer a direct channel for broader scientific engagement. This presentation introduces an automated workflow that connects official statistical data with the decentralized social media ecosystem using R. We present `csoie_bot` , an open source automated bot designed to fetch, process, and visualize official data from the Central Statistics Office (CSO) of Ireland, and dynamically publish these visual updates to Bluesky via the AT Protocol (https://github.com/damien-dupre/csoie_bot and https://bsky.app/profile/csoiebot.bsky.social). Built entirely within the R environment, the bot leverages modern data manipulation and visualization libraries to create clear, accessible graphics that translate complex public metrics into easily digestible visual content. The automated pipeline utilizes GitHub Actions to execute scheduled scripts, ensuring a continuous stream of public updates without requiring manual intervention or continuous local server hosting. This regular presentation will detail the technical architecture of the bot, focusing on how R interacts with modern web APIs and social media protocols. We will discuss the specific challenges of automated data visualization, such as maintaining design consistency across varying data formats, managing API authentication securely in public repositories, and configuring robust error handling for scheduled cloud execution. By sharing the development process and open source repository of `csoie_bot`, this talk aims to provide a reproducible template for data scientists, statisticians, and researchers looking to automate their public science communication and maximize the real-world impact of their data insights.

Automating Educational Statistics Production Processes Using R and Shiny: A Case Study from Turkish Statistical Institute

Fethi Saban Ozbek[1]

1 - Turkish Statistical Institute (Turkey)

Official educational statistics play a critical role in shaping national policies and measuring socioeconomic development. Traditionally, compiling and analyzing certain steps within the production process of educational statistics retained a reliance on manual steps using spreadsheet software like Microsoft Excel, alongside distinct platforms such as SAS and Toad for Oracle to a certain extent. Standardizing and modernizing these resource-intensive workflows across disparate tools offers significant potential for enhancing institutional agility. To modernize this data production infrastructure and enhance corporate efficiency, the Education Statistics Department at the Turkish Statistical Institute (TurkStat) initiated a national project focused on automating data production pipelines through the R programming language and Shiny package. Under this ongoing initiative, dedicated R projects and interactive Shiny applications have been successfully developed and deployed for key statistical products, encompassing national education statistics, formal education statistics, school life expectancy, higher education employment indicators, and human development index. Through the developed Shiny applications, data browsing and inspection workflows have been migrated to a user-friendly interface. Instead of dealing with complex database queries or manual data transfers, users can directly pull database tables by simply clicking a single button on the interface, instantly triggering all data processing steps via underlying R scripts. This integrated R and Shiny automation, replacing legacy workflows based on SAS, Toad for Oracle, and Excel, has yielded substantial advancements in statistical operations. By converting repetitive data cleaning, integration, formatting, advanced analysis, and visual presentation (visualization) tasks into reproducible R scripts, the project has effectively minimized the risk of human error and enhanced data accuracy. Furthermore, the automation of these routine workflows has unlocked considerable time and labor savings, enabling institutional resources and human capital to pivot away from routine processing and toward high-value analytical tasks, advanced modeling, and deeper quality assurance. Crucially, the accelerated processing timelines provide data production teams with enhanced operational visibility, allowing for proactive monitoring and timely interventions. This study presents the structural framework of TurkStat’s automation transition, documents the R and Shiny methodologies used to optimize heterogeneous systems, and discusses the broader organizational benefits on institutional efficiency.

Bayesian Small Area Estimation for Multinomial Health Indicators Using R: An Application to Official Statistics

Ioannis Nikolaidis[1], Stefanos Giakoumatos[2]

1 - Hellenic Statistical Authority (ELSTAT) (Greece), 2 - University of Peloponnese, Department of Finance and Accounting, Greece (Greece)

The increasing demand for reliable and timely statistics at detailed geographical and demographic levels has made Small Area Estimation (SAE) a key component of modern official statistics. Policymakers and public health authorities require disaggregated indicators to support evidence-based decision making, monitor inequalities, and efficiently allocate resources. However, in many small domains, direct survey estimators are unreliable or even unavailable due to insufficient sample sizes, particularly when categorical variables contain several response levels. This limitation is well documented in the SAE literature (Fay & Herriot, 1979; Rao & Molina, 2015). This study develops a Bayesian unit-level SAE framework for estimating a three-category health indicator derived from the Greek Health Survey (ELSTAT). The categorical response is modeled through a multinomial logistic mixed-effects model with area-specific random effects, allowing borrowing of strength across domains. Bayesian inference is implemented in R using the brms package, which relies on Stan and Hamiltonian Monte Carlo (HMC) for efficient posterior simulation in complex hierarchical models (Bürkner, 2017). The general idea of borrowing strength through mixed models is consistent with classical SAE approaches such as Battese et al. (1988). A key challenge in official statistics is that only marginal population totals for auxiliary variables are typically available from census data, while the full joint distribution is unknown. To reconstruct consistent population structures within each area, Iterative Proportional Fitting (IPF) is applied. The IPF procedure is implemented in R using the mipfp package (Templ et al., 2017), based on the classical raking methodology of Deming and Stephan (1940). To account for sampling variability and complex survey design, bootstrap resampling and direct estimator computations are implemented in R using the survey package (Lumley, 2004). This ensures that design-based uncertainty is properly incorporated into the simulation framework alongside model-based uncertainty. The proposed framework integrates posterior simulation from the Bayesian model, IPF-based population reconstruction, and stochastic benchmarking procedures to ensure coherence between small area estimates and national aggregates. Benchmarking is applied within each posterior simulation replicate, enabling full propagation of uncertainty across all stages of the procedure. The SAE methodology is grounded in the general framework of small area estimation described in Rao and Molina (2015). The contribution of this work is threefold. First, it extends Bayesian unit-level SAE methodology to multinomial outcomes with multiple response categories. Second, it provides a fully reproducible simulation-based framework implemented entirely in open-source R software, combining brms, Stan, IPF algorithms, posterior predictive simulation, and the survey package. Third, it demonstrates the applicability of modern SAE methods for official statistics production in situations where direct estimators fail due to sparse data structures and unstable domain-level estimates. The methodology is illustrated using data from the Greek Health Survey, highlighting its relevance for producing coherent and reliable official statistics at fine regional levels. Results are presented through spatial maps showing the geographical distribution of the estimated health categories across small areas, providing an intuitive visualization of regional disparities. Spatial visualization is implemented in R using the sf package for spatial data handling and the tmap package for thematic mapping (Pebesma, 2018; Tennekes, 2018). In addition, correspondence analysis is applied to explore associations between areas and health categories using the FactoMineR and factoextra packages, enabling dimensionality reduction and clear graphical representation of categorical relationships (Lê et al., 2008; Kassambara & Mundt, 2020). Keywords: Small Area Estimation; Bayesian statistics; Multinomial models; Official statistics; R software; brms; Stan; mipfp; survey; Iterative Proportional Fitting; Health surveys; Spatial analysis; sf; tmap; Correspondence analysis; FactoMineR. Selected References Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83(401), 28–36. Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1), 1–28. Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics, 11(4), 427–444. Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. Journal of the American Statistical Association, 74(366), 269–277. Kassambara, A., & Mundt, F. (2020). factoextra: Extract and visualize the results of multivariate data analyses. R package version 1.0.7. Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. Journal of Statistical Software, 25(1), 1–18. Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software, 9(1), 1–19. Pebesma, E. (2018). Simple Features for R: Standardized support for spatial vector data. The R Journal, 10(1), 439–446. Rao, J. N. K., & Molina, I. (2015). Small Area Estimation (2nd ed.). Wiley. Templ, M., Meindl, B., & Kowarik, A. (2017). mipfp: Multidimensional Iterative Proportional Fitting and Contingency Table Updating. R package. Tennekes, M. (2018). tmap: Thematic maps in R. Journal of Statistical Software, 84(6), 1–39.

Behind-the-Scenes Complexity, Clear Insights Up Front: Longitudinal Minimum Income Benefit Analysis in a French Local Authority, from Database to Simple Indicators

Elie Chosson[1]

1 - Conseil Départemental de l’isère (France)

We wish to present the longitudinal statistical work carried out within the Isère Departmental Council (a french local authority), using the example of tracking RSA (« Revenu de Solidarité Active ») benefit trajectories (the same method is now applied to other social policies, such as child welfare (ASE) or support for autonomy). The objective is to show that by using R, a local authority can acquire complex information, otherwise scarcely available, but nevertheless decisive in the management of social policies. Thus, this experience has shown that, with regard to RSA, the usual statistics calculated on “stocks” underestimate the dynamism of beneficiaries. To this end, we propose to present the data source (1), then the R workflow, its evolution over time, and the challenges it addresses (2), and finally to illustrate the uses of the database to meet information needs within the local authority (3). (1) The raw data are transmitted monthly to the Department by the local social security agency (the « Caisse d’Allocations Familiales »), in xml format. They are very comprehensive for describing households, but as they are not consolidated, they may contain errors, in particular short “false exits” that need to be corrected. Today, the database, stored as a grouping parquet, covers 113 months from January 2017 for more than 87000 beneficiaries. (2) Initially, processing was made possible by the """"""TraMineR""""""" library, which is very effective for quickly producing longitudinal information. The database was stored as a simple list of monthly data.frames in “RData” format. We will show how a set of factors (GDPR compliance, increase in data volume, internal upskilling) have led to improvements in database management and statistical processing. Today, the R code responsible for updating and utilizing the database—organized within functions in an internal package—relies on a thoughtfully balanced combination of “arrow”, “data.table”, and “tidyverse”. (3) Thus, the database is ready to be used in various ways within the institution: producing indicators for regular deliveries by the statistics department, carrying out comprehensive presentations of RSA benefit trajectories (seniority and outcomes of beneficiaries, typology of trajectories by profile), and finally evaluating the impact of experimental support schemes (modelling RSA exit rates compared between “test” and “control” groups in a non-randomised framework). We will be able to illustrate these different uses. Finally, throughout this complex process, the main difficulty is to simplify longitudinal information and make it easily usable by professionals.

Beyond code: supporting teams through the transition to Open Source

Raphael Simon[1], Vincent Biot[1], Nathalie Rigollot[1], Meriem Said[1]

1 - ATIH (France)

More than a third of ATIH's (Technical Agency for Information on Hospitalization) staff are statistician-engineers, and the agency has historically relied on proprietary software to process the large, complex datasets generated by the PMSI, France's national hospital discharge information system. Since 2023, the agency has undertaken a strategic transition toward open-source solutions, in line with its Strategic Information Systems Plan (SDSI) and following the lead of other public administrations such as INSEE (the French National Institute of Statistics and Economic Studies). The motivations are familiar but consequential: digital sovereignty, reduced licensing costs, tool diversification, and alignment with current data science standards (R, Python, SQL). This transition is not simply a matter of replacing one piece of software with another: it reshapes working practices, professional identities, and the organization of work itself. The central challenge is human, not technical. ATIH's statisticians are domain experts, so the goal is to build on their expertise rather than undermine it. The agency adopted a participatory approach, involving experienced staff early on in defining the agency's R coding standards and in building new technical reference frameworks. Upskilling rests on a comprehensive set of measures: foundational training courses, centralized documentation, internal communities of practice, peer knowledge-sharing sessions, themed workshops, code audits, and four hackathons bringing together the entire statistician workforce around cross-cutting challenges. The change also extends to infrastructure: the systematic introduction of version control, new collaborative practices, and the adoption of code forges and CI/CD pipelines. This segment, often underestimated, requires dedicated support. Finally, the migration must contend with a legacy of statistical programs accumulated over more than twenty-five years. Using the `datadiff` package and an AI-assisted code translation methodology developed in collaboration with ThinkR and Theodo allows the team to make steady progress on this large volume of programs while confirming the stability of their outputs. In conclusion, a successful open-source transition depends far more on people than on technology: support, recognition, structure, and modernization are the keys to a successful migration in a public administration.

Beyond parametrised reports: Adventures automating country notes at the OECD

Maria Paula Caldas Rivera[1], Rodolfo Ilizaliturri López[1]

1 - Organisation de Coopération et de Développement Economiques = Organisation for Economic Co-operation and Development (France)

Each year, the Education and Skills Directorate of the OECD prepares dozens of country notes to accompany its largest statistical publications, including Education at a Glance, the Teaching and Learning International Survey, and the Survey of Adult Skills. Through the years, data analysts across teams have come up with increasingly better ways to streamline this process using RMarkdown, quarto and officer. However, automation is only a small part of the picture. In our presentation, we will speak about some of the less visible constraints faced by large statistical agencies (e.g. complying with existing corporate standards and tooling, fragmented pipelines, incorporating feedback from many stakeholders) and how our data analysts attempt to tackle them. From our favourite {officer} hacks, to our creative uses of Git and the delights of creating Lua filters with Claude, we will share how our toolkit has been evolving and how we hope it continues to evolve in the next years.

Beyond Use Cases: Modular R Packages for Automated Coding and Data Extraction

Nina Niederhametner[1], Eileen Vattheuer[1], Marlene Steiner[1]

1 - Statistics Austria (Austria)

Official statistics increasingly rely on artificial intelligence to automate labor-intensive production processes. While many successful applications exist, solutions are often developed independently for individual use cases, resulting in duplicated effort and limited reusability. To address this challenge, Statistics Austria has developed two modular platforms, implemented as R packages, that provide reusable AI infrastructure for statistical production. The first platform supports automated coding of unstructured text. It supports tasks such as the classification of open survey responses, economic activity coding, and privacy screening of free-text fields. The platform combines large language models, fine-tuned machine learning models, confidence scoring, and human-in-the-loop workflows within a common framework. The second platform addresses the extraction of structured information from heterogeneous PDF documents, including financial statements, utility bills, and administrative records. It integrates OCR, rule-based extraction, table extraction, and LLM-based information extraction. Both platforms are implemented as modular R packages that allow project-specific pipelines to be configured with minimal development effort and deployed, for example, as REST APIs. By shifting the focus from individual use cases to reusable AI infrastructure, the platforms aim to offer implementation through configuration rather than extensive redevelopment. This talk presents the design principles, architecture, and implementation of both R packages and discusses experiences gained from applying them across multiple statistical domains. The work illustrates how reusable R-based infrastructure can facilitate a scalable adoption of AI methods in official statistics while reducing development effort and promoting methodological consistency.

Bridging R and Python for AI-Driven NLP

Georgios Lykos[1], Spyridon Dimas[1], Georgia Panagopoulou[1], Michail Vikentios[1]

1 - Hellenic Statistical Authority (ELSTAT) (Greece)

The classification of free-text descriptions from administrative sources into standardized statistical codes is a fundamental, yet resource-intensive task in official statistics. Whether applied to health diagnostics, economic activities, or occupational data, integrating state-of-the-art Natural Language Processing into traditional R-centric statistical production pipelines presents significant challenges. We will present a hybrid, production-ready pipeline that successfully combines the advanced Deep Learning capabilities of Python with the robust interface and orchestration of R. Utilizing the reticulate package, we developed a system that fine-tunes and deploys multilingual BERT-based Large Language Models directly from the RStudio environment. The model processes unformatted text from administrative registers, assigning confidence scores and extracting the top-five most probable statistical classes for each entry. To ensure the strict data quality and transparency required in official statistics, we implemented a """"""Human-in-the-Loop""""""" architecture using R Shiny. This interactive dashboard allows human reviewers to dynamically filter predictions by confidence level, visually evaluate the model’s top suggestions, and manually override or input custom codes when the algorithm’s certainty is low. This approach demonstrates how statistical offices can leverage the R ecosystem not only to orchestrate advanced Python-based AI models but also to build intuitive review tools that maintain necessary human oversight. The presentation will cover the architectural design of the pipeline, the handling of multilingual administrative data, and the practical benefits of the Shiny application in drastically reducing manual coding efforts while preserving high statistical accuracy.

Building an internal package ecosystem for data processing

Viktoria Kis[1], Erika Lee[1], Simon Normandeau[2], Sara Otavo Peña[1]

1 - Directorate for Education and Skills (France), 2 - Directorate for Education and Skills (France)

At the OECD, our team runs several annual data collections of administrative education data from around 40 countries. When each data collection has its own pipeline, fixes and improvements rarely travel between them, and as pipelines drift, maintaining them becomes more difficult. To answer this challenge, our team created a small ecosystem of R packages, with four core packages handling a different step of a common processing pipeline: reading raw submissions, validating data, deriving indicators, and uploading to an SDMX database. The resulting system is more flexible (new data collections can be easily added) and a fix in one place is a fix everywhere. Our presentation will cover the questions we faced through this process and our responses to the ongoing challenge of creating a long-lasting codebase. What functionalities to include in a single package and what to split? What are the implications for maintainability and governance? What skills does our team need to maintain this new system?

BuSuCo: A package for business survey coordination

Johannes Straubinger[1]

1 - Statistics Austria (Austria)

The purpose of the R package BuSuCo is to allow the user to apply different business survey coordination systems. Business survey coordination is an important task for NSIs that wish to control the overlap of business samples over time and between different surveys. A number of different coordination systems have been developed explicitly for business surveys by various European NSIs. A detailed overview and analysis of these coordination systems can be found in Straubinger (2025): Methods to Coordinate Survey Samples in the Context of German Business Statistics (PhD thesis). The BuSuCo package provides functions to implement each one of the coordination systems discussed in Straubinger (2025) together with a function for creating a customised data set for testing purposes as well as some auxiliary functions. A beta version of the package is already available at https://gitlab.rlp.net/wisostat/busuco. A final version ready for publication at CRAN is planned in the near future.

Calibration of Survey Weights in R: A Comparative Overview of Available Packages

Marcin Szymkowiak[1], Kamil Wilak[1]

1 - Poznań University of Economics and Business; Statistics Poland (Poland)

Calibration is one of the most widely used techniques for adjusting survey weights in official statistics. By forcing the weighted sample totals of selected auxiliary variables to match known population totals, calibration reduces non-response and coverage bias and improves the precision of survey estimates, while ensuring consistency between survey-based and register-based figures published by national statistical institutes. R offers a number of packages implementing calibration, differing in their underlying algorithms, flexibility, and intended use. The aim of this presentation is to compare the main R packages dedicated to calibration of survey weights, for instance: survey, which embeds calibration functions (calibrate, rake, postStratify) within its survey-design framework; sampling, providing the lower-level calib function with several distance functions (linear, raking, logit, truncated); ReGenesees, offering a more elaborate, production-oriented calibration environment with explicit handling of calibration domains, bounded weights and convergence diagnostics, etc. and packages addressing calibration in non-probability and mixed-source surveys, such as NonProbEst and nonprobsvy. The packages will be compared on the basis of a simulation study covering several calibration scenarios (varying numbers and types of auxiliary variables, calibration domains, and levels of misalignment between sample and population totals). The comparison will focus on functionality (available distance functions, support for bounded g-weights, complex sample designs, and population-totals specification), computational performance (execution time and scalability with the number of auxiliary variables and sample size), numerical stability and convergence behaviour, and overall ease of use in a production setting at a National Statistical Office. Resulting weight distributions and the precision of calibrated estimates will also be assessed across packages. The presentation is intended for practitioners in official statistics who are evaluating which R tool best fits their weighting workflow, as well as for R users interested in a practical, evidence-based comparison of calibration software.

ckm: an easy application of cell-key method to protect tabular data

Nadège Ferrer-Pradines[1], Julien Desclodure[1], Dimitri Lemaire[1]

1 - Institut national de la statistique et des études économiques (INSEE) (France)

The Cell Key Method (CKM) is a statistical technique used to protect the confidentiality of tabular data by perturbating all cells in it. The ckm package provides tools to apply the CKM in R, enabling users to select the best set of parameters and to generate perturbed counting tables from microdata. After generating the individual keys, the user can directly build the perturbed table. A single function returns a list containing the perturbed table, the transition matrix, the risk and utility measures and the original table. Additionnal functions allow to explore various sets of parameters.

codelist: working more efficiently with code lists in R

Jan van Der Laan[1]

1 - Statistics Netherlands / CBS (Netherlands)

Classifications with corresponding code lists are very common in Official Statistics. A code list consists of a set of permissible codes with corresponding meta data such as labels, parent-child relations and descriptions. Examples are NACE for classifying industrial sectors and regional classifications. Although code lists are common in statistics, the functionality in R for working with these is very limited. Base R only has the factor which is basically a character vector with a restraint on the values that can occur in the vector. It does not allow for a mapping between arbitrary codes and labels. There is also the labelled object from the haven package. This object is mainly intended to be used with the read functions in the haven package to preserve categories from applications such a SPSS and only allows for a mapping between codes and labels. In this paper I will introduce the codelist package. This package has the following features: define code lists with multilingual labels and descriptions; codes can be in simple hierarchies (each code can have one parent code); define codes that can be interpreted as missing values; is supplies methods to assist in working with codes such as casting to higher levels in a hierarchy and converting between codes and labels; and validity checking when comparing and assigning values.

Creating population pyramids in less than 100 lines of code

Michael Neutze[1]

1 - Federal Statistical Office of Germany (Germany)

Official statistics have many uses for population pyramids but there's a general lack of suitable software to create them. Even ggplot examples are astonishingly rare, lack in fidelity or are too complex code-wise. To fill this gap we will come up with easy to understand examples that result in production ready graphics, even for print layouts like scientific journals and comparable publications. We will start with basic diagrams and add comparisons e.g. over time. Different approaches will be discussed and, in the end, users should be able to adapt code examples to best fit their needs and refine afterwards. The tutorial is aimed at beginner to intermediate level. A basic understanding of ggplot2 is preferable. On the other hand if you are already a frequent user of gtable and can write grobs from memory I can probably not teach you much. Your computer should have the following packages installed: readxl, tidyverse, scales, ggtext, grid, gridtext, gridExtra, gtable and patchwork. Example data and code will be provided via https://github.com/wahlatlas/uRos2026 Attendees are welcome to bring their own data, no matter which topic (demography, migration, traffic accidents etc.) but take the time to find data available in one year age-bands to make the most of the data visualisation. However, we will also develop solutions for pyramids made from age-group data.

Developing an R Package for Sample Weight Estimation

Guillaume Osier[1], Claude Lamboray[2]

1 - Institut national de la statistique et des études économiques du Grand-Duché de Luxembourg (STATEC) (Luxembourg), 2 - STATEC (Centre Administratif Pierre Werner B.P. 304 L-2013 Luxembourg Luxembourg)

As part of its strategy, Luxembourg's National Institute of Statistics and Economic Studies (STATEC) is investing in making its statistical production more streamlined, efficient and standardised. To support this objective, the institute has already developed in-house R packages addressing core statistical processes, such as imputation (STATECimputation) and variance estimation (STATECvariance). Building on this approach, the next step is the development of an R package dedicated to the calculation of sample weights for survey data. The package aims to provide survey managers with practical guidance and a harmonised framework for producing robust sample weights across different surveys. This presentation will briefly outline the overall architecture of the package and its main features.

DuckDB beyond R: demystifying the analytical database in your toolbox

Ines Hiverlet[1]

1 - Institut national de la statistique et des études économiques (INSEE) (France)

DuckDB has rapidly become a cornerstone of modern data analysis workflows, offering a lightweight, high-performance analytical database that runs directly on a laptop without the complexity of traditional database administration. While many R users encounter DuckDB through its seamless R interface, understanding what lies beneath the wrapper can unlock new ways of thinking about data processing and performance. This talk introduces DuckDB from two complementary perspectives. First, we will explore how the DuckDB R package integrates with familiar tools such as dplyr and Arrow, enabling analysts to work with datasets that exceed memory limits while maintaining an idiomatic R workflow. Second, we will step outside R and examine DuckDB through its command-line interface, using the CLI to reveal the database concepts, execution model, and query engine that power the R experience. By demystifying DuckDB as both an R package and a standalone database system, attendees will gain a deeper understanding of how modern analytical engines work, when DuckDB is the right tool for a problem, and how to move confidently between interactive analysis, reproducible pipelines, and SQL-based exploration.

Estimate distributions with fat and skewed tails with the FatTailsR package

Patrice Kiener[1]

1 - InModelia (5 rue Malebranche, 75005 Paris France)

The FatTailsR package generalizes the logistic distribution to fat and skewed tails thanks to a third and a fourth parameter that both have a physical meaning. I will explain the design of this new distribution and two methods for the parameter estimation, one ultra-fast non-parametric method and one parametric method. Both methods allow to estimate simultaneously the central part of the distribution and the left and right tails, departing from the classical Hill estimator. The result is much more accurate. As the quantile function has a closed form, estimating the quantiles at any level is very easy. Examples will be taken in stock markets since log-returns of assets usually exhibit skewed and fat tails. For instance S&P 500 has a tail index of 3.2 plus some skewness.

Event-sequence analysis of fertility histories with TraMineR, INDSCAL and DISQUAL

Elisabeth Morand[1]

1 - Institut national d’études démographiques (France)

Sequence analysis is a standard tool for life-course social science research, using TraMineR package implementation. Using it on panel survey which permit to have retrospective and prospective sequence aim to compare group and how the geometrical structure is shared from a wave to another. The data are the French Generations and Gender Survey (ERFI): the same respondents were interviewed three times, in 2005, 2008 and 2011. Each fertility history is a sequence of dated events such as unions, first birth, later births. From one wave to the next, respondents do not rewrite their past; they just add a few events, often quite literally a baby. We distinguished the first wave into two groups, those who answered all the 3 times from the others, so that attrition can be looked at rather than ignored. For respondents of all waves, dissimilarities between event sequences were computed in TraMineR (optimal matching of event sequences, OME) and clustered them into a trajectory typology. The three dissimilarity matrices describe the same individuals, and INDSCAL (the smacof package) returns one shared space of fertility patterns and one set of weights per wave, showing how that space stretches as births accumulate. The sequences only grow over time, so the three configurations resemble one another by construction; that resemblance is the thing I am trying to measure and the difference with the space using more individual sequences but only on one wave. I summarise it with two old and well-understood numbers: the RV coefficient for how close the configurations are overall, and the adjusted Rand index for how much the clusterings agree. The study was also conducted on the new GGP survey realised in 2024, to see if the 2005 sequence space is more or less the same as the 2024 one. Finally, Saporta’s discriminant analysis on qualitative variables (DISQUAL) is used to see how far age group, sex and occupation account for which trajectory someone follows. The study is conducted on France but can be compared to another country.

From code to print-ready publication: Automating layout at the French Ministry of Justice

Adam Marsal[1], Fanny Allard[1]

1 - Ministère de la Justice [Paris, France] (France)

The production of formatted statistical publications is a major challenge for official statistical services: it often requires substantial resources, particularly through the use of external providers for layout work, and limits the reproducibility of outputs. To address these constraints, and as part of the modernisation of its production processes, the Statistical Service of the French Ministry of Justice (SSER) has developed an in-house tool based on the R ecosystem. This tool takes the form of an R package containing reproducible publication templates, along with helper functions for chart formatting and page layout. It produces paginated, print-ready outputs without the need for any external layout software. The publications generated automatically comply with the SSER's visual identity guidelines: a two-column layout, mandatory typography, institutional headers and footers, and a defined colour palette. The result is a professional PDF, laid out without any manual intervention and ready for publication on the ministry's website. A first template covering the Infos Rapides Justice collection has been deployed, enabling the production of around fifteen publications. Using this tool reduces layout costs and turnaround times by removing the systematic reliance on external providers, and greatly improves the reproducibility of publications.

From SAS to R, one small step for data production, one giant leap for official statistics

Pierre Lamarche[1,2], Clément Rousset[3]

1 - Centre de Recherche en Économie et Statistique (5, Avenue Henry Le Chatelier91120 Palaiseau France), 2 - Institut national de la statistique et des études économiques (INSEE) (France), 3 - INSEE (France)

Exiting SAS has turned to be a tremendous effort for Insee, with the particularly unsatisfying purpose to maintain the exact same level of quality in the production process, without the obligation of paying the expensive licence costs. It has stood for a significant investment in terms of training and time dedicated to conversion of legacy code, which sums up to almost 100 FTE over the three years the project has last. Yet this successful project is paving the way for the modernisation of the statistical production, as it comes with the deployment of modern datascience platforms, the promotion of good practices regarding coding, and in a general perspective, a shift towards datascience practices and an increased ability in adopting new techniques for the production of official statistics. This change has also an impact on the frontier between IT and statistics, as statisticians gain in their ability in deploying automated and secured processes. The evolution we foresee for Insee has a whole has been made possible by an ambitious investment on IT infrastructures and software, and encompasses the open-source development of in-house solutions such as Onyxia, which turns to be relevant in the context of official statistics, but not only.

Harvesting Data as Crops Are Harvested : An Automated R Pipeline for Timely Yield Estimates

Hanan Touzani[1]

1 - Ministère de l'agriculture, de l'agroalimentaire et de la souveraineté alimentaire (France)

This project focuses on a survey about crop yields on arable land. The main goal of this survey is to provide data in order to help analysts to make short-term agricultural situation reports. As the survey covers several different crops and because of weather and geographic differences, data collection happens at different times depending on the region and the crop. Indeed, some regions harvest and report their data much earlier than others, and the need for quick forecasts varies by crop and area. To manage this complex need, we developed an automatic R program using packages such as survey or ReGenesees. As soon as new data are collected by the data collection team, the program automatically updates and calculates the estimates. It provides statistics such as average yields by crop and by area, confidence intervals and precision levels. Moreover, the program automatically checks if the data currently available for a specific crop in any area is reliable enough to be used, or if the sample is too small. This allows each region to report its figures as soon as they are ready.

Integrating Time Series Processing into an Automated Production Workflow in R

Suzanne Heidsieck[1]

1 - Service statistique ministériel de la sécurité intérieure (France)

Since 2023, SSMSI has redesigned the production process for monthly delinquency indicators to maximize automation and facilitate the transition to R. The workflow is managed through a Git repository and is organized around a main script located at the root of the project, together with several secondary scripts dedicated to specific tasks and output generation. The process relies on a current-month directory and archived directories from previous months, ensuring a stable structure throughout the production cycle. The program is structured into four R modules. The first module initializes the paths, variables, and functions required to run the workflow. Object names containing date information are automatically generated from the current date. The second module computes all raw monthly series from three databases covering infractions, victims, and suspects. Fourteen of the fifteen indicators exhibit seasonal patterns and therefore require seasonal adjustment in order to remove seasonal and trading-day effects that may impede short-term movements, long-term trends, and cyclical behavior. The third module relies on a cruncher that updates a workspace which refreshes the data and models' parameters, and generates the corresponding outputs. Through the JDemetra+ interface, the producer reviews and validates any proposed modifications to the pre-specified models. Finally the fourth module generates all required outputs which are mostly outputs consisting of Excel tables containing the information needed to prepare the monthly publication, including the latest indicator values and their variations. The module also produces Quarto reports, particularly for daily series covering the previous three months, allowing the producer to detect potential data-quality issues that may not be visible in the aggregated monthly series. The monthly production process is complemented by an annual review campaign aimed at revising the parameters used in the decomposition models and identifying the most appropriate specifications for each series, including parameter re-identification and re-estimation, predictors' set selection, as well as outliers detection.

KOMA – An R package for Bayesian estimation of simultaneous equation models

Samad Sarferaz[1], Merlin Scherer[1]

1 - ETH Zurich (Switzerland)

Central banks and forecasting institutions rely on simultaneous equation models for their interpretability and coherent, theory-driven forecasts. KOMA (KOF Macroeconomic Analysis) is an R package for Bayesian estimation of such models using Metropolis-within-Gibbs sampling. It provides text-based model specification, customizable prior distributions, full posterior inference, conditional forecasting, out-of-sample evaluation, and ready-to-use tables and plots.

Large Language Models as Assistants for Time Series Analysis in Official Statistics: Opportunities for Innovation and Capacity Building

Jelena Suzić[1]

1 - Department for Time Series Analysis and Composite Indicators Development (SORS) (Serbia)

Time series analysis is a key component of official statistics, supporting seasonal adjustment, short-term economic analysis, forecasting and evidence-based policymaking. At the same time, National Statistical Institutes face growing demands for innovation, advanced analytical capabilities and effective transfer of methodological knowledge to new generations of statisticians. Recent advances in Large Language Models (LLMs) offer new opportunities to support both statistical production and capacity building. Beyond code generation, LLMs can act as interactive methodological assistants, helping users understand statistical concepts, explore analytical approaches, develop reproducible workflows, and accelerate the implementation of analytical solutions. This paper presents practical experiences with the use of LLMs in time series-related activities within the Statistical Office of the Republic of Serbia. Applications include support for seasonal adjustment using JDemetra+ and R, development of ARIMA-based forecasting models, generation of analytical and visualization scripts, preparation of methodological documentation and development of training materials for statisticians. Particular attention is given to the role of LLMs as learning assistants that facilitate onboarding, support knowledge transfer and help staff acquire programming and analytical skills in the field of time series analysis. The paper is framed within the broader objectives of the Decision-Making Support System at the National Level (DMSS NL), currently developing under the IPA 2022 programme – EU4SORS “Development of a Modern Statistical System”. As a strategic initiative focused on modernizing official statistics and promoting innovative analytical solutions, DMSS NL provides a suitable environment for exploring emerging technologies and new approaches that can strengthen analytical capacities and support methodological innovation. Based on practical examples, the paper evaluates the opportunities and limitations of LLM-assisted work, including productivity gains, improved accessibility of advanced methodologies, support for innovation, and enhanced learning processes. Challenges related to statistical validity, reproducibility, transparency, governance, and the need for expert validation of AI-generated outputs are also discussed. The findings suggest that LLMs can serve as valuable assistants in strengthening analytical capacities, accelerating methodological development, and supporting the modernization of time series production processes, while maintaining the central role of expert judgement in official statistics.

Linking Administrative Data Without Unique Identifiers: A randomForest Approach in R

Christopher Nölting[1], Murad Hajiyev[1]

1 - Federal Statistical Office of Germany (Germany)

The presentation describes an R-based approach to link administrative data with the statistical business register when unique identifiers are missing. Central to the solution is the randomForest package, whose ensemble learning and random feature selection mechanisms enable robust predictions. The model’s ability to evaluate feature importance played a key role in refining the matching process. String similarity was addressed using the stringdist package. The implementation achieved a remarkable 98.7 % top-1 accuracy, significantly outperforming traditional scoring methods. The presentation covers the R implementation, including parameter optimization and techniques for handling imbalanced data.

Making gustave production-ready: multistage designs, wrapper inspection and estimation diagnostics.

Thomas Delclite[1]

1 - Statbel, l'office belge de statistique (Belgium)

Variance estimation is an important part of survey quality, but it is hard to do well when the sampling design is more complex than simple random sampling. For many years, several statistical offices have used the SAS macro Poulpe (Caron, 1998) for this task. Its R successor is the gustave package (Chevalier & Larbi, INSEE): transparent, reusable and reproducible, but not always easy to learn. This talk is an experience report from an ongoing project at Statbel, the Belgian statistical office, where we are moving the variance estimation of our household surveys from Poulpe to R. I am not the author of gustave. My aim is to share what we learned from real production work, and to present the tools we had to build along the way. The starting point was a gap. The qvar() function makes variance estimation easy in the most common single-stage case, but our surveys are multistage. I present qvar_ms(), a declarative interface for an arbitrary number of sampling stages, with non-response correction at any subset of stages, calibration at any stage, and stratification crossed with the parent units. It implements the same Rao (1975) decomposition as Poulpe, and every case is validated both against a hand-built wrapper and against Poulpe reference output. Moving to production raised a second question: how do you trust a wrapper that someone else defined? A gustave wrapper is a function that carries a whole survey design frozen in its environment. This makes it convenient to use, but opaque to read. I present two answers. First, estimation-time diagnostics: variance components with the exact effect of calibration, design effect and its Kish decomposition, degrees of freedom, non-response and domain checks. Second, inspect_wrapper(), which rebuilds the pipeline that will actually run from the technical data of the wrapper itself, and reports it as a self-contained document for audit and quality reporting.

Making Singular Spectrum Analysis More Accessible: A Step-by-Step Framework and R Implementation

Eulalie Delaune[1]

1 - Institut national de la statistique et des études économiques (INSEE) (France)

Singular Spectrum Analysis (SSA) is a family of non-parametric methods used for signal processing that emerged in the 1980s. It provides a framework for decomposing a time series into interpretable components, such as trend, seasonal and noise These methods perform well under just a few hypotheses, and do not require that the series be stationary. However, they offer a lot of flexibility, making them difficult to implement in practice, as their performance strongly depends on several user-defined parameters and on the grouping of elementary components. This paper presents a detailed and pedagogical overview of Basic SSA applied to time series analysis, with several of its extensions. Particular attention is given to its practical implementation, including specific visualizations helpful for parameter selection. Several extensions are discussed: Toeplitz SSA, Sequential SSA, automatic parameter selection procedures, and the integration of Independent Component Analysis (ICA) to improve component separability. Beyond reviewing existing methods, this paper aims to provide a practical framework for understanding and applying SSA. It also describes a forthcoming R package designed to make SSA more transparent by exposing each step of the analysis and facilitating the visualization of grouping strategies, making parameter selection easier. Several extensions are also implemented. The package is intended to be complementary to existing SSA packages, focusing on its interpretability and effectiveness in real-world applications.

Measuring Spatial Inequalities in University Accessibility Using R and GTFS Data: Evidence from Rome

Armando D'aniello[1], Gianluigi Salvucci[1]

1 - Istituto Nazionale di Statistica (Italy)

Access to higher education is not only a matter of institutional supply, but also of effective reachability. In metropolitan contexts, the spatial distribution of public transport services can either mitigate or amplify territorial inequalities, acting as a structural determinant of educational opportunity. This study presents an analysis of transit-based accessibility to universities in Rome, Italy, integrating open data sources within a fully reproducible R workflow. The analysis combines three main data sources: GTFS (General Transit Feed Specification) data from Rome's public transport operator (June 2025), the 2021 Istat Census Tract boundaries used as elementary spatial units, and OpenStreetMap road and pedestrian network data. Additionally, a set of socioeconomic vulnerability indicators, produced by Istat for sub-municipal areas in the context of a parliamentary inquiry on urban peripheries, is incorporated to contextualize accessibility outcomes within a broader spatial justice framework. The methodological workflow is implemented in R using a set of specialized packages. The r5r package provides an interface to the R5 routing engine, enabling multimodal travel time computation (WALK + TRANSIT) from census tract centroids to university locations, with departure time set to a weekday morning peak hour (08:00). A maximum trip duration of 61 minutes and a maximum walking time of 20 minutes are imposed. The sf package handles spatial data operations, rmapshaper performs geometry simplification, and tmap is used for cartographic visualization. The Osmosis command-line tool, called within the pipeline, is used for spatially cropping the OSM .pbf extract to the buffered study area. Accessibility is operationalized as a binary indicator: census tracts from which at least one university campus can be reached within the defined time threshold are classified as accessible. Results show that 56.6% of Rome's census tracts (approximately 61% of the resident population) fall within this threshold, while the remaining areas, concentrated in peripheral zones, remain underserved. Cross-tabulation with the Istat vulnerability index reveals a statistically significant association (χ^2 test, $p < 0.001$) between inaccessibility and social vulnerability, with a Pearson φ coefficient of -0.31. Among inaccessible tracts, 37% belong to vulnerable sub-municipal areas, compared to only 11% among accessible ones. Over 334,000 residents live in tracts that are simultaneously inaccessible and socially vulnerable. These findings confirm accessibility as an effective proxy for territorial equity and demonstrate the value of integrating GTFS-based routing, official statistical geographies, and social indicators within a transparent, open-source R environment for evidence-based urban and transport planning.

Mental health and social participation: A data-driven analysis using R

Camelia Maria Dan[1], Cristina Rodica Boboc[1]

1 - The Bucharest University of Economic Studies / Academia de Studii Economice din Bucureşti (Romania)

Promoting mental well-being has become a major priority in educational and social contexts, as young people and adults face increasingly intense emotional, social, and professional pressures. This study investigates the relationship between perceived social support, frequency of social interactions, and mental well-being, with a focus on how the quality of interpersonal connections contributes to psychological resilience. The data source used is the European Working Conditions Survey (EWCS 2024), a large-scale dataset collected by Eurofound that covers a wide range of aspects of European citizens' professional and social lives. The entire workflow, from data import and cleaning to indicator development and statistical modeling, is implemented in R, utilizing packages from the tidyverse ecosystem for data manipulation, as well as specialized packages for analyzing complex surveys. The dependent variable is a measure of mental well-being constructed using the WHO-5 Well-Being Index methodology, applied to the relevant items from the EWCS 2024. The main independent variables include perceived social support (the availability of a trusted person, indicators of interpersonal trust), the frequency of social interactions (meetings with friends and relatives, participation in community activities), and the quality of relationships (satisfaction with personal relationships, feelings of loneliness). Control variables include age, gender, educational attainment, employment status, and household composition. The methodology includes descriptive statistics broken down by demographic groups, a correlation matrix between the dimensions of social support and mental well-being, linear regression models for estimating the relative impact of each factor, and an analysis of heterogeneity across countries. Preliminary results indicate that perceived social support is a strong predictor of psychological well-being, with significant variations across countries and demographic groups. The study highlights the importance of leveraging data from official European statistics to analyze the social dimensions of mental health and demonstrates how R can support the entire analytical chain, from raw data to interpretable results. The findings have direct implications for public policy, suggesting the need for institutional interventions that strengthen social connectedness and support networks, particularly in educational settings.

Modernizing Global Survey Data Processing: Reproducible R Pipelines at the IEA

Winfried Moser-Nussgruber[1], Kamil Kowolik[1]

1 - International Association for the Evaluation of Educational Achievement (IEA) (Germany)

Large international educational studies such as TIMSS and ICCS collect and process highly complex assessment and questionnaire data from hundreds of thousands of respondents worldwide. Historically, the International Association for the Evaluation of Educational Achievement (IEA) managed these massive datasets through long-established, relational database infrastructures. As online data collection and study complexity increase, the IEA has modernized its approach to meet new agile demands. This presentation details our successful evolution toward script-driven, reproducible workflows for upcoming studies, including ICCS 2027 and TIMSS 2027. At the center is dpeR, a stable, domain-specific R-package tailored exclusively to the IEA's complex survey structures. By transitioning from intermediate database states to in-memory processing, we achieved substantial performance gains—accelerating complex data transformations to mere minutes while improving flexibility. This shift significantly reduced maintenance debt by streamlining the codebase: pure production code was reduced from 12,000 to 4,000 lines, and study-specific scripting decreased from 15,000 to 6,500 lines. Furthermore, replacing a proprietary, IEA-maintained point-and-click interface (28,000 lines) with RStudio, an externally maintained open-source IDE (0 lines) eliminated another major source of institutional overhead. For international stakeholders, data integrity is paramount. To ensure this, dpeR acts as a secure single source of truth. In contrast to the legacy infrastructure, which operated without automated tests, dpeR is actively safeguarded by 13,800 lines of rigorous unit testing code. This formally verified foundation enables data managers to independently control their pipelines and seamlessly integrate advanced statistical procedures. Ultimately, this project demonstrates that modernization is not solely a technical challenge, but a cultural one. Successfully transitioning to code-driven workflows required breaking down operational silos and fostering knowledge sharing between methodological, operational, and technical experts. Although this initially transferred substantial technical responsibilities directly to data managers without team expansion, the resulting reusable pipelines eliminated historical overtime and guarantee drastically reduced configuration efforts for subsequent survey cycles. While the dpeR framework remains proprietary IEA intellectual property, this case study serves as a blueprint demonstrating how large-scale institutions can collaboratively build rapid, secure, and reproducible survey architectures by combining rigorous software engineering standards with the analytical power of R.

National Accountant: A Reproducible R–DuckDB Framework for Balancing Supply and Use Tables with Hard and Soft Constraints

Alpay Kocak Necmettin[1]

1 - Hacettepe University = Hacettepe Üniversitesi (Turkey)

National statistical offices increasingly need production-grade, reproducible pipelines to compile and balance Supply and Use Tables (SUTs) and related input-output systems, while also accommodating emerging extensions such as digital economy breakdowns, TiVA-style indicators, and experimental modules aligned with updated international standards. In practice, many SUT workflows still rely on spreadsheets and manual adjustments, limiting transparency, scalability, and repeatability. This presentation introduces “National Accountant”, an R-based Shiny application that operationalizes SUT balancing as a configurable optimization problem with a DuckDB backend. The framework is designed as a generic balancing studio applicable to SUT, IO, Make-Use, and similar matrix systems. It supports a relative adjustment objective at the cell level, hard linear constraints to enforce accounting identities, and soft constraints with reliability weights to represent partially trusted data sources. Flexible lower and upper bounds, including non-negativity, are also supported. To ensure scalability, the system relies on sparse matrix representations and OSQP-based quadratic programming. A key feature is its metadata-driven design. A lightweight YAML specification defines periods, layers, block structures, and matrix dimensions, allowing the same engine to be reused across different datasets and institutional contexts. The application includes interactive tools for constructing SUT linkages without coding, alongside diagnostics that highlight infeasibilities, constraint tensions, and data inconsistencies. Beyond core balancing, the framework integrates optional extension modules when available in the database. These include experimental digital economy reconciliations, TiVA compilation workflows, and a numerically robust proxy for Local Content of Exports (LCX). Importantly, the system remains fully functional even when such extensions are absent, enabling gradual implementation across statistical offices with varying levels of data availability. The presentation demonstrates an end-to-end workflow, from data ingestion and validation to balanced outputs and derived indicators, highlighting reproducibility, modularity, and deployment considerations using R and Shiny. The approach offers a practical pathway for modernizing national accounts production systems.

phsshiny: An R Package for Standardisation and Accessibility in the Publication of Health Data

Ciara Gribben[1], Russell Mccreath[1], Bateman Mcbride[1]

1 - Public Health Scotland (United Kingdom)

Public Health Scotland (PHS) publishes over 250 public dashboards, most related to official statistics. It's vital that these products align to governance and regulatory standards, including accessibility, access management, and statistical release processes. However, despite support through guidance and processes, dashboard quality and accessibility vary across the organisation. This is a significant challenge for an organisation as large as PHS and has real implications for compliance and public trust. The phsshiny R package was developed as part of a centralised review of dashboard development practice. The package contains a suite of functions to produce information-focussed applications in a highly standardised, accessible and reproducible format. Implemented using a Reproducible Analytical Pipeline (RAP) approach, the package handles accessibility, styling, security, metadata, and general organisational governance requirements automatically. This ensures these requirements are consistently met, reduces the duplication of effort to get there, and allows app producers to focus on meaningful analysis rather web development principles, styling, performance, security, etc. The functions within the package create a 'declarative' framework to create an information application. The standard Shiny format is standardised into simple functions that handle layout and components, keeping accessibility and open standards at the core. Charts, as an example, include an option to show and download the underlying data, requiring no additional code or work. Building on the open standards, requirements for metadata and governance, information about the organisation, cookies, and accessibility are all applied with templates automatically. As a proof of concept, an existing management information publication, previously using flexdashboard with over 60 visualisations, was rebuilt using phsshiny in collaboration with key stakeholders. This resulted in an interactive, accessible, and fully compliant application, built from a codebase that is focussed entirely on analytical content. Other projects show a reduction in the codebase size of around 80%, applications can be put together as a prototype in a couple of hours, and all with consistently positive feedback around the usability and impact on working practices. This package is currently being rolled out to other projects, considering additional requirements and iteratively improving based on feedback. The goal is to establish a consistent, accessible, and maintainable publication standard across all PHS outputs. This presentation will share the principles behind phsshiny, sharing a transferrable framework to standardise analytical product development, and the lessons learned along the way.

pickmdl3: An R package implementing the X-12-ARIMA pickmdl procedure within the JDemetra+ R framework

Jens Kristoffer Haug[1], Øyvind Langsrud[1], Ane Seierstad[1], Susie Jentoft[1], Ivar Hamre Jørn[1], Quang Pham Dinh[1]

1 - Statistics Norway (Norway)

Seasonal adjustment is of central importance to the analysis of economic time series in official statistics. Within the European Statistical System, JDemetra+ is the recommended software, available in R through the packages RJDemetra and the newly released rjd3. At Statistics Norway, we have transitioned from US Census Bureaus's X-12-ARIMA to JDemetra+ in recent years. To ensure methodological consistency we have developed the R package pickmdl3 which implements the pickmdl model selection procedure from X-12-ARIMA within the rjd3 framework. The pickdml model selection procedure differs from the automdl procedure in JDemetra+ by restricting the choice of RegARIMA model to an ordered set of five parsimonious models. This restriction may lead to more stable model selection and thus less future revisions to seasonally adjusted data. A limitation of the pickmdl approach, however, is that if none of the listed models fits the data well, the procedure selects an inadequate RegARIMA model, potentially conflicting with the ESS's guidelines on Seasonal adjustment. To address this issue, the pickmdl3 package provides the option to fall back on the automdl procedure when none of the five listed models proves adequate. The pickmdl3 package further includes functionality to support an orderly production process when using the X-12-ARIMA methodology in rjd3. The package includes functions for seasonal adjustment of multiple series based on a data frame with model specifications, as well as a graphical user interface for creating, editing, and saving such data frames. This functionality provides an efficient overview and contributes to a well-organized production process as the data frames with specification can be saved for versioning.

postlink: An R Package for Post-Linkage Data Analysis

Priyanjali Bukke[1], Gauri Kamat[2], Jiahao Cui[2], Roee Gutman[2], Martin Slawski[1]

1 - University of Virginia (United States), 2 - Brown University (United States)

The advent of the digital age has reshaped the processes by which data are collected, shared, and analyzed. An important trend within this landscape is the integration of disparate data sources, often collected independently and stored in isolation. Consequently, methods for data integration are needed to realize the value of all existing data sources in combination. Among these methods, Record Linkage (RL) plays a pivotal role by enabling the connection of individual records across multiple data sets. RL spans a wide array of contemporary applications, including the linkage of survey data, administrative and electronic health records, as well as historical censuses and vital records. Open-source software to facilitate RL when records are linked using quasi-identifiers is widely available. However, tools for analyzing the resulting linked files that account for linkage uncertainty are lacking. In particular, there is lack of software in the secondary analysis setting in which RL and downstream analysis are conducted separately. In this talk, we present postlink, an R package dedicated to analyzing data integrated using error-prone record linkage. A central feature of the package is its extensible, formula-based architecture, offering a suite of post-linkage counterparts to R's standard modeling toolkit. We conclude with a discussion of ongoing methodological extensions and automated maintenance and end-user workflows, alongside opportunities for community-driven software development.

Preprocessing Categorical Variables through Edit-Induced Equivalence Classes in ARTIC

Simona Toti[1], Romina Filippini[1]

1 - Istituto Nazionale di Statistica (Italy)

The identification and treatment of non-sampling errors is a fundamental step in the production of Official Statistics, as it contributes to improving the accuracy and overall quality of statistical estimates. In this context, ARTIC (Automatic Random Error Treatment and Imputation for Categorical Variables), a standardized procedure implemented in R, has been developed for the automatic detection and correction of random errors in categorical data, following the Fellegi-Holt (F-H) framework. The Editing and Imputation (E&I) process consists of three sequential phases: (i) data checking to identify records affected by errors, (ii) error localization to determine the variables responsible for each inconsistency, and (iii) imputation of the identified errors. These phases are implemented using R packages such as *validate*, *validatetools*, *errorlocate*, and *VIM*, together with a set of ad hoc functions. Within the F-H framework, the localization phase aims to identify the smallest set of values that must be modified in order to restore consistency. The *errorlocate* package formulates this task as a Mixed-Integer Programming (MIP) problem and relies on a MIP solver to obtain an optimal solution. The computational time required by the `locate_errors()` function depends on both the number of edit rules and the number of categories involved in those rules. To improve computational efficiency, we propose two preprocessing functions that reduce the number of categories considered during the optimization process. Fellegi and Holt suggested exploiting *edit-induced equivalence classes* to simplify the imputation phase. These classes partition the domain of a variable into groups of values that have identical effects on the satisfaction or violation of all edit rules involving that variable. Building on this idea, we observe that the minimum-change localization problem yields the same solution whether it is formulated on the original variables or on variables recoded according to their edit-induced equivalence classes. Based on this property, we developed a set of R functions that take the original data and edit rules as input and produce a recoded version of both the data and the rules, where variable categories are replaced by their corresponding equivalence classes. The proposed recoding can be applied as a preprocessing step before error localization. Results from a real-world application show that this approach reduces computational burden while preserving the correctness of the localization process.

procR: New Features for End-to-End Table Production in Official Statistics

Younes Saidani[1], Ngoc-Han Tran[1]

1 - Federal Statistical Office of Germany (Destatis) (Germany)

National statistical offices routinely produce cross-tabulations of microdata—descriptive summary tables that present key indicators such as counts, sums, or means, often broken down by demographic, geographic, or socioeconomic dimensions. These tables form the backbone of official publications and data dissemination. The package {procR} being developed at Destatis provides a unified, R-native workflow for producing custom cross-tabulations of microdata – from variable formatting through aggregation to publication-ready output. This talk presents areas of active development since the initial release, most prominently expanded output options and support for replicate weights: 1) Aggregate tables produced by {procR} can now be exported as publication-ready Excel, CSV and HTML tables. Variable hierarchies are visually represented using nested row and column headers. Furthermore, accessible and machine-readable tables are also automatically produced. 2) Survey-weighted estimates in official statistics frequently require variance estimation via replicate weights. {procR} allows efficiently computing point estimates and standard errors using replicate weights without exploding memory footprint. These and further additions like integrated statistical disclosure control move {procR} closer to its goal of covering the complete table production workflow – from raw microdata to dissemination – within the R ecosystem. The package is openly developed at https://gitlab .opencode.de/ysaidani/procr. Keywords: package; tabulation; crosstab; excel; html; statistical disclosure control; sas

Protecting linked tables with non-nested hierarchies: a use case with rtauargus

Nadège Ferrer-Pradines[1], Julien Desclodure[1], Dimitri Lemaire[1]

1 - Institut national de la statistique et des études économiques (INSEE) (France)

The rtauargus package provides an R interface for τ-Argus, a reference tool to apply statistical disclosure control to tabulated (aggregated) data. While τ-Argus is a tool designed to be used via a Java GUI or called using a batch-file (script-file), rtauargus package allows to create inputs (rda, arb, hst and tab files) from data in R format. It generates the sequence of instructions to be executed in batch mode (arb file) and launches a τ-Argus batch in command line, then retrieves the results in R. These different operations can be executed in one go, but also in a modular way. They allow to integrate the tasks performed by τ-Argus in a processing chain written in R. rtauargus contains other additional functionalities, some of them unique to this package. In this presentation, we will focus on the protection of linked tables (tables sharing common cells that must be protected consistently) with tab_multi_manager(), in presence of non-nested hierarchies. We will also introduce a modification on tab_multi_manager() released in version 1.3.4 that increases its efficiency, especially on large lists of linked tables (100 tables and more).

Publishing in the R Journal

Mark van Der Loo[1,2]

1 - Eurostat (Luxembourg), 2 - Leiden Institute of Advanced Computer Science [Leiden] (Netherlands)

Since 2021, the uRos conference has offered the possibility to submit papers for a special section in the R Journal. The conference has a tight connection with the journal, with several members of the organizing committee serving or having served on the Journal's editorial board. Until now the efforts of the uRos conference have resulted in just a single published paper. Drawing on my experiences as editor for the R Journal, this talk aims to help prospective authors preparing successful submissions. I will give a look behind the scenes, explaining how the editorial team is organized, and elaborate on the editorial process. It will be explained how papers are selected for review and on what grounds editorial decisions are based. I will also highlight some common pitfalls that have lead to rejection and how to avoid them.

R Shiny Platform for Detecting, Validating, and Reporting River Overflow Periods

Vahe Movsisyan[1], Madoyan Habet[2], Ina Karapetyan[1]

1 - Datamotus LLC (Armenia), 2 - American University of Armenia (Armenia)

Reliable flood-timing indicators are essential for environmental statistics, early warning communication, and routine hydrometeorological reporting. We present an R Shiny platform that operationalizes the full workflow for river overflow analysis: automatic detection of spring overflow start and end dates, expert validation, hydro-meteorological interpretation, and production of report ready statistical outputs. The detection module uses a transparent, high-level event segmentation approach on daily discharge series. Discharge dynamics are smoothed to reduce short-term noise, seasonal reference behavior is estimated from annual flow structure, and candidate high-flow periods are evaluated to isolate the principal overflow episode for each monitoring station. To strengthen interpretability, the platform integrates nearby meteorological station data, including temperature and precipitation. The descriptive analytics layer provides station specific and comparative summaries of relationships between discharge and meteorological drivers. A geospatial module maps hydrological and meteorological stations in one interface, enabling users to inspect station proximity, network coverage, and upstream/downstream context while reviewing results. The workflow is explicitly human-in-the-loop: after automatic detection, domain experts can inspect hydrographs, compare neighboring stations on the same river system, and manually adjust detected dates when local hydrological knowledge indicates refinement is needed. This step provides practical quality assurance and strengthens trust in operational use. A key final outcome is automated generation of standardized report tables with additional derived indicators across stations. These products are designed for periodic institutional reporting used by the Hydrometeorology and Monitoring Center of Armenia. The framework is intended for long-term institutional maintenance.

Rapid Crisis-Oriented Survey Visualization in Official Statistics: Flood Damage in Polish Agriculture Using R

Kamil Grabarczyk[1]

1 - Statistics Poland (Statistical Office in Zielona Góra) (Poland)

The 2024 floods in the Oder river basin significantly affected agricultural regions in southern and western Poland, creating an urgent need for rapid statistical insight into the scale and spatial distribution of damages. In response, Statistics Poland incorporated additional flood-related questions into the regular Business Tendency in Agriculture Survey (AK-R) conducted in January 2025. Respondents were asked to assess the impact of flood damage on their agricultural holdings using a three-level severity scale ranging from minor impacts to threats to farm stability. This presentation demonstrates how R was used to transform newly collected survey responses into interactive geospatial analytical outputs supporting interpretation, communication, and potential policy-oriented applications. The workflow integrated official survey microdata with administrative boundary datasets and hydrographic layers using the sf ecosystem and related spatial analysis packages. Two complementary visualization approaches were developed. The first approach used randomized point allocation within municipalities to represent affected agricultural holdings while preserving confidentiality and minimizing disclosure risk. This method allowed the visualization of spatial patterns without revealing actual farm locations or sensitive respondent characteristics. The second approach applied density-based spatial smoothing techniques to produce regional intensity maps identifying clusters of reported damages and broader flood impact zones. The presentation discusses the methodological trade-offs between analytical usefulness and statistical confidentiality, particularly in the context of official statistics dissemination. It also highlights the flexibility of existing survey infrastructure in responding to unexpected crisis situations and demonstrates how open-source tools can support rapid analytical production within public statistical institutions. The resulting visualizations revealed clear spatial concentration patterns corresponding to flood-affected regions, particularly in the Opolskie, Dolnośląskie, Śląskie, Lubuskie, Małopolskie, Wielkopolskie, and Świętokrzyskie voivodeships. The developed workflow provides a reusable framework for future rapid-response statistical visualization tasks related to environmental events, agricultural shocks, or other emerging phenomena requiring timely spatial analysis. The presentation additionally reflects on reproducibility, scalability, and the role of R-based geospatial workflows in modernizing analytical communication in official statistics.

rebalancedNoise: Perturbative Statistical Disclosure Control for Magnitude Tables in R

Bernhard Meindl[1]

1 - Statistics Austria (Austria)

Publishing table data in official statistics requires protecting respondent privacy while keeping the data useful. Traditional suppression-based methods often hides large amounts of safe data to protect sensitive values, while standard noise-injection methods can break the mathematical totals across complex hierarchies. This contribution introduces rebalancedNoise (https://github.com/sdcTools/rebalancedNoise), an open-source R package that implements the EZS method (Sabolová, Almberg, & Others, 2025). The package applies record-level noise combined with a balancing algorithm. This approach protects sensitive cells under strict disclosure control rules, but minimizes the noise added to non-sensitive cells. As a result, all row and column totals in the hierarchy remain perfectly consistent and correct. `rebalancedNoise` uses a clear object-oriented setup that connects directly with standard R packages in the domain such as sdcHierarchies (Meindl, 2026). It can handle multiple numerical variables at the same time to ensure consistent results across the entire dataset. To help users check data quality before publication, rebalancedNoise includes a built-in $summarize() function. This tool breaks down the table into overall, sensitive, and rebalanced sections, showing changes in the Mean Absolute Percentage Error (MAPE) and data distributions. This allows data managers to easily see how well the noise was minimized and check for any large data distortions before releasing the tables. References Meindl, B. (2026). sdcHierarchies: Create and (interactively) modify nested hierarchies. https://doi.org/10.32614/CRAN.package.sdcHierarchies Sabolová, R., Almberg, D., & Others. (2025). Using perturbative methods for magnitude tables in statistical disclosure control. Proceedings of the UNECE Work Session on Statistical Data Confidentiality. Retrieved from https://unece.org/sites/default/files/2025-10/SDC2025_Sf_Sweden_Almberg_D.pdf

Reconciling Local and National Demographic Accounts in R

Daniel Ward[1]

1 - Office for National Statistics (United Kingdom)

Producing demographic estimates that are both locally detailed and nationally coherent is a persistent challenge in official statistics. National statistical systems increasingly rely on noisy, incomplete and sometimes conflicting administrative data sources to estimate population stocks and migration flows. Recent Bayesian demographic accounting work in R, including the dpmaccount package, has shown how these sources can be combined to generate coherent estimates with explicit uncertainty. This is a major advance for demographic estimation, especially where local-level variation and uncertainty quantification are central to decision-making. However, important operational challenges remain. In practice, official statistics often require some national totals to be treated as fixed counts rather than uncertain targets. For example, annual Long-Term International Migration (LTIM) totals or cross-border flows may need to be enforced exactly, even while estimates are being produced for local authority, age and sex groups. Standard demographic accounting approaches can struggle in this setting, because methods that preserve local structure and uncertainty do not automatically guarantee exact agreement with externally imposed macro-level totals. This work addresses that gap by developing new R-based methods for reconciling local and national demographic accounts while preserving as much inferential richness as possible. Rather than relying solely on a single Bayesian framework, we explore a range of constraint-based reconciliation methods that disaggregate fixed national totals to lower levels without breaking consistency elsewhere in the system. These include constrained optimisation approaches, iterative proportional fitting and state-space modelling. A key aim is not only to force coherence, but to do so in a way that remains useful for official statistics production. By iteratively reconciling distributions rather than point estimates alone, we can generate uncertainty intervals that remain meaningful even under strict national constraints. This offers a practical route through one of the central tensions in demographic estimation: how to produce rich local models without sacrificing the hard consistency requirements of the wider operational statistical system. The presentation will outline the underlying ideas, show how they are implemented in R, and discuss the broader implications for the next generation of demographic estimation in official statistics.

regimeseats: An Adaptive State-Space Extension for TRAMO-SEATS in Official Statistics

Cheyenne Amoroso[1], Germán Aneiros[1], Carolina García-Martos[1], Jan Koopman Siem[1], José A. Vilar[1]

1 - University of A Coruña = Universidade da Coruña (Spain)

Seasonal adjustment is a key task in official statistics to ensure a correct interpretation of the current economic situation. Following Eurostat Guidelines, National Statistical Institutes rely on TRAMO-SEATS, a well-established ARIMA model-based approach that assumes parameter stability over time. However, time series covering long periods are frequently affected by structural breaks arising from economic crises or institutional reforms. Traditional time-invariant ARIMA models typically address these sudden shifts by saturating the time series with deterministic outliers, such as level shifts or temporary changes. While this strategy successfully stabilizes in-sample estimation, it artificially suppresses the stochastic variance of the shock. To address these critical shortcomings within the R ecosystem, we present regimeseats, a novel R package that introduces an auto-adaptive state-space extension for traditional TRAMO-SEATS models. Designed specifically with the rigorous demands of NSI production pipelines in mind, the package seamlessly bridges existing, widely-used R infrastructure. It integrates the robust preprocessing and linearization capabilities of rjd3tramoseats with the advanced state-space filtering and smoothing algorithms provided by the KFAS package. By unifying these tools, regimeseats offers a robust framework capable of handling non-stationary volatility and structural breaks without discarding the established seasonal adjustment methodology. The core methodology of regimeseats relies on an Augmented Box Model that facilitates a smooth, logistic interpolation between distinct pre- and post-crisis regimes. Instead of relying on manual interventions or the subjective allocation of dummy variables, the package features an automated structural break detection algorithm driven by rolling residual volatility. This allows the model to identify volatility clusters and adapt its parameters autonomously. Furthermore, regimeseats is tailored for the strict constraints of official statistics: it crucially allows practitioners to entirely freeze historical pre-crisis models. This feature ensures that historical revisions are minimized while the model simultaneously adapts dynamically to new economic realities at the end of the series. In this presentation, we demonstrate the theoretical foundations of the regimeseats package and showcase its practical application to a set of official Spanish macroeconomic indicators as the Gross Domestic Product (GDP) or the Industrial Production Index (IPI), among others. These series differ in frequency and sectoral coverage and are affected by structural breaks of varying nature and intensity, which makes them particularly well suited to illustrate the package.

Reproducible Statistical Publications with Quarto

Christophe Dervieux[1]

1 - Posit PBC (United States)

Statistical offices produce a wide variety of publications — bulletins, indicators reports, dissemination documents — that must meet different format requirements depending on the audience: an HTML page for web dissemination, a formatted PDF for print or archiving, a Word document for editorial review. Maintaining separate documents for each format is time-consuming and introduces inconsistencies when content needs to be updated. Quarto is an open-source publishing system — the next generation of R Markdown — that addresses this directly. A single Quarto document combines R code and narrative text: tables, figures, and inline numbers are computed from data, and the same source file can be rendered into multiple output formats simultaneously. Updating the analysis updates all outputs at once, with no copy-paste between tools. This talk demonstrates this workflow through a concrete example built around real open data. Starting from existing R code that retrieves and processes data from a public source, we build a Quarto document that produces both a publication-ready HTML report and a professionally formatted PDF — with captioned figures, automated tables, and cross-references — from a single source file. The talk focuses on what makes this practical: how little additional markup is needed beyond existing R code, and what each output format offers for statistical dissemination. No prior experience with Quarto is required. Statisticians already using R for data processing will find it a natural extension of their workflow. Those familiar with R Markdown will recognise the approach and see what Quarto adds: consistent multi-format support, better PDF output, and an active ecosystem increasingly used across the official statistics community.

samplyr: A Tidy Grammar for Survey Sampling Design in R

Dicko Ahmadou[1]

1 - World Bank - UNHCR Joint Data Center on Forced Displacement (Denmark)

In national statistical offices, sampling designs are carefully documented in methodological reports, with an explicit logic connecting strata, clusters, sampling stages, and selection probabilities. This structure is often lost in the R code that implements the design, where the plan gets scattered across intermediate objects, successive calls, and technical details. The more complex the design, the wider the gap between the documentation and the code, making implementations harder to review, audit, and reuse across survey rounds. samplyr was built to close this gap. The package provides a declarative grammar inspired by the tidy ecosystem, where each verb maps to a sampling operation: sampling_design() to initiate a plan, stratify_by() to stratify and allocate, cluster_by() to define clusters, draw() to specify method and size, add_stage() to chain stages, and execute() to draw the sample from a frame. The design is expressed in code that reads like the methodological report that describes it, and the same object documents the plan, draws the sample, and carries the information needed for estimation with the survey package. The talk will illustrate this approach with designs common in official statistics, including stratified multistage household surveys. We will show how an explicit expression of the design improves transparency and reproducibility, reduces implementation errors, and supports knowledge transfer within statistical offices. samplyr grew out of work with national statistical offices in West and Central Africa, where sampling has often relied on proprietary tools such as SPSS (Complex Samples) or SAS (Surveyselect), or on spreadsheet-based selection that limits the complexity of feasible designs. By offering a free, open-source grammar that remains readable for statisticians coming from these tools, samplyr supports a sustainable transition to R for the full sampling workflow, from design to selection to estimation, and aims to make sampling code read the way sampling designs are written. Link: https://dickoa.gitlab.io/samplyr/

Scaling Local LLM Coding Assistants for R, Python and SAS in Official Statistics

Bernhard Fischer[1]

1 - Federal Statistical Office of Germany (Germany)

Large language models are increasingly used as coding assistants, but their adoption in official statistics is constrained by confidentiality, reproducibility and governance requirements. Sending source code, metadata or statistical production logic to external services is often not possible. At the same time, statistical offices face a growing demand for support in R and Python programming, automation of analytical workflows and migration away from legacy environments such as SAS. This contribution presents our technical approach for deploying local LLMs as a scalable internal coding assistance service for statistical programming. The setup is based on centrally hosted models exposed through a single API endpoint. It separates model serving, access management and monitoring, allowing the service to be used by a larger number of developers and statisticians while keeping interactions within the organization’s infrastructure. The talk includes practical use cases for R, Python and SAS-related workflows. These include generating and explaining R code, refactoring existing scripts, translating selected SAS patterns into R, and assisting with interactions between R and SAS. A specific example is the use of domain-specific “skills” that teach the model how to access a SAS system through the saspy Python package and a corresponding R wrapper. Such predefined skills can also provide reusable instructions, code templates and organizational conventions, enabling the model to support workflows that are not covered well by general-purpose training data. Beyond individual coding tasks, the contribution discusses technical and operational challenges when scaling local LLM-based assistance: model selection, GPU resource management, latency, concurrent usage, context management, prompt and skill governance, logging, evaluation and user training. Special attention is given to the limitations of LLMs in production-related statistical work, including hallucinated code, incomplete SAS-to-R translations and the need for systematic validation and code review. The presentation aims to share lessons learned from moving from experimentation with local models towards an operational service for many users. It argues that local LLMs can become a practical component of modern statistical IT environments if they are combined with clear governance, domain-specific skills and integration into existing R, Python and SAS workflows.

Small Area Estimation of Poverty in the Practice of Statistics Poland

Marcin Szymkowiak[1], Kamil Wilak[1], Tomasz Józefowski[1], Nobuo Yoshida[2]

1 - Poznań University of Economics and Business; Statistics Poland (Poland), 2 - World Bank (United States)

The European Survey on Income and Living Conditions (EU-SILC) is the primary source of poverty information published by Statistics Poland, both for the country as a whole and at the voivodeship (province) level. However, owing to the limited sample size in the relevant cross-classifications, EU-SILC data do not allow reliable poverty indicators to be published at lower levels of spatial aggregation. At the same time, the growing demand for accurate poverty maps - used, among other things, in the allocation of development funds - calls for small area estimation (SAE) methods and for combining data from different sources (EU-SILC, the census, administrative registers). In a collaboration between the Statistical Office in Poznań, Statistics Poland and the World Bank, poverty indicators were estimated at the level of subregions (NUTS 3) - a level not yet officially published in Poland - using univariate and multivariate Fay-Herriot models. All computations were carried out in R, with the central tool being the EU SAE Dashboard: a single Shiny application developed by the World Bank, drawing on experience gained in cooperation with Statistics Poland, that guides the user through the successive stages of small area estimation of poverty - from data preparation and diagnostics, through the selection of explanatory variables and model estimation, to assessing the quality of the estimates and visualising them on choropleth maps. The aim of the presentation is not only to present the modelling results, but above all to showcase R as a complete environment for producing official poverty maps. The presentation will cover the data preparation stage and the architecture of the EU SAE Dashboard application, as well as the key R packages used at the individual stages: emdi and povmap for estimating the SAE models, and sf, tmap, leaflet and ggplot2 for processing and visualising spatial data.

Spipe: a lightweight workbench for building and monitoring reproducible analytical pipelines in R, illustrated on the GEM Luxembourg production and analysis chain

Maxime Pettinger[1]

1 - STATEC Research (Luxembourg)

Official statistics production often relies on long, monolithic scripts that are hard to read, test, reuse and maintain. Reproducible analytical pipelines (RAP) offer a way forward, but adopting them on an existing production or analysis chain raises practical questions: how to modularise legacy code, how to assemble steps into a readable workflow, and how to monitor and iterate on it efficiently. This talk presents Spipe, a lightweight R package developed while converting segments of the Global Entrepreneurship Monitor (GEM) Luxembourg production and analysis chain into R, following what we call a """"""pipeline programming style""""""" grounded in RAP principles. The original chain was split into modules rewritten as small, pipeline-friendly functions, complemented by helpers to build and monitor pipelines, now bundled in Spipe. Spipe is a thin layer over R’s native pipe that acts as a workbench for building, running and monitoring data production and analysis chains of small to moderate size. The modular blocks it helps shape are plain functions that can later be orchestrated by dedicated tools such as targets or airflow, easing the transition from exploratory scripts to packaged code. Using the GEM use case, the talk shows how Spipe instruments a pipeline: wrapping any function or expression into a monitored block, marking a pipeline’s start and end, logging and timing each step, checkpointing and serialising intermediate outputs, skipping or """"""passing through""""""" expensive segments by reusing cached results, and pausing or stopping a run interactively. Several of these mechanisms, notably the skipping/""""""passing through""""""" of segments, leverage the flexibility of R’s S3 object system. The demonstrated pipeline recodes and derives variables, attaches labels, produces several hundred weighted proportion and mean tables with confidence intervals (using survey/srvyr), reshapes the resulting nested results by S3 class, and plots them. The session emphasises readability, reproducibility and a smooth path from interactive exploration to documented, packaged, monitorable pipelines. Spipe is still work in progress and not yet available on a public repository; it builds on common CRAN packages, and this set of dependencies may still evolve. Attendees will leave with reusable patterns, and an outlook on Spipe, to turn their own production code into reproducible, monitorable pipelines that downstream orchestrators can pick up.

for Official Statistics: A Method for Data Editing Using Large Language Models

Bjørkholt Solveig[1], Islam Mousumi Mona[1], Torvholm Frøystad Sarah[1], Johansen Odin[1], Javed Maryam[1]

1 - Statistics Norway (Norway)

In business statistics, annual reports are a core source of information for data cleaning and editing. Official statistics producers both validate and impute numbers using data drawn from companies' annual reports. This editing work is time consuming due to the unstructured format of these documents, which leaves producers with the manual task of looking up numbers and comparing them to survey responses. A standardized data source of core numbers from annual reports would reduce the burden on both respondents and statistics producers, offering efficiency gains without compromising statistical quality. Yet, extracting structured information from annual reports has proven challenging. Earlier approaches, using open-source Optical Character Recognition (OCR) and rule-based parsing, have been unreliable given the non-standardized layout and occasionally poor quality of these reports. We therefore assess a new method using large language models to perform targeted structuring of annual reports for data editing. With the ellmer package, we demonstrate how these models can be integrated into an API-based pipeline that extracts and structures information from the reports and adds it to a common database, automating the manual task of number extraction. We compare Gemini 2.5 Pro with local models to evaluate the cost–quality tradeoff, and discuss operability in closed production environments. To illustrate the method, we use external economy statistics as a case study, a statistic characterized by a complex survey questionnaire and substantial editing requirements. We demonstrate how the method can be integrated into an existing production pipeline to streamline editing tasks. Looking further ahead, the approach could also reduce survey burden by enabling direct imputation from annual reports into official statistics pipelines, so that companies whose reports already contain the relevant numbers need not report them separately.

The Metadata Quality Assistant (MQA): an R/Shiny prototype for supporting the preparation of SIMS quality reports

Vasiliki Spiliopoulou[1], Christina Pierrakou[1], Konstantinos Molfetas[1]

1 - Hellenic Statistical Authority (Greece)

In official statistics, the quality of reference metadata is as important as the quality of the data themselves. The European Statistical System uses the Single Integrated Metadata Structure (SIMS) as its common conceptual framework for quality and reference metadata reporting. In practice, NSIs and ONAs still compile SIMS reports as long Word templates filled in manually by domain authors, which may result in empty fields, placeholder text, and descriptions that are either too brief to be informative or too vague to be useful, making the subsequent quality-control review slow and subjective. We present the first prototype of the Metadata Quality Assistant (MQA), an open-source R/Shiny application developed at the Hellenic Statistical Authority (ELSTAT) to support the preparation of SIMS quality reports. MQA is positioned as an aid for metadata authors rather than an audit tool: in its first phase, its purpose is to perform completeness checks on each field before the report is sent for internal review, not to replace the human reviewer. The application is built around a clear separation between a stable R/Shiny engine and an extensible rule catalogue. Authors upload a completed report in .docx format directly through the browser. The application walks the document's tabular structure, locates each SIMS sub-concept by its section code and extracts the user-supplied text without altering the source file. The extracted content is then evaluated against the rule catalogue: mandatory-field presence, minimum text length, placeholder detection and keyword coverage. Keyword matching is accent-insensitive through a Greek-Latin transliteration step, which removes a recurring source of false negatives for Greek-language reports. The interface returns a scorecard of total, valid, critical and improvement-needed fields together with a colour-coded, field-by-field audit trail that the author can use as a checklist. In the current proof of concept the rule catalogue is maintained as a spreadsheet; as the catalogue grows, the rule catalogue is expected to migrate to a medium better suited to large-scale extension and governance. The presentation will outline the design choices behind MQA, the role of the rule catalogue in keeping the assessment logic flexible, and directions for further development of the tool.

The RJDverse Package Suite for Time Series Econometrics in Official Statistics

Karsten Webel[1], Anna Smyk[2]

1 - Deutsche Bundesbank (Germany), 2 - Institut national de la statistique et des études économiques (France)

Raw economic data recorded periodically by national statistical agencies or central banks often undergo processing before publication. Univariate sub-annual time series with nonignorable seasonal fluctuations, such as monthly industrial production, are typically seasonally adjusted, so that their underlying long-term trends and sudden short-lived unusual movements can be recognised more easily. Systems of (transformed) univariate time series, such as (seasonally adjusted) quarterly national accounts, are usually reconciled to establish, or restore, contemporaneous and/or temporal constraints. A similar treatment is often required for data that measure the same economic variables but originate from different sources, such as business surveys conducted regularly at annual and quarterly frequencies. The JDemetra+ time series software implements numerous methods and tools for fulfilling these and other data processing tasks. This Java-based program is mainly developed at the National Bank of Belgium in cooperation with the French National Institute for Statistics and the Deutsche Bundesbank. It has been recommended by Eurostat and the European Central Bank for the harmonised production of official data within the European Statistical System and the European System of Central Banks since February 2015. Around 2020, the main developers also created a first set of core functions for making the underlying Java computing routines more easily accessible in R, primarily to facilitate testing. Over time, however, these initial code chunks have morphed into a whole ecosystem of R packages baptised the RJDverse on the GitHub platform in 2024. Motivated by achieving CRAN status in 2026, we provide an overview of the various packages, which can be broadly categorised into general versus specific purpose tools. Packages belonging to the first category mainly bundle auxiliary and utility functions and, hence, constitute the foundation of the RJDverse. They provide, for example, filtering and smoothing algorithms for state space models as well as functions for wrangling with both JDemetra+ workspaces and input data, for generating exogenous regression variables, and for calculating standard diagnostics, such as normality and seasonality tests. In contrast, packages belonging to the second category bundle functions for specialised tasks and, hence, constitute the pillars of the RJDverse. Their implemented methods cover, for example, nonparametric and ARIMA model-based seasonal adjustment, trend-cycle estimation, benchmarking and reconciliation, regression-based temporal disaggregation and interpolation, calendarization of flow time series, revision analysis, and nowcasting by means of dynamic factor models. Our overview is purely nontechnical and prioritises illustration, including code snippets, over methodological rigour.

The Survey Fieldwork Digital Twin: An R-Based Early-Warning System for Mixed-Mode Official Statistics

Hakile Resulbegoviq[1,2], Vuk Čađenović[1]

1 - Damar Institute (Montenegro), 2 - University of Ljubljana (Slovenia)

Survey errors often become visible before the final dataset is produced. In mixed-mode official surveys, early signs of quality problems can appear during data collection itself: very short interviews, increasing item nonresponse, slow regional sample realization, interviewer-specific response patterns, web breakoffs, mode differences, or validation errors that repeat across cases. Standard data cleaning can detect many of these problems, but often too late for retraining, questionnaire adjustment, or reallocation of fieldwork resources. This paper develops the idea of a survey fieldwork digital twin as an R-based system for monitoring mixed-mode official surveys while fieldwork is still in progress. The system builds a reproducible reference model of expected fieldwork behaviour and compares it with incoming data from PAPI, CATI, CAPI and CAWI modes. It brings together survey responses, paradata, sample progress, interviewer information, validation rules and methodological notes in one workflow. The system monitors differences between expected and observed patterns in sample progress, interview duration, item nonresponse, straightlining, validation errors, interviewer-level indicators and mode-specific response distributions. These differences are not treated as automatic errors, but as signals that require review. Each signal is linked to a structured review process in which survey methodologists and field supervisors assess the evidence, decide whether an intervention is needed, and record the decision. The decision log is the main element that separates this approach from a standard monitoring dashboard. It records what was flagged, why it mattered, who reviewed it, what action was taken, and whether the issue was resolved. This creates a documented link between fieldwork data, quality control decisions and later methodological evaluation. The paper presents a practical R workflow for building a prototype of such a system. The workflow covers data harmonisation, expected-progress tables, validation checks, early-warning thresholds, interviewer diagnostics, mode comparisons, dashboard outputs and automated reporting using Quarto or R Markdown. The approach is intended for national statistical offices and research institutes that need reproducible quality control without relying on expensive proprietary systems. By moving quality assurance closer to the fieldwork phase, the survey fieldwork digital twin helps make mixed-mode survey production more timely, auditable and reproducible. Its contribution is the connection between fieldwork signals and documented methodological decisions, rather than the creation of another monitoring dashboard.

The {ProduceR} package: streamlining and improving the reliability of statistical production

Vincent Reduron[1]

1 - Direction de la recherche, des études, de l’évaluation et des statistiques (France)

The {ProduceR} package offers a set of five functions designed to meet common needs in statistical production - a term referring to the creation of reliable and usable statistical data from raw information. These functions were developed from the practical experience of their author, who has worked in statistical agencies for 20 years. The underlying philosophy is the following. Statistical production generally involves manipulating complex data, particularly when the raw information consists of administrative data, but also when it consists of survey data. This information is almost always found in a set of tables, as tabular representation is the canonical model for statistics. High-quality production requires understanding the structure of these tables and their content, as well as mastering the relationships between them. However, in day-to-day work, one is reluctant to carry out checks if they take a long time to program; manual exploration often replaces a comprehensive check. Nevertheless, it is highly important to check the data at different stages of production: for example, joining two tables must be preceded by verifying their unique keys, followed by checking for the absence of duplicates or unwanted missing values. Thus, the five functions of {ProduceR} were developed to be concise and accessible, so that the producer does not hesitate to perform extensive checks. They are intentionally limited in number and targeted at the main needs encountered in real-world scenarios (analysis of duplicates and missing values, summary analysis of variables, table comparison). The package is available on CRAN at the URL https://CRAN.R-project.org/package=ProduceR.

Towards metrics for awesome software packages: first results

Olav ten Bosch[1], Mark van Der Loo[2,3]

1 - Statistics Netherlands (Netherlands), 2 - Eurostat (Luxembourg), 3 - Leiden University (Netherlands)

The """"Awesome List of Official Statistics Software"""" is a collaborative repository of free and open-source software (FOSS) components utilized in the production of, or providing access to, official statistics. Much of this list comprises R-packages showcased at statistical conferences, functioning as a shared institutional memory that facilitates the reuse of existing tools in future projects. Building on discussions from 2025 regarding the maturation of the official statistics software landscape, this presentation explores the development of quality metrics for these packages. We propose evaluating software based on characteristics such as alignment with statistical standards, domain neutrality, granularity, software independence, uncertainty propagation, and privacy-by-design. Crucially, these metrics integrate the """"Statistical Open-Source Software Guiding Principles,"""" endorsed by the UNECE Conference of European Statisticians (CES) in 2025. An enhanced awesome list incorporating these metrics aims to promote a more mature and standardized FOSS ecosystem, characterized by high-quality, composable components that maximize functionality for statistical organizations. In this presentation, we report on the evolution of the list over the past year and share early results from our implementation of these metrics. By mapping UNECE principles to candidate metrics, we have identified a subset that can be calculated automatically using metadata from packaging systems, version control systems, and official statistics journals. We conclude with reflections on our progress and the future of the initiative. Disclaimer: The views expressed in this research are those of the authors only and do not necessarily reflect the opinions of their respective institutes.

Unsupervised Learning Approaches for Identifying Structural Anomalies in European Regional Statistics

Bogdan Oancea[1], Ciprian Alexandru[2]

1 - University of Bucharest (Romania), 2 - Ecological University of Bucharest (Romania)

Ensuring the coherence and analytical quality of regional socio-economic statistics is a central task for national statistical institutes and for the European Statistical System. Standard validation procedures, such as range edits, ratio checks and univariate outlier detection, remain essential for identifying implausible or extreme values in individual indicators. However, they are less effective when the objective is to detect unusual multivariate configurations across several regional indicators. This paper presents an R-based, fully reproducible framework for identifying structurally atypical regional profiles in Europe using publicly available Eurostat data. We construct a cross-sectional dataset of NUTS 2 regions for 2022, covering four core socio-economic indicators: GDP per capita in purchasing power standards, unemployment rate, tertiary educational attainment and population density. The analysis is implemented entirely in R and combines data acquisition, preprocessing, anomaly detection, visualisation and reporting in a transparent workflow. We compare five anomaly detection approaches: univariate z-scores, Mahalanobis distance, Isolation Forest, Local Outlier Factor and One-Class Support Vector Machines. A region is classified as structurally anomalous when it is identified by at least three of the five methods. The results show that unsupervised machine learning can complement traditional validation tools by identifying regions whose multivariate profiles differ substantially from the broader European pattern. The detected cases include highly developed metropolitan economies such as Brussels, Vienna, Berlin and Prague, regions affected by persistent socio-economic disadvantages such as parts of Slovakia, Hungary, Castilla-La Mancha and Extremadura, as well as Istanbul, whose profile differs markedly from that of EU capital regions. These cases should not be interpreted as data errors. Rather, they indicate meaningful structural divergence that may require further analytical, methodological or policy attention. The contribution of the paper is methodological and practical. It demonstrates how R can be used to build scalable, reproducible and interpretable workflows for multivariate validation and exploratory quality assessment in official statistics. The proposed framework can be integrated into existing statistical production and validation systems, supporting early detection of unusual regional configurations while preserving transparency and reproducibility.

Using LLM’s for automated translation of Excel files and SAS scripts to R

Andreas Jahn[1]

1 - Destatis (Germany)

Many statistical production processes at the Federal Statistical Office currently rely on a heterogeneous software landscape that includes SAS, Microsoft Excel files, and R-based parallel processing environments. As part of a broader modernization strategy, R is intended to become the leading platform for data preparation, analysis, and statistical production. A key challenge in this transition is the migration of existing business logic that is embedded in complex Excel workbooks and SAS programs. This presentation introduces an agent-based approach that leverages Large Language Models (LLMs) to support the automated translation of binary Excel files into reproducible and maintainable R code. The proposed framework combines specialized software agents that analyze workbook structures, identify calculation logic, extract transformation rules, and generate corresponding R implementations. By decomposing the migration task into smaller, domain-specific subtasks, the approach aims to improve both transparency and translation quality while reducing manual effort. Building on this foundation, the framework is extended to support the automated translation of SAS scripts into R. In contrast to conventional code conversion tools, the agent-based architecture integrates multiple stages of validation, quality assurance, and testing. Generated R code is automatically reviewed, compared against expected outputs, and subjected to predefined test cases to ensure functional equivalence with the original SAS implementation. The workflow therefore not only automates code generation but also addresses critical requirements for reliability and reproducibility in official statistics. The presentation discusses the overall architecture, implementation experiences, and initial results from practical migration scenarios. Particular attention is given to the role of LLMs in understanding legacy analytical workflows, the challenges associated with translating procedural and spreadsheet-based logic, and the opportunities offered by agentic AI systems for large-scale software modernization. The proposed approach demonstrates how LLM-driven agents can support the transition toward an R-centered statistical production environment while maintaining quality standards and reducing migration costs.

Using R to provide a new interface to an old time series program

Guy Mélard[1]

1 - Univerité libre de Bruxelles, Faculté SBS-EM & ECARES (Belgium)

ANSECH (from the French terms: ANalyse des SEries CHronologiques) is a Fortran program [1, 2, 3] designed to handle time series modeling with a few peculiarities. In particular, it accepts time-dependent models with marginal heteroscedasticity, deterministic seasonality, and direct fitting of a large number of exponential smoothing methods, including damped trend models, etc. The program can also run Monte Carlo experiments. From a numerical point of view, there is now an optional 128-bit computing engine that provides standard errors for parameter estimates that are much more precise than usual [4]. Two additional reasons to keep ANSECH working are that there is now a strong theory [5] supporting ARIMA models with time-dependent coefficients, and they are considered seriously for model-based seasonal adjustment [6]. Unfortunately, the interfaces built on top of ANSECH [7, 8] are difficult to maintain, and we should abandon them. The idea, therefore, is to use R as a new interface. We have considered doing this in levels: first, running the input command file and recovering the output file with appropriate new plots; second, building an interactive interface to construct the input command file; and third, in the simplest cases, passing arguments directly. That way, all the existing potential of ANSECH will be salvaged with minimal effort. Reprogramming it in R would be time-consuming, and the resulting program would be very slow and imprecise. References [1] Bossier, F. et Mélard, G. (1978), Une application économique des modèles ARIMA non stationnaires, Publications Econométriques, 11, 1978, 13–31. [2] Mélard, G. (1982), Software for time series analysis. In: Caussinus, H., Ettinger, P., Tomassone, R. (eds) COMPSTAT 1982 5th Symposium held at Toulouse, Physica, Heidelberg, pp. 336–341. https://doi.org/10.1007/978-3-642-51461-6_50. [3] Mélard, G. (2012), Time Series Analysis: ANSECH 1973-2012, Communication aux Journées Internationales Analyse statistique : Théorie et Applications, June 4-6, Oujda, Maroc. https://www.researchgate.net/publication/254453491_ANSECH_Logiciel_pour_l'analyse_des_series_chronologiques. [4] Mélard, G. (2024), Standard errors in nonlinear models like time series models, ITISE 2024, 15th-17th July, 2024, Gran Canaria (Spain), https://itise.ugr.es/PresentacionPDF/ITISE2024_Slides_presentation_232.pdf [5] Alj, A, Azrak, R., and Mélard, G. (2025), General estimation results for tdVARMA array models, Journal of Time Series Analysis 46, 137−151. https://doi.org/10.1111/jtsa.12761 [6] Mélard, G. and Palate, J. (2025), Seasonal adjustment using time-dependent ARIMA models: implementation in JDemetra+ and an R package, INSEE/OECD Workshop on Time Series Analysis and Statistical Disclosure Control Methods for Official Statistics, December 3-4, 2025 - OECD, Headquarters (Paris), France.https://www.researchgate.net/publication/407756756_Seasonal_adjustment_using_time-dependent_ARIMA_models_implementation_in_JDemetra_and_an_R_package [7] Mélard, G. (1990), Méthodes de prévision à court terme, Editions de l'Université de Bruxelles, Bruxelles, et Editions Ellipses, Paris, 1990, 468 pages ; 2e édition 2007, 537 pages, avec CD-ROM. [8] Mélard, G. and Pasteels, J.-M. (2000), Automatic ARIMA modeling including interventions, using time series expert software, International Journal of Forecasting, 16, 497–508.

weightflow: a declarative, auditable recipe for survey weighting in R

Juan Pablo Ferreira[1,2]

1 - National Institute of Statistics (INE) (Uruguay), 2 - Universidad de la República de Uruguay = University of the Republic of Uruguay [Montevideo] (Uruguay)

Survey weighting is a multi-step process: design base weights are adjusted for unknown eligibility, ineligibility, within-household selection and nonresponse, then calibrated to population totals, trimmed and rounded. In practice it often lives in long, fragmented scripts that are hard to review, reproduce and defend. weightflow, now on CRAN, expresses the whole process as a single explicit, pipeable recipe. Borrowing the specification/execution split of the recipes/tidymodels ecosystem, a weighting recipe is inert: each step (step_unknown_eligibility, step_nonresponse, step_calibrate, step_trim_weights, and others) is declared in run order, inspected, and only then estimated with prep(). The cascade reads top to bottom and is audited step by step, matching how statistical offices document the weights they release. It covers the standard toolkit and several extensions: nonresponse via weighting classes or response propensities (logistic regression, CART, random forests or boosting, with optional cross-fitting), and automatic R-indicators to monitor representativity; calibration by raking, post-stratification, linear/GREG, model-assisted (Wu-Sitter) and ridge-penalised methods, including domain-partitioned calibration and tidy data-frame totals (for instance population projections) supplied without hand-built cell variables, plus integrated household calibration (one weight per dwelling) and Potter's MSE-optimal trimming. Person- or household-level adjustment is set through a single cluster argument. A distinctive feature is recipe-aware variance estimation: a bootstrap resamples PSUs within strata and re-runs the entire recipe on each replicate, so standard errors reflect the sampling design and every weighting adjustment at once, not the design alone. The replicate weights hand off to survey and srvyr and are portable to other software an office may use. The talk works through the pipeline end to end on real microdata from Uruguay's continuous household survey (ECH): it induces realistic eligibility and nonresponse, weights the survivors back with integrated household calibration, and validates the resulting poverty-rate estimate against a known population value, with design-based confidence intervals. weightflow has no hard dependencies beyond base R, and generates self-contained HTML weighting reports. The aim is a transparent, reproducible and production-oriented approach to survey weighting for official statistics.

What makes the R community unique?

Romain Lesur[1]

1 - Insee (France)

Among the many open-source languages, R occupies a singular position, not merely because of its statistical capabilities, but because of the community that has grown around it. What makes the R community truly distinctive is not technical: it is sociological. Other languages have active communities, diversity initiatives, and resources for learners. What distinguishes R is the density and institutional depth with which three dimensions coexist: pedagogy, inclusion, and reproducibility. A community built on pedagogy. R was created in 1993 by two statisticians at the University of Auckland, and this academic origin has shaped the community's priorities. Teaching and knowledge transfer have always been central concerns, producing a large body of freely available learning resources written for audiences with limited programming backgrounds. This matters for NSIs, where statistical expertise is common but programming skills are unevenly distributed. A community built on inclusion. Most open-source languages have diversity initiatives. What sets R apart is how formally these are embedded in its governance. R Forwards is a task force of the R Foundation itself, not a peripheral community group. R-Ladies, founded in 2012, now counts over 47,000 members across 159 cities and 47 countries, considerably larger than comparable initiatives in other languages. rainbowR supports and connects LGBTQ+ R users. AfricaR and MiR (Minorities in R) Community focus on underrepresented geographic and ethnic groups. These are not informal initiatives: they are coordinated, sustained, and formally embedded in the governance of R itself. A community built on reproducibility. With Sweave, knitr, R Markdown, and Quarto, R has made reproducible workflows accessible with relatively little setup. Other languages offer comparable tools, but the integration of code, narrative, and output has been a design priority in R for longer, and the culture around it runs deep. This aligns naturally with official statistics, where methodological transparency is a baseline requirement. By joining the R community, statistical institutes do not simply gain a tool: they connect with a global network that shares their commitment to open science, rigorous methodology, and the democratisation of statistical knowledge.

What’s new for time-to-event data in tidymodels

Hannah Frick[1]

1 - Posit, PBC (United States)

If you are interested in when an event occurs, not just if the event occurs, but for some of your observations you might still be waiting for the event - then your modelling problem is neither a classification problem, nor a regression problem, it’s a censored regression problem. Survival analysis is the tool for that, even when your event isn’t """"death""""" but rather a building permit application being processed, a pet being adopted, or a complaint being resolved. The tidymodels framework is a collection of packages for safe, performant, and expressive supervised predictive modeling on tabular data. The framework’s consistency makes switching between models easy, and its guardrails against common pitfalls, such as overfitting due to data leakage, make it safe. It covers the entire modelling workflow: preprocessing and feature engineering, models, resamples, performance metrics, and tuning. tidymodels supports survival analysis across the entire framework with dedicated models and metrics, allowing the same ease and expressiveness as for classification and regression, across all steps of the modeling process. We’ve been adding more functionality and infrastructure in the areas of models and metrics. Come hear what’s new and learn how to extend the framework further!

{stlpro}: Seasonal Adjustment with Cross-Validated Filter Selection

Daniel Ollech[1]

1 - Deutsche Bundesbank (Germany)

Seasonal adjustment is a critical step in time series analysis, with the STL (Seasonal-Trend decomposition using LOESS) method widely used due to its flexibility and robustness. However, the choice of the seasonal filter parameter (s_window), which governs the smoothness of the estimated seasonal component in STL, remains a challenge. Traditional approaches rely on fixed defaults or ad-hoc rules, which may not generalise well across diverse time series. This paper introduces the {stlpro} package. It builds on the STL-method by incorporating cross-validation for data-driven selection of the s_window parameter. To our knowledge, this is the first implementation of STL that integrates cross-validation-based seasonal filter selection in a fully automated way. The package addresses key practical challenges, including: Filter diversity: Candidate s_window values are selected based on differences in squared gains of the regression weights, ensuring meaningful variation in seasonal smoothing. Missing data handling: The weights used in the LOESS regressions are adapted appropriately to handle missing values. Computational efficiency: Cross-validation is embedded within STL’s iterative steps, balancing accuracy and speed. A simulation study evaluates the performance of the cross-validation implementation across time series of varying lengths and seasonal patterns. Results show that the cross-validation based selection returns appropriate seasonal filters for most series, often choosing the filter that yields the lowest mean absolute difference between simulated and estimated seasonal component. Applications to real-world data highlight the method’s performance, especially in terms of selecting an appropriate seasonal filter, on time series used in official statistics. The {stlpro} package supports both multiplicative and additive decompositions and accommodates non-integer seasonalities, such as those observed in weekly time series. Keywords: Seasonal adjustment, STL, cross-validation, time series, filter selection

From Data Acquisition to Harmonised Outputs: A Reproducible R Workflow for Official Statistics

Ciprian Alexandru

Ecological University of Bucharest

Statistical production workflows often evolve incrementally. What begins as a collection of R scripts for downloading, cleaning and transforming individual datasets can gradually become difficult to maintain, reproduce and audit, particularly when data originate from heterogeneous sources and require different processing rules.

This presentation describes, from an R user and developer perspective, the redesign of such a workflow into a modular and reproducible R-based pipeline.

The approach separates the workflow into two R packages with distinct responsibilities. The first package, manages data acquisition from heterogeneous sources, including APIs, databases and externally provided files. The second, implements the subsequent cleaning, harmonisation and refinement stages. This separation makes it possible to distinguish source-specific data access from statistical processing logic and to develop and test both components independently.

A metadata-driven approach is used to reduce dataset-specific code. YAML configuration files describe data sources, parameters and processing requirements, while common functions translate these configurations into reproducible operations. The workflow combines package-based development with targets for pipeline orchestration, {renv} for dependency management, structured logging for monitoring and diagnostics, and Git-based version control. Intermediate and final datasets are stored together with information supporting data lineage and reproducibility.

Particular attention is given to practical challenges encountered by an R user when moving from exploratory scripts towards a production-oriented workflow: handling heterogeneous data structures, defining common data models, separating acquisition from transformation, implementing validation checks, managing data vintages, and designing logs that can subsequently support workflow monitoring and visualisation.

Lightning Talks

Automatic classification of pastoral-use parcels / Classification automatique des parcelles à usage pastoral

Romuald Thion[1]

1 - Service Statistique Ministériel de l'Agriculture et alimentation (France)

Identifying the pastoral nature of grasslands and pastures in the *Registre Parcellaire Graphique* (RPG, derived from geolocated Common Agricultural Policy subsidy declarations) is difficult: pastoral activities — i.e., extensive livestock farming on spontaneous forage resources, including summer pastures, alpine pastures, or rangelands — are highly diverse. Although each parcel is associated with a declared crop, which should allow pastoral areas to be identified, declaration practices are not homogeneous across departments of France. This constitutes a major limitation to reliably characterising pastoral areas. We present a processing pipeline for the pastoral classification of RPG parcels built on the tidymodels machine learning framework. The target variable is derived from the spatial intersection between RPG parcels and the areas surveyed by the enquête pastorale (pastoral survey), conducted since 2012, which serves as ground truth. The predictors come from CAP declarations (crop types, parcel geometries) enriched with topographic (elevation, slope, aspect…), pedological, and climatic features obtained from INRAE's Rural Development Observatory (ODR — Observatoire du Développement Rural). The gradient boosting algorithm XGBoost was selected for classification. We obtain very encouraging results for both binary and ternary classification (where pastoral areas are further split into two sub-categories): 89.0% specificity, 74.5% sensitivity, a classification error of 21.8 thousand hectares out of 132 thousand test hectares, with the three-class classification adding only 1.8 thousand hectares of misallocation. These models, trained on 396 thousand hectares, allow pastoral use to be predicted across the entire 2,900 thousand hectares of the Massif Central - a highland region covering about 15% of mainland France. Applications include recoding declared CAP crop codes, extending the prediction to the Pyrénées, Jura, or Vosges massifs, and refining typologies of grazing livestock farms. —- French version —- L'identification du caractère pastoral des prairies et pâturages du Registre Parcellaire Graphique (RPG) (issu des déclarations géolocalisées des aides de la PAC) est *difficile* : les activités pastorales, c'est-à-dire l'élevage extensif valorisant les ressources fourragères spontanées (estives, alpages, parcours), sont diversifiées. Bien que chaque parcelle soit associée à une culture déclarée, permettant en théorie d'identifier les surfaces pastorales, les pratiques de déclaration ne sont pas homogènes selon les départements. Cela constitue une limite importante à l'utilisation directe du RPG pour caractériser de manière fiable les surfaces pastorales. Nous présentons une chaîne de traitement pour la classification pastorale des parcelles du RPG sur le framework d'apprentissage automatique tidymodels. La variable cible est issue de l'intersection spatiale entre les parcelles RPG et les surfaces de l'enquête pastorale conduite depuis 2012 qui constitue la vérité terrain. Les prédicteurs sont issus des déclarations PAC (types de culture, géométries des parcelles) enrichies avec des caractéristiques topographiques (altitude, pente, exposition…), pédologiques et climatiques obtenues par l'Observatoire du Développement Rural. L'algorithme XGBoost de *gradient boosting* est retenu pour la classification. Nous obtenons des résultats très encourageants en classification binaire et ternaire (où les surfaces pastorales sont raffinées en deux sous-catégories) : 89.0% de *specificity*, 74.5% de *sensitivity*, une erreur de classification de 21.8 milliers d'hectares sur 132 de tests, la classification en trois postes n'ajoute que 1.8 millier d'hectares mal ventilés. Ces modèles entrainés sur 396 milliers d'hectares permet de prédire l'usage pastoral sur l'ensemble des 2 900 milliers d'hectares Massif central. Les applications comprennent le recodage des codes cultures déclarés à la PAC, la déclinaison de la prédiction sur les massifs des Pyrénées, du Jura ou des Vosges ou le raffinement des typologies d'exploitations herbivores.

Common pitfalls when switching from SAS to R

Pierre Lamarche[1,2]

1 - Centre de Recherche en Économie et Statistique (5, Avenue Henry Le Chatelier91120 Palaiseau France), 2 - Institut national de la statistique et des études économiques (INSEE) (France)

During the project conducted at Insee for exiting SAS, several common pitfalls were identified rooted in the specific behavior of SAS and the discrepancies between this language and the modern datascience languages. The very usual problems are caused by: - the specific way for dealing with missing values; - the row-wise approach for data treatment in SAS; - the macro language in SAS that mimics a functional approach with its imperfections. This presentation will describe briefly those specific issues and ways to deal with them in a systematic manner.

Estimating French election results like the pollsters from partial vote counts on election night: a transparent, reproducible R approach based on open public data

Thomas Guinhut[1], Eulalie Delaune[1], Romain Winckell[2]

1 - Ecole Nationale de la Statistique et de l'Analyse de l'Information [Bruz] (France), 2 - Institut national de la statistique et des études économiques (INSEE) (France)

On every French election night since 1965, the result is announced at 8 p.m., the moment the last polling stations close, as required by law. These early estimates are produced by private polling institutes whose procedures are not published and rest, by their own account, on decades of political and historical expertise rather than on a documented statistical method. The question we raise is whether that black box in a moment of genuine democratic significance is necessary, or whether open data and fully reproducible code can do the same job in the open. We reproduce the process of sampling and estimation ex post for the 2022 presidential first round, within a strictly statistical framework and entirely in R, using only publicly available data: the official first-round results for 2017 and 2022, Insee's 2023 open-data file linking anonymised elector addresses to their polling station, the last Filosofi gridded socio-fiscal data, and the Insee density grid. A central step is the geographic data linkage: matching elector addresses, socio-fiscal grid squares and polling-station boundaries, which Insee's recent MapvotR package and R's spatial and data-linkage ecosystem make remarkably straightforward. Combining these sources lets us characterise the roughly 65,000 metropolitan polling stations and build a sampling frame. We then simulate election night with a two-stage design: a few hundred polling stations are drawn, and up to a hundred ballots of each are counted. Six first-stage designs are compared over 1,000 Monte Carlo replications, ranging from simple random sampling to stratified and balanced sampling. Our results show that naive sampling is structurally biased, because only stations closing early are eligible while late-closing urban stations are excluded. Well-chosen academic designs, combined with calibration on auxiliary variables, almost entirely correct this bias. Crucially, accuracy comes from the quality of the design, not merely from a larger sample. In this purely theoretical setting, our open methods appear at least as reliable as the institutes' published 8 p.m. estimates. Our aim is not to discredit professionals, but to show that a method anyone can read, audit and rerun can match expertise-driven estimation. This raises a broader question for the official statistics community: could national statistical institutes, as Insee has begun to do, publish regularly the polling-station boundaries together with fine-grained socio-economic data? Doing so would turn election-night estimation into a public good, letting results be known quickly, reliably, and through trustworthy methods. A profoundly democratic act.

Facilitating R-Based Statistical Validation Through EZR: A Case Study in Melanoma High-Frequency Ultrasound Assessment

Kyrylo Chalyy[1], Roman Hryshchenko[2], Olga Bogomolets[3], Viktor Stepanenko[2]

1 - Department of Natural Sciences and Information Technologies, Shupyk National Healthcare University of Ukraine, Kyiv, Ukraine (Ukraine), 2 - Department of Dermatology and Venereology, Bogomolets National Medical University, Kyiv, Ukraine (Ukraine), 3 - Dr. Bogomolets Institute of Dermatology and Cosmetology, Kyiv, Ukraine (Ukraine)

Objective: To demonstrate how EZR (Easy R), a graphical user interface built on the R statistical environment, can facilitate the application of advanced statistical methods by clinical researchers without programming experience. High-frequency ultrasound (HFUS) assessment of cutaneous melanoma was used as a practical case study to illustrate the implementation of an R-based validation workflow. Background: R provides a powerful and flexible framework for statistical analysis; however, its adoption in clinical research may be limited by the need for programming skills. EZR extends the capabilities of R through a menu-driven interface designed specifically for medical statistics. This approach enables clinicians and biomedical researchers to apply validated R-based methods in routine practice while promoting transparency, reproducibility, and wider adoption of statistical standards. Materials and Methods: Data from 90 patients with dermoscopically suspected cutaneous melanoma were analysed using EZR. Twenty-two patients underwent HFUS examination using the DUB Skinscanner (22 MHz), while 68 examinations were performed using the General Electric LOGIQ E R8 system equipped with a 10–22 MHz linear probe. Histopathological examination (HPE) served as the reference standard. The EZR workflow included descriptive analyses, estimation of discrepancy rates with 95% confidence intervals, comparison of diagnostic performance using odds ratios, Spearman rank correlation analysis, and linear regression modelling to evaluate agreement between ultrasound-derived and histopathological measurements of tumour invasion depth. EZR provided access to R-implemented procedures through a menu-driven interface without requiring users to write statistical code. Results: Discrepancies between ultrasound and HPE findings were observed in 63.6% (95% CI: 42.9–80.3%) of DUB examinations and 14.7% (95% CI: 8.2–25.0%) of GE examinations. The odds of disagreement were 10.1 times higher for the DUB system (OR 95% CI: 3.4–30.4). For the GE system, EZR-based analyses demonstrated a very strong association between ultrasound and histopathological measurements (Spearman's $\rho = 0.929$, $p < 0.001$). Linear regression analysis further confirmed good agreement between the two methods ($R^2 = 0.874$). Contribution and Conclusions: This case study illustrates how EZR can serve as an effective gateway to the R ecosystem for clinicians and other non-programming researchers. By enabling access to robust R-based statistical procedures through an intuitive graphical interface, EZR lowers barriers to the adoption of reproducible analytical practices. The experience presented here demonstrates that advanced statistical validation can be integrated into routine biomedical research without dedicated programming expertise and may provide a model for other applied domains seeking to expand the use of R.

From Data to Publication : An Automated R-Based Framework for Consumer Price Index Reporting

Zineb Aoulad Hammou Achou[1], Zakariae Halhal[1]

1 - High Commission for Planning (HCP), Tangier–Tetouan–Al Hoceima Regional Directorate (Morocco)

The HAKAMA 2 project, led by the High Commission for Planning (HCP), focuses on strengthening public governance through improved statistical production, survey systems, and institutional communication. Within this project, the Regional Directorate of Tangier-Tetouan-Al Hoceima (DRTTA) focused on improving the processing and use of Consumer Price Index (CPI) data for official reports. The collaboration with INSEE led to a revision in project objectives, emphasizing the integration of CPI data into the regional BDS-Tanger database and the generation of regional CPI reports. R-scripts were developed for automating the data integration process. Subsequently, efforts expanded to automate the analytical report generation, facilitating simultaneous production of the BDS-Tanger input template and the CPI report.

Further microdata anonymisation tools in R

Marco Caputi[1]

1 - Istituto Nazionale di Statistica (Italy)

National Statistical Institutes use a wide variety of dissemination channels to make statistical data accessible to various types of users and to maximise the value of public information assets. These include, for example, interactive online databases for aggregate data, maps, graphs, customised processing services, Data Centres, and anonymised microdata files. In particular, anonymised microdata files allow users to perform statistical analyses that would not be possible with aggregate data, while maintaining adequate guarantees regarding statistical confidentiality through the application of Statistical Disclosure Control methods. These files can be accessed both by researchers belonging to recognised institutions (Scientific Use Files) and by non-specialist users for study or work purposes (Public Use Files). Tools such as mu-Argus and the R package sdcMicro are specifically designed for the production of anonymised microdata files. The sdcMicro package, in particular, offers a wide variety of Statistical Disclosure Control measures and methods, alongside the ability to operate entirely within the R environment. The purpose of this work is to provide additional features designed both to expand the options available within the anonymisation process and to meet specific needs related to the criteria adopted by the Italian National Institute of Statistics. These features relate to the disclosure risk measurement phase as well as the data protection phase of the process; furthermore, the fact that these additional tools operate entirely within the R environment allows them to be used alongside the sdcMicro package. In particular, regarding the risk assessment phase, the flexibility in managing key variable sets to model the disclosure scenario is enhanced, and further risk measures, such as t-closeness, are made available. Finally, concerning the data protection phase, additional perturbative methods are introduced.

How many R packages in CRAN do you know?

Patrice Kiener[1], Christophe Dutang[2]

1 - InModelia (5 rue Malebranche, 75005 Paris France), 2 - École nationale supérieure d’informatique et de mathématiques appliquées (France)

CRAN is an evolving object with approximatively 6 new packages per day, and some of them are very useful! But, without any help, you probably do not know them. The easy to use RWsearch package gives a full control to download the most recent version of CRAN database, compare it with a previous version, identify the new packages, search for keywords in the package name, title, description, author or maintainer and present the results in a concise document, either in txt, pdf or html format. At a higher level, the package allows to quickly create a reliable literature review for a given topic without asking to generative AI. We have been extensively using RWsearch since 2018 to identify the packages related to distributions and the Distribution task view is now ranked number 2 by the number of referenced packages. RWsearch proves useful every day.

ILSAmerge and ILSAstats: Two new R packages for international large-scale assessments

Andrés Christiansen[1]

1 - IEA Hamburg (Germany)

International Large-Scale Assessments (ILSA) in education, such as TIMSS, PIRLS, ICILS, and ICCS, provide valuable insights into educational systems worldwide. However, analyzing ILSA data poses unique challenges, including how the data is structured and the use of plausible values and replicate weights for estimating the standard errors. We introduce two R packages to address these challenges: ILSAmerge and ILSAstats. ILSAmerge simplifies the process of downloading and merging ILSA data across multiple countries into a single dataset. By automatizing the integration of country-specific files, this package eliminates the need for manual data handling or the requirement of multiple steps. While testing against competing packages, ILSAmerge was the only CRAN package compatible with all studies of the IEA (International Association for the Evaluation of Educational Achievement). ILSAstats is designed to analyze ILSA data. The package calculates means, proportions, quantiles, correlations, and single-level and multi-level regression models while accounting for plausible values and replicate weights, adhering to each ILSA's methodological guidelines. While testing against competing packages, ILSAstats was the only CRAN package able to reproduce all official results of all studies of the IEA. Together, ILSAmerge and ILSAstats provide a comprehensive toolkit for researchers and analysts working with ILSA data, streamlining data preparation and statistical analysis

Manual Rework Volume Monitoring and Forecasting Tool

Yves Coltier[1]

1 - Institut national de la statistique et des études économiques (INSEE) (France)

The monitoring and forecasting tool for manual rework volume is developed in RShiny and is still in the experimental phase. It provides a user-friendly interface and aims to simplify the interpretation of various metrics used to assess the performance of a supervised learning model. Additionally, it facilitates trade-offs between annotation quality for a survey and the manageable volume of manual rework. The tool allows users to import data (in Parquet or CSV format) containing at least the following: - The target (manually annotated label), - The best model's prediction, - The associated confidence score, - The weight of each observation (a default weight of 1 is applied if unspecified). After mapping the file's columns to the expected variables, the data is processed and saved locally with an identifier code for the survey. Users can delete a survey or clear the cache to optimize performance. Following the import, the tool generates: - A graph of the correct automatic coding rate; - A histogram comparing initial codes with prédictions; - A summary table of discrepancies. It also displays: - A confusion matrix to visualize correct and incorrect predictions by class at an aggregated level; - A table of the top 10 most frequent misclassifications. Finally, the tool assists in selecting an optimal confidence threshold, below which the model's predictions are considered unreliable, and the text must be sent for manual review. To support this decision, the tool estimates: - The expected overall quality of the survey; - The volume of manual rework required, based on the confidence threshold, while integrating parameters such as the human error rate or the use of term lists enabling direct coding. The project is available on GitHub: https://github.com/ycoltier/monitoring-ml.

Mapping the Gendered Impact of Economic Shocks: A Spatial Analysis of Mancession vs. Shecession across European Regions

Andreea-Denisa Trascan[1,2], Camelia-Maria Dan[1]

1 - The Bucharest University of Economic Studies / Academia de Studii Economice din Bucureşti (Romania), 2 - National Scientific Research Institute for Labor and Social Protection (Romania)

Economic crises do not affect labor markets uniformly. Their consequences are shaped by the sectoral structure of regional economies and by the unequal distribution of women and men across economic activities. The 2008 global financial crisis disproportionately affected male-dominated sectors such as construction and manufacturing, giving rise to what has been described as a mancession. More than a decade later, the COVID-19 pandemic produced a markedly different pattern, with severe disruptions concentrated in service, retail, hospitality, and tourism sectors where women are more heavily represented, generating a shecession. This paper examines how these two crises generated distinct spatial patterns of labor market vulnerability across European regions. Using Eurostat data at the NUTS-2 level for a two-decade period spanning the 2008 and 2020 crises, the analysis combines information on gender-specific employment rates, educational attainment, and sectoral economic structures to investigate the geography of crisis-related inequalities. These spatial dynamics are explored through bivariate choropleth maps to identify variations in the intensity and distribution of gendered employment losses, examining how regional economic specialization shapes exposure to different macroeconomic shocks. To complement the visual analysis, the paper employs a Difference-in-Differences framework using panel data techniques. These models are designed to assess the extent to which educational attainment and sectoral dependency influence regional labor market resilience during periods of economic disruption. More specifically, the analysis tests whether regions characterized by a strong concentration of construction and manufacturing activities were more vulnerable to the employment losses associated with the 2008 financial crisis, while regions with a greater reliance on consumer-facing services experienced stronger adverse effects on female employment during the COVID-19 pandemic. By combining geospatial visualization with causal inference methods, this study provides new evidence on the evolving geography of gender inequality. More broadly, it demonstrates how analytical tools within the R ecosystem can support the integration of official statistics, spatial analysis, and policy-relevant labor market research.

R you useful for maintaining information systems ?

Alice Mainguené[1]

1 - Institut national de la statistique et des études économiques (INSEE) (France)

R is extremely useful for maintaining the information systems, such as France’s business directory, Sirene. First, it enables us to work on very large databases. Since the directory was established in the 1970s, it now contains records for over 31 million companies, including their histories and administrative procedures. With R and the RPostgres package, we can easily analyse the data, extract information for our partners and detect issues within the directory. Second, R enables us to make bulk calls to web services, using the httr package. The directory relies on APIs and web services that we can call individually using the Swagger tool. However, as statisticians, we know that large-scale testing is often necessary to detect low-frequency phenomena, such as errors that occur only in specific types of administrative procedures. With R, we can remotely and efficiently call these web services in bulk. For example, last year, we tested 45,000 administrative procedures in the conformity web service with R. This allowed us to validate a new type of XML file from one of our partners. Finally, R simplifies the mass creation of JSON files with the jsonlite package. These files serve as input for one of our batch processes, helping us correct data in the directory. For example, they helped us correct 12,000 postal codes in the adresses of companies located in the overseas territories.

RISQ: Representativity Indicators for Survey Quality

Reijer Idema[1]

1 - Statistics Netherlands (Netherlands)

Missing data due to nonresponse is a threat to the quality of statistics based on surveys. In the EU funded project Representativity Indicators for Survey Quality (RISQ), measures were developed for the degree to which respondents resemble the population. These measures are actively used to assess and improve survey designs. In this presentation, we present a new R-package that implements representativity indicators for survey quality.

Sampling Designs and Sample Quality Control: Using Quarto Documents in Agricultural Surveys

Michael Levi-Valensin[1]

1 - Ministère de l'Alimentation, de l'Agriculture et de la Pêche [Paris, France] (France)

Designing a sample survey involves a large number of steps that must be carefully followed: defining the target population, stratification, allocation, sample selection, variance estimation, and so on. During the testing phases, it is not always easy to quickly access all of these characteristics while avoiding the risk of errors. Generating a Quarto document from the saved information makes it possible not only to visualize key information in the form of tables, charts, or maps, but also—and above all—to check consistency: selected sample sizes, compliance with precision requirements, estimates of aggregates, and so forth. We will present examples of HTML reports produced for the 2025 PKGC survey on agricultural practices in arable crop production and for the latest 2026 ESEA survey on the structure of agricultural holdings.

Semi-automated editing of large volumes of foreign trade data for the environmental Economy-Wide Material Flow Account

Gorka Campandegui García[1], Julia Maestro López[1]

1 - Instituto Nacional de Estadística (INE) - Spanish National Statistical Institute (Spain)

The compilation of Economy-Wide Material Flow Accounts (EW-MFA) requires accurate estimates of the weight associated with imports and exports of goods. Traditionally, these estimates have been based on aggregated foreign trade information, limiting the ability to detect errors and identify transactions that do not represent physical flows of materials. To improve the quality of these estimates, the Spanish National Statistical Institute (INE) has developed a procedure based on administrative records from the Spanish Tax Agency at the highest level of disaggregation available. The implemented solution is designed to process approximately 100 million records per year through a modular architecture developed in R, combining widely used packages from the statistical community with tools developed in-house. The methodology integrates the generation of descriptive information, imputation of missing values, automatic outlier detection, and expert review of records identified as potentially problematic. The use of microdata enables the detection of anomalies that remain hidden in traditional aggregated data and the use of auxiliary variables available in customs records to identify errors and inconsistencies more accurately. In addition, the high level of detail facilitates the identification of quasi-transit operations and other transactions that do not involve a physical inflow or outflow of materials into or from the national territory, preventing their inclusion in EW-MFA aggregates. The process has been designed as a semi-automated editing workflow in which detection parameters are calibrated to reduce the number of cases requiring manual review and to focus expert effort on records with the greatest potential impact on aggregate results. The results show a substantial reduction in the manual review burden with improved identification of inconsistencies and transactions that should be excluded from material flow accounting. This work demonstrates how the combination of detailed administrative records, statistical imputation and editing techniques, and large-scale data processing tools in R can simultaneously improve the quality, efficiency, and reproducibility of statistical production processes in the field of environmental statistics.

SmartSurveyAgent: A New Approach to Survey Data Collection in Official Statistics

Simona Cafieri[1]

1 - Italian National Institute of Statistics (Italy)

Conversational artificial intelligence is increasingly regarded as a promising tool for data collection in statistical surveys, yet its methodological implications in the field of official statistics remain largely unexplored. This paper presents SmartSurveyAgent, a self-administered conversational data collection mode that integrates conversation-based data collection within the framework of the Total Survey Error (TSE) paradigm. The approach treats conversation not merely as a technological interface, but as an observable and measurable component of the survey design, enabling systematic analysis of interaction dynamics through structured paradata. These indicators are linked to key dimensions of TSE, including measurement error, item nonresponse, and response burden. Implemented in R and interfaced through a bot API, the system guides respondents through a conversational response process, manages user sessions, and records both survey responses and interaction metadata in compliance with institutional standards. Fully integrable into statistical survey production systems and interoperable with traditional data collection modes such as CAWI, CATI, and CAPI, SmartSurveyAgent extends the possibilities of mixed-mode designs in official statistics by introducing a formally defined and methodologically observable conversational approach.

Template and Function: A Tabular Configuration Pattern for Large Scale Operations on Files with Varying Schemas in Official Statistics

Daniel Rakotomalala[1]

1 - Direction de l’évaluation, de la prospective et de la performance (France)

DEPP-B1 is the unit at DEPP (the French Ministry of Education’s statistical service) supporting public policy evaluation and education research. In this role, we regularly execute operations that apply uniform logic to collections of files whose schemas are not consistent across units: files originate from several producing offices with different internal conventions. Typical operations include format conversion, anonymization for researcher access, conditional variable selection or observation filtering, and variable and modality harmonization. Counts often reach the hundreds. Variable names drift across files and schemas evolve across years. Writing one ad-hoc script per file does not scale; writing a single generic pipeline cannot accommodate the structural variation. We describe a pattern, Template-and-Function, that emerged from this production work. The Template is a tabular configuration whose rows are units (files, variables, modalities) and whose columns are parameters (for instance, variables to anonymize and output paths). The Function operates on a single row’s parameters and produces a side effect: a written file, a transformation, a check. Iteration over the Template is mechanical. In R terms, it amounts to designing large-scale operations around a data frame (df) and a function (f) passed into expressions such as apply(df, 1, f, …) or purrr::pmap_dfr(df, f, …). Crucially, the Template is itself data: readable, writable, editable, version-controlled, and archivable as audit evidence of the operation. Around the iteration loop, one can control execution behaviors such as error robustness via tryCatch (a failed row does not block the whole operation), logging via a success column appended to the Template, and optional parallelization. We illustrate the pattern on four production cases at DEPP-B1: (1) file conversion; (2) anonymization and pseudonymization of individual-level datasets for legally compliant access by researchers; (3) observation filtering and variable selection across files where the relevant variable names differ across units; (4) variable and modality harmonization across multiple waves of data, typically for longitudinal cohort construction. Each case applies the same pattern with different Template schemas and Function bodies. We discuss scalability, fault isolation, testability on subsets, and the auditable lineage that a persisted Template provides. The pattern is not a new method. It builds on standard R codes for iteration (apply, purrr::pmap) and error handling (tryCatch). What it adds is the discipline of externalizing the iteration’s configuration outside the R session: the Template persists, is auditable, and is editable.

The Effects of Skills on Gender Wage Inequality in the Digital Economy

Gianina-Maria Petrașcu[1], Andreea-Denisa Trascan[1,2]

1 - The Bucharest University of Economic Studies / Academia de Studii Economice din Bucureşti (Romania), 2 - National Scientific Research Institute for Labor and Social Protection (Romania)

Understanding the sources of gender pay inequality remains one of the key challenges facing contemporary labour markets. Using data from the second cycle of the Programme for the International Assessment of Adult Competencies (PIAAC 2023), this study examines how multidimensional skills used at work are related to pay differences between men and women across eight European countries. Specifically, the analysis uses the OECD's eight official indicators of skills use at work: reading, writing, numeracy, ICT skills, problem solving, task discretion, learning at work, and influencing skills. Gender-specific Mincerian wage equations were estimated using OLS with country fixed effects, controlling for age, education, and literacy and numeracy proficiency. A Blinder-Oaxaca decomposition was then applied to assess the relative importance of skill endowments and returns to skills in explaining gender pay gaps.The comparative cross-country framework provides insights into how the labour market valuation of skills varies across institutional contexts. The results indicate that all eight dimensions of workplace skill use are positively associated with wages for both men and women. Returns to reading, writing, learning at work, and influencing skills are higher for women, while returns to numeracy, ICT skills, and task discretion are slightly higher for men. The decomposition results show that differences in skill endowments contribute little to the overall gender pay gap, as women report similar or higher levels of workplace skill use across most dimensions. Consequently, differences in returns to skills account for the majority of the observed pay disparities. However, substantial cross-country heterogeneity is evident, with the relative importance of endowment and return effects varying across institutional settings. Overall, the findings suggest that addressing gender pay inequality requires attention not only to skills development but also to how skills are rewarded in the labour market.

AI-Assisted Scientific Literature Review with LitRefineR Using OpenAlex

Ioana Barbulescu[1], Ana-Maria Ciuhu[2], Valentina Vasile[3], Andrei Suciu[4], Razvan Vasile[5]

1 - Romanian Academy, School of Advanced Studies of the Romanian Academy, Doctoral School of Economic Sciences, National Institute for Economic Research "Costin C. Kirițescu", Institute of National Economy, Bucharest, Romania (Romania), 2 - Institute of National Economy, Romanian Academy / Romanian Institute of Statistics (Romania), 3 - Institute of National Economy, Romanian Academy (Romania), 4 - Specialist software architecture, Independent scholar (Romania), 5 - The Bucharest University of Economic Studies (Romania)

Compound literature searches need to identify publications that address multiple concepts together. We evaluated lexical and semantic retrieval for the query "AI AND agriculture" using a stratified corpus of 200k OpenAlex publications. BM25 served as the lexical baseline. The BGE facet-min method scored each concept separately and ranked publications by their lower similarity score. Relevance judgments distinguished publications that substantively addressed the topic from incidental mentions and unrelated records. In the ranking restricted to human-judged publications, BGE facet-min achieved an average precision of 0.636 and a precision at ten of 0.70, compared with 0.222 and 0.10 for disjunctive BM25. In a separate comparison of the exported BM25-AND and BGE facet-min result sets, the two methods shared 56 of 269 distinct publications (Jaccard similarity 0.208). Relevance screening retained 102 publications for scientometric analysis in R. We used human judgments where available and Opus labels for the remaining publications. The analysis examined publication years, field classifications and the frequency of terms in titles. The proposed R package will automate searches with two or three concepts across CORE, PubMed, arXiv, Semantic Scholar and OpenAlex. It will deduplicate records across sources and use AI to assess which concepts each publication addresses and whether it treats them together. Future applications include "AI AND official statistics" and "AI AND labour market".

Use of R Packages for Cartography: The Case of Road Traffic Accident Statistics in Luxembourg

Guillaume Osier[1], Alexi Dere[1], Lucile Bodson[1]

1 - Institut national de la statistique et des études économiques du Grand-Duché de Luxembourg (STATEC) (Luxembourg)

Luxembourg's National Institute of Statistics and Economic Studies (STATEC) publishes annual indicators on road traffic accidents occurring within the country. These statistics are based on police reports completed whenever a fatal accident or an accident involving personal injury takes place within Luxembourg's territory. The resulting dataset provides detailed microdata at both the accident and victim levels. In addition to information on the circumstances and consequences of each accident, the data include the geographic coordinates (longitude and latitude) of the accident location, enabling detailed spatial analyses. In this presentation, we present several use cases illustrating how R packages—primarily sf and leaflet—can be employed for the cartographic visualization of road traffic accidents. We show that these tools are not only valuable for producing interactive and informative maps, but also constitute important tools for statistical analysis of road accidents. In particular, they facilitate spatial analyses, such as identifying and counting accidents within specific geographic areas or administrative boundaries.

VIM::vimpute: A Unified Imputation Engine for Flexible Missing Data Workflows in R

Alexander Kowarik[1], Eileen Vattheuer[1], Matthias Templ[2], Johannes Gussenbauer[1], Nina Niederhametner[1]

1 - Statistics Austria (Austria), 2 - Fachhochschule Nordwestschweiz FHNW (Switzerland)

We present vimpute(), a unified imputation engine for multivariate missing data in the R package VIM. This lightning talk provides an update to the presentation given at uRos2025, highlighting recent developments and the consolidation of several imputation approaches around a common computational core. The function allows multiple variables to be imputed in a single workflow, using variable-specific models tailored to the type and structure of each target variable. It builds on the mlr3 ecosystem and therefore benefits from a modern and extensible modelling framework, including Task, Learner and tuning objects, a consistent interface, and built-in support for parallelisation. The method performs automatic iterative imputation with convergence checks and can optionally include hyperparameter tuning. Additional options include predictive mean matching, bootstrap-based uncertainty, and multiple imputation with (m > 1). Existing VIM functions such as kNN(), rangerImpute(), xgboostImpute() and regressionImp() already rely internally on vimpute(), making it the common computational core behind several established imputation approaches.

Why Do Some Counties Perform Better? Educational Hubs, High-Tech Development and Skills Mismatch in Romania

Cristina Boboc[1,2], Gianina Maria Petrașcu[1]

1 - The Bucharest University of Economic Studies, Romania (Romania), 2 - Institute of National Economy (Romania)

The increasing importance of skills as a determinant of productivity, competitiveness, and inclusive economic growth has intensified concerns regarding the alignment between labour market demand and workforce capabilities. Across the European Union, skills mismatch has become a central policy issue, reflecting the growing divergence between the qualifications and competencies possessed by workers and those required by employers. In Romania, these challenges are amplified by demographic decline, labour migration, regional disparities, technological change, and persistent structural imbalances in the labour market. This paper examines skills mismatch in Romania and its territorial disparities across counties. Drawing on the conceptual framework developed by Cedefop, the analysis considers various forms of mismatch, including overeducation, undereducation, overskilling, underskilling, and vertical mismatch. To assess Romania's relative position, the study employs the European Skills Index (ESI) and benchmarks county-level labour market structures against the five best-performing EU Member States. A structural skills mismatch indicator is constructed by measuring the distance between the occupational and sectoral composition of Romanian counties and the benchmark profile derived from leading European performers. The empirical analysis is conducted in R and combines data processing, indicator construction, benchmarking techniques, and spatial visualization to assess regional patterns of skills mismatch. The results reveal considerable heterogeneity across Romanian counties, indicating that skills mismatch is a geographically differentiated phenomenon. Counties characterized by less diversified economic structures and lower levels of development tend to display higher mismatch scores and weaker labour market outcomes. At the same time, the ongoing digital and green transitions are increasing demand for new competencies, raising the risk of skills obsolescence and widening existing territorial disparities. The findings indicate that Romania continues to face significant difficulties in aligning workforce competencies with labour market requirements. Persistent skill shortages coexist with relatively low employment rates, while educational attainment does not always translate into appropriate labour market outcomes. The analysis further suggests that regional labour market structures play a critical role in shaping the intensity of skills mismatch across counties. The paper argues that reducing skills mismatch requires a coordinated policy approach that strengthens the links between education, training, and labour market demand. By providing a county-level assessment based on a European benchmark framework, the study offers valuable evidence for place-based policy design and contributes to the broader debate on skills governance, regional development, and labour market resilience in the context of digital and demographic transitions.

{IssueTrackeR}: Handling Tickets in R

Tanguy Barthelemy[1]

1 - INSEE (France)

Version control for projects using Git on platforms such as GitHub or GitLab has become established as a best practice and is now a standard in software development. In statistical projects, the automation of production pipelines using languages such as R or Python also increasingly relies on these tools, particularly their project management systems: creating tickets, using labels, and defining intermediate steps to structure the development process. The R package **{IssueTrackeR}** (@issuetracker-r-package) allows you to retrieve, organize, and analyze GitHub issues directly in R. It offers tools that make it easier to explore them, for example, to track the progress of resolving issues reported by users or to perform advanced searches in the issue history using keywords. **Keywords (3–5)**: Package - GitHub - ticket management - GitHub API ## Development ### Features **{IssueTrackeR}** allows you to centralize tickets from one or more GitHub repositories (public or private, belonging to different users or organizations) into a single database. Once imported, this information can be filtered to: - Search for questions and answers with a common theme (to see if someone has already asked) - Select the most urgent, high-priority, or unanswered topics Tickets will also be analyzed to track and visualize trends in the number of open and closed tickets over time. ### Structure Tickets, labels, and milestones are represented (using appropriate S3 classes and methods) as `data.frames` rather than `lists` in order to take advantage of R's array processing capabilities. The package is available on GitHub, R-Universe, and on CRAN. ### API GitHub The package uses the GitHub API via the **{gh}** package to retrieve data and associated metadata. ### Coming soon… Currently, it is only possible to retrieve information from GitHub, but eventually, the package may be extended to other project management platforms, notably GitLab.

{loomer} — A high-performance auxiliary R package for analyzing the Danish Lifelines dataset

Ilya Kashnitsky[1]

1 - Statistics Denmark (Denmark)

The Danish Lifelines dataset is a comprehensive, event-driven administrative register tracking individuals' presence in the Danish population from 1986 to 2025. Each record represents a contiguous period of residence, segmented by birth, death, immigration, and emigration events, enabling rich longitudinal demographic research across 9.5 million of individuals who lived, died, moved in and out of Denmark in the last four decades. However, transforming raw date-based records into analysis-ready datasets at this scale demands substantial computational resources and specialized data-wrangling expertise. We introduce {loomer}, an auxiliary R package purpose-built to interface with the Lifelines dataset. Leveraging the fastverse ecosystem — particularly collapse and kit — {loomer} delivers aggregations and data manipulations orders of magnitude faster than base R or tidyverse equivalents. Core features include: a built-in codebook with instant decoding of integer-encoded variables (enabling 4× faster data reads); optimized cross-sectional population snapshot functions; cohort construction utilities for targeting sub-populations by demographic trajectory; and dedicated migration dynamics tools. To address strict data access restrictions that hinder reproducibility, {loomer} ships with a 1% anonymized synthetic population that mirrors the full dataset's structure and event frequencies, enabling local development and testing prior to secure server deployment. Together, these features significantly lower the barrier to entry for demographic research with Lifelines.

EMOS Presentations

AI, Robotization and Perceived Occupational Displacement across European regions - an R based Spatial Econometrics Analysis

Vlad-Andrei Petre[1], Smaranda Cimpoeru[1]

1 - Bucharest University of Economic Studies (Romania)

Artificial intelligence and robotization are generating major occupational transformations in labour markets, affecting both work organization and individuals' perceptions of job displacement risks. While existing research has mainly focused on the objective effects of digital technologies on employment, productivity and occupational restructuring, the perceptual dimension of job displacement remains less explored, especially at regional level. This study analyses the determinants of perceived occupational displacement associated with AI and robotization across 228 NUTS 2 regions of the European Union. The main objective is to identify the perception-based and socioeconomic factors associated with perceived job displacement, while accounting for spatial autocorrelation and regional heterogeneity. The analysis also evaluates whether local spatial regression models provide a more detailed understanding of the phenomenon compared with classical and global spatial specifications. The analysis combines perception indicators derived from Eurobarometer surveys 101.3 and 101.4 with official regional statistics from Eurostat. The dependent variable measures the share of respondents who perceive AI and robotization as technologies that may replace jobs. The explanatory variables include perceived digital skills for future work, the perceived increase in the pace of work tasks due to AI and robotization, positive perceptions regarding the use of AI and robots in the workplace, negative expectations regarding the personal job situation, negative perceptions of the national employment situation, the unemployment rate, tertiary education and employment in high-technology and knowledge-intensive services. The methodological framework combines exploratory spatial data analysis with spatial econometric models. The analysis is implemented entirely in R, using dedicated packages for spatial analysis and spatial modelling, including sf, spdep, spatialreg and GWmodel. Spatial autocorrelation is tested using Moran's I statistic, followed by a classical OLS specification and a broad set of global and local spatial regression models, including SAR, SEM, SDM, SLX, SDEM, SARAR, GWR and MGWR. The results confirm significant spatial autocorrelation, indicating that perceived job displacement is not randomly distributed across European regions. Among the tested models, MGWR provides the strongest explanatory performance, showing that the relationships between explanatory variables and perceived occupational displacement vary across regions and operate at different spatial scales. The findings indicate that unemployment, tertiary education, perceived digital skills and the perceived increase in the pace of work tasks generated by AI and robotization are positively associated with perceived occupational displacement in many regions. In contrast, employment in high-technology and knowledge-intensive services and positive perceptions regarding the use of AI and robots in the workplace are associated with lower perceived displacement risk. By revealing geographically differentiated patterns, the study supports territorially differentiated policies aimed at workforce adaptation, reskilling and the valorisation of human capabilities that remain difficult to replace by AI. At the same time, it emphasizes the usefulness of an R-based spatial modelling framework to analyze regional socioeconomic phenomena characterized by spatial dependence and spatial heterogeneity.

An R Framework for AI-Assisted Questionnaires in Official Statistics: An Application to Financial Literacy Surveys

Valeria Rossi[1]

1 - Università degli studi della Campania """"Luigi Vanvitelli""""" = University of the Study of Campania Luigi Vanvitelli (Italy)

This research proposes a methodological and computational framework for AI-assisted mixed-mode survey data collection in official statistics, with an application to financial literacy surveys. The study investigates how alternative survey administration modes affect response behaviour, engagement, nonresponse mechanisms, and data quality by comparing AI chatbot-based data collection with traditional modes, including CAWI, CATI, and CAPI. The analysis is conducted within a reproducible R-based framework for evaluating mode effects in mixed-mode surveys. The empirical application is based on a pilot survey administered to university students. The questionnaire combines a simplified set of ISTAT census variables with an adapted MiFID II framework to collect socioeconomic covariates, financial knowledge, investment objectives, and risk attitudes. Respondent profiles are identified through clustering techniques applied to risk attitudes and socioeconomic covariates collected via the survey. Representative investment portfolios generated in R are incorporated into the questionnaire as an elicitation tool to analyse preferences and decision-making behaviour. The pilot questionnaire covers three thematic dimensions: socio-demographic characteristics, financial literacy and investment behaviour, and survey-mode preferences. Clustering is performed on rank-aggregation composite indicators computed separately within each dimension, rather than on PCA component scores. This choice ensures greater consistency, robustness, and interpretability, as each composite indicator maps directly onto one theoretical dimension. In contrast, PCA scores are linear combinations that can blend variables across dimensions, making them harder to interpret and less closely tied to the three-dimensional theoretical framework. From a methodological perspective, the research integrates design-based inference, IPW, G-computation, and doubly robust estimators to account for selection bias and mode effects in mixed-mode surveys. A novelty of the research is the development of a unified causal inference framework that integrates mixed-mode survey methodology, AI-assisted data collection, investor profiling, and behavioural finance within a reproducible R-based analytical architecture. Unlike the existing literature, which typically examines these dimensions separately, the project investigates their joint effects.

Developing a territorial One Health indicator using R

Carmen Limone Fenicia[1]

1 - Università degli studi della Campania """"Luigi Vanvitelli""""" = University of the Study of Campania Luigi Vanvitelli (Italy)

The project aims to develop a multidimensional territorial composite indicator within the One Health framework, based on a set of interrelated dimensions, including health, lifestyles, risk behaviours, socioeconomic and living conditions, consumption patterns, access to services and subjective well-being. The final objective is their combination into a synthetic but interpretable territorial One Health index intended to support the analysis of geographical and population-level inequalities, including differences across Italian regions, macro-areas and population groups, thereby identifying territories with specific combinations of vulnerabilities and strengths. The research is structured in two main analytical stages. In the first stage, composite indicators will be constructed for each domain, potentially organised into theoretically defined subdomains. Elementary variables will be selected according to their conceptual relevance, harmonised and standardized. Correlation analysis and dimensionality-reduction methods will be used to examine the empirical relationships among variables, assess whether the proposed subdomains are supported by the data and inform possible weighting and aggregation strategies. Different approaches, including equal weighting and PCA-based weighting, will be compared in terms of interpretability, stability, sensitivity and robustness. The domain-specific indicators will subsequently be integrated into the overall territorial One Health index, while preserving a sufficiently transparent structure to allow the contribution of individual domains to be examined. The second analytical stage will investigate the relationships between the domain indicators, subdomains and selected health outcomes. Where an appropriate outcome variable is available, supervised regularisation methods such as Ridge regression, LASSO and Elastic Net may be applied to identify the components most strongly associated with or relevant for predicting, the selected outcomes. These methods are therefore intended as methodological support for identifying the most informative relationships among the different components of the framework. By integrating these dimensions, the resulting index may provide an applied tool for analysing territorial inequalities, comparing health-related profiles and identifying areas in which multiple disadvantages accumulate. It may also help distinguish whether similar health outcomes are associated with different combinations of behavioural, social, economic or environmental determinants across territories and population groups. A possible future development will involve scenario-based analyses after the identification of territorial clusters or profiles. Such analyses could explore how changes in lifestyle factors or specific domain indicators might affect the overall One Health index. This extension could contribute to the evaluation of prevention strategies and provide evidence to support the design of territorially targeted policies aimed at reducing health inequalities.

R-Based Visualization of Continuous-Time Markov Migration Models

Anna Boukouvala[1], George Tsaklidis[1]

1 - Aristotle University of Thessaloniki (Greece)

Migration statistics play a key role in understanding population dynamics and regional transitions. Continuous-time Markov models provide a flexible and widely used framework for representing migration processes in official statistics; however, their interpretation can become challenging as system complexity increases. This work presents an R-based approach for the analysis of continuous-time Markov migration models, grounded in a mathematical formulation of migration dynamics within the probability simplex and providing a structured representation of system evolution over time. Beyond standard stochastic modelling, it introduces a geometric perspective that supports a more intuitive interpretation of migration dynamics in terms of flows on the simplex. The implementation in R is designed to support the analysis of migration systems and enables flexible exploratory analysis and visual inspection of model outputs. Overall, the proposed framework highlights how R can be used not only for the estimation and simulation of continuous-time Markov migration models, but also for enhancing their interpretability in the context of official statistics.

Subsampling Non-Response Follow-Up in the Canadian Housing Survey: Comparing the Historical Hansen-Hurwitz Method and Non-Probability Estimation

Thomas Guinhut[1]

1 - Guinhut Thomas (France)

Social surveys worldwide face declining response rates and rising collection costs; interviewer follow-up (CATI or CAPI), while effective at raising participation, remains the most expensive collection stage. In recent cycles of the Canadian Housing Survey (CHS), a Statistics Canada household survey, CATI has absorbed an important share of the budget relative to the small share of responses it yields, raising the question of whether a lighter, cheaper follow-up strategy can preserve estimation quality. This five-month internship at Statistics Canada, part of the EMOS-labelled Master's degree at ENSAI (France), compares three strategies in R using real data from the respondents of the last CHS: (1) the current CHS design, reproduced as a reference; and, under the same subsampling design – only a subsample of non-respondents to the first, online self-response mode receives full CATI follow-up – (2) the historical Hansen-Hurwitz (1946) estimator and (3) a non-probability approach, treating CATI-followed units as a probability reference sample and remaining online self-response-only units as a non-probability sample corrected for the unknown selection mechanism. Among several techniques for this last approach, only a Chen, Li and Wu (2020) pseudo-likelihood estimator has been implemented so far. For the remainder of the internship, other non-probability methods available in R will be explored, in particular the 'nonprobsvy' package by Chrostowski, Chlebicki and Beręsewicz (2025), presented at uRos2024, which implements inverse probability weighting, mass imputation and doubly robust estimators; this package has not yet been used, but the first Chen-Li-Wu results already confirm that this type of estimation is viable. Preliminary simulations show results comparable between the non-probability approach and the Hansen-Hurwitz method, supporting the feasibility of this framework. Bias increases across the four non-response scenarios reproduced (MAR or MNAR for each of the two collection modes). A conditional non-response adjustment distinguishing early and late respondents, in the spirit of Hansen-Hurwitz, reduces bias more than the classic pooled adjustment when the MNAR assumption holds for the first, online mode; this improvement propagates to the non-probability estimator. These results were obtained without holding the collection budget constant and, as expected, variance increases. The internship continues until the end of August, and remaining work will address a second, less intuitive question: whether, at constant budget, reinvesting the cost savings generated by interviewing far fewer non-respondents, e.g. into a larger self-response sample or more intensive CATI follow-up, could allow a gain in quality from subsampling non-respondents to the first mode.

Typology of business ecosystems in European Union countries

Dalidovich Ulada[1]

1 - Bratislava University of Economics and Business / Ekonomická univerzita v Bratislave (Slovakia)

The aim of this study is to analyse the transformation of entrepreneurial ecosystems across the countries of the European Union in the pre-crisis and post-crisis periods and to develop their dynamic typology using selected data science methods and multivariate analysis. The study focuses on identifying different types of entrepreneurial ecosystems, tracking their changes over time, and predicting the possible future development of countries within the European entrepreneurial space. The analysis is based on panel data for EU countries. The input variables include indicators of business demography, entrepreneurial activity, growth potential and innovation performance obtained from international databases, primarily Eurostat Business Demography and Structural Business Statistics, complemented, where appropriate, by OECD Entrepreneurship Indicators. The methodological framework combines dimensionality reduction techniques, cluster analysis, cluster validation methods and selected predictive classification models. The statistical analysis will be carried out using the R statistical environment, providing a reproducible framework for data processing, multivariate analysis and predictive modelling. Particular emphasis is placed on comparing the pre-crisis and post-crisis periods, identifying changes in the structure of country groups and analysing transitions of countries between different types of entrepreneurial ecosystems. The predictive component aims to estimate the probable future development of countries, including possible changes in their membership in the identified ecosystem types. Special attention will be paid to Slovakia and the Visegrád Group (V4), examining their position within the European typology, their development trajectory and their potential future direction.

Author Index

Authors and co-authors are listed alphabetically by their names as printed. References give the first page of each abstract.

Adam Marsal: 25

Alexander Kowarik: 58

Alexi Dere: 58

Alice Mainguené: 54

Alpay Kocak Necmettin: 31

Ana Markovic: 15

Ana-Maria Ciuhu: 57

Andreas Jahn: 44

Andreea-Denisa Trascan: 54, 57

Andrei Suciu: 57

Andrés Christiansen: 53

Ane Seierstad: 32

Anna Boukouvala: 63

Anna Smyk: 41

Antunez Kim: 9

Armando D'aniello: 29

Barragán Andrés Sandra: 6

Bateman Mcbride: 32

Bernhard Fischer: 38

Bernhard Meindl: 36

Bjørkholt Solveig: 40

Bogdan Oancea: 43

Camelia Maria Dan: 30, 54

Carmen Limone Fenicia: 63

Carolina García-Martos: 37

Cetinkaya-Rundel Mine: 6

Cheyenne Amoroso: 37

Christina Pierrakou: 40

Christophe Dervieux: 37

Christophe Dutang: 52

Christopher Nölting: 27

Ciara Gribben: 32

Ciprian Alexandru: 43, 48

Claude Lamboray: 23

Clément Rousset: 25

Cristina Boboc: 59

Cristina Faricelli: 12

Cristina Rodica Boboc: 30

Dalidovich Ulada: 64

Damien Dupre: 15

Daniel Ollech: 47

Daniel Rakotomalala: 56

Daniel Ward: 36

De Jonge Edwin: 10

Dervieux Christophe: 6

Desclodure Julien: 7

Dicko Ahmadou: 38

Dimitri Lemaire: 22, 34

Dunja Katić: 15

Duygu Dedeoglu: 14

Edwin de Jonge: 12

Eileen Vattheuer: 20, 58

Elie Chosson: 18

Elisabeth Morand: 24

Erika Lee: 13, 21

Eulalie Delaune: 28, 50

Fanny Allard: 25

Ferrer-Pradines Nadège: 7

Fethi Saban Ozbek: 16

Gauri Kamat: 33

George Tsaklidis: 63

Georgia Panagopoulou: 20

Georgios Lykos: 20

Germán Aneiros: 37

Gianina-Maria Petrașcu: 57, 59

Gianluigi Salvucci: 29

Gorka Campandegui García: 55

Guillaume Osier: 23, 58

Guy Mélard: 45

Hakile Resulbegoviq: 42

Hanan Touzani: 25

Hannah Frick: 47

Iliass Taleb: 13

Ilya Kashnitsky: 60

Ina Karapetyan: 35

Ines Hiverlet: 23

Ioana Barbulescu: 57
Ioannis Nikolaidis: 17
Islam Mousumi Mona: 40
Ivar Hamre Jørn: 32
Jan Koopman Siem: 37
Jan van Der Laan: 22
Javed Maryam: 40
Jelena Suzić: 15, 27
Jens Kristoffer Haug: 32
Jiahao Cui: 33
Johannes Gussenbauer: 58
Johannes Straubinger: 21
Johansen Odin: 40
José A. Vilar: 37
Juan Pablo Ferreira: 46
Julia Maestro López: 55
Julien Desclodure: 22, 34
Kamil Grabarczyk: 35
Kamil Kowolik: 30
Kamil Wilak: 22, 39
Karsten Webel: 41
Konstantinos Molfetas: 40
Kyrylo Chalyy: 51
Lucile Bodson: 58
Madoyan Habet: 35
Marcello D’orazio: 12
Marcin Szymkowiak: 22, 39
Marco Caputi: 52
Maria Paula Caldas Rivera: 19
Mark van Der Loo: 34, 43
Marlene Steiner: 20
Martijn Tennekes: 11
Martin Slawski: 33
Matthias Templ: 58
Matthieu de Canteloube: 11
Maxime Pettinger: 39
Melinda Tokai: 15
Meriem Said: 19
Merlin Scherer: 26
Michael Levi-Valensin: 55
Michael Neutze: 23
Michail Vikentios: 20
Michonneau François: 6, 9
Murad Hajiyev: 27
Nadège Ferrer-Pradines: 22, 34
Nathalie Rigollot: 19
Ngoc-Han Tran: 34
Nina Niederhametner: 20, 58
Nobuo Yoshida: 39
Olav ten Bosch: 43
Olga Bogomolets: 51
Oliver Hauke: 14
Osman Sert: 14
Øyvind Langsrud: 32
Panaiotis Thelma: 8
Patrice Kiener: 24, 52
Pérez Bote Adrián: 6
Pierre Lamarche: 25, 49
Powell Rebecca: 10
Priyanjali Bukke: 33
Quang Pham Dinh: 32
Raphael Simon: 19
Razvan Vasile: 57
Rebecca Killick: 5
Reijer Idema: 55
Renato Magistro: 12
Rodolfo Ilizaliturri López: 19
Roee Gutman: 33
Romain Lesur: 46
Romain Winckell: 50
Roman Hryshchenko: 51
Romina Filippini: 33
Romuald Thion: 49
Russell Mccreath: 32
Samad Sarferaz: 26
Sara Otavo Peña: 13, 21
Simon Couch: 5
Simon Normandeau: 13, 21
Simona Cafieri: 56
Simona Toti: 33
Smaranda Cimpoeru: 61
Spyridon Dimas: 20

Stefanos Giakoumatos: 17
Susie Jentoft: 32
Suzanne Heidsieck: 26
Tanguy Barthelemy: 59
Tennekes Martijn: 7
Thomas Delclite: 28
Thomas Guinhut: 50, 64
Tomasz Józefowski: 39
Torvholm Frøystad Sarah: 40
Vahe Movsisyan: 35
Valentina Vasile: 57
Valeria Rossi: 62
Van Der Laan Jan: 10
Vasiliki Spiliopoulou: 40
Velásquez Isabella: 10
Viktor Stepanenko: 51
Viktoria Kis: 13, 21
Vincent Biot: 19
Vincent Guyader: 11
Vincent Reduron: 42
Vlad-Andrei Petre: 61
Vuk Čađenović: 42
Winfried Moser-Nussgruber: 30
Younes Saidani: 34
Yves Coltier: 53
Zakariae Halhal: 51
Zimmer Stephanie: 10
Zineb Aoulad Hammou Achou: 51

Practical Conference Information

uRos2026 · Paris, France · 18–20 November 2026

Tutorials: Wednesday, 18 November.
Conference sessions: Thursday, 19 November, and Friday, 20 November.

Venue and address

Pierre Mendès France Conference Center
139 Rue de Bercy
75012 Paris
France

Hosted by INSEE at the premises of the Ministry of Economics, Finance and Industrial and Digital Sovereignty.

Conference rooms

Room	Programme use
Auditorium	Plenary sessions and track A
Bloch-Lainé 1 (access “cour”)	Track B
Bloch-Lainé 2 (access “rue”)	Track C

Tutorials also take place in these rooms. Consult the conference programme for individual session times and room assignments.

Conference contact

Email: uros2026@insee.fr
Conference website: https://r-project.ro/conference2026.html

Publication contact

Email: urosconf@gmail.com
Book of Abstracts website: https://uros.r-project.ro/