Book of Abstracts (2025)

uRos (Use of R in Official Statistics)
24-26 November 2025, Bucharest, Romania

Contents

A framework for adopting/developing R packages that implement sound statistical methods for production of official statistics

Authors

- Marcello D'Orazio (Italian National Institute of Statistics (ISTAT))

Abstract

The Italian National Institute of Statistics (Istat) is strongly committed to improving its statistical production processes by integrating traditional and non-traditional data sources. This modernization requires cost-effective statistical software tools that implement sound methodologies. For this reason, Istat has adopted R and, more recently, Python alongside proprietary software to increase flexibility, foster innovation, reduce costs, and align with national and European government initiatives. This has led to the in-house development of R packages. These packages, together with those developed by other national statistical offices, facilitate the adoption of recent methodological innovations in the production of official statistics. Due to the increased use of open-source tools like R and Python, Istat recently decided to define a clear strategy for open-source statistical tools. The first step in this strategy was the development of guidelines and methodological governance, which were published at the end of 2024. The methodological governance, developed within the Methodological Directorate, aims to provide a framework for the development or adoption of statistical open-source software tools (OSST) developed externally. In practice, it is tailored to the development of additional R (Python) packages, as well as the adoption and possible adaptation of packages developed externally. Essentially, a standard procedure and corresponding flowchart are defined, covering everything from acquisition to internal and external endorsement, dissemination, and promotion. This procedure relies on the DevOps framework applied to implementing a new methodology in an R (or Python) package. Here, "new methodology" is intended broadly, including improving established methods to enhance quality, performance, etc. The procedure is supplemented by a note on licensing issues that will need to be addressed by IT and legal experts at the organizational level in the future. This was an important step toward comprehensive governance for successfully using open-source statistical tools in official statistics production. The next step is to define the corresponding IT counterpart, which is a key factor in the successful adoption of these tools.

References

No References available

ARTIC: Automatic Random error Treatment and Imputation for Categorical variables

Authors

- Simona Toti (Italian National Institute of Statistics (ISTAT))
- Romina Filippini (Italian National Institute of Statistics (ISTAT))

Abstract

The identification and treatment of non-sampling errors is a fundamental step in the production of Official Statistics, aimed at ensuring higher accuracy and quality of the resulting estimates. Traditionally, the Editing and Imputation (E&I) process represents one of the most time-consuming phases. In the context of Surveys, developing generalized tools that can be easily adapted to different questionnaires offers valuable support throughout the entire process. To this end, a standardized procedure in R, called ARTIC (Automatic Random error Treatment and Imputation for Categorical variables), has been developed for the automatic detection and correction of random errors for categorical variables, based on the Fellegi-Holt methodology. Fully implemented in R, the procedure is customizable, extendable, and easy to share. The Editing and Imputation (E&I) process for random errors consists of three sequential phases: checking the data to identify units affected by errors; localizing the erroneous variables within each unit; and imputing the identified errors. In ARTIC, each step of the process—checking, localizing, and correcting random errors—is carried out using R libraries such as validate, validatetools, errorlocate, VIM, and ad hoc functions. The function parameters are defined directly by the user. Starting from a raw dataset, and based on user-defined consistency rules, the procedure produces output files (data and reports) for each phase of the process. The final output is a dataset corrected for random errors in accordance with the predefined rule set. To make ARTIC accessible to non expert user in R programming, interactions such as data loading, parameter specification, etc. are managed through interactive web applications developed with the Shiny library. The complete ARTIC procedure consists of the following three components: a main script to be launched via RStudio, a file containing the ad hoc functions, and the Shiny apps. The procedure is accompanied by a detailed manual that provides theoretical background on the implemented methods and guides the user through all steps of the procedure, describing the required input structure and the contents of the outputs.

References

No References available

Automatic Enforcement of Data Confidentiality in Official Statistics Using R Shiny

Authors

- Fethi Saban Ozbek (Turkish Statistical Institute)

Abstract

In Türkiye, data confidentiality in official statistics is governed by the "Regulation on Procedures and Principles Regarding Data Confidentiality and Confidential Data Security in Official Statistics," published in the Official Gazette dated 20.06.2006 and numbered 26204. In accordance with this legislation and in a manner compatible with EU legislation, the protection of confidential information in tabular data is mandatory. According to Article 6 of the regulation, a cell in a tabulated dataset is considered confidential if (a) it is based on fewer than three statistical units, or (b) although it is based on three or more units, one unit accounts for more than 80% or two units together account for more than 90% of the total value in that cell (primary suppression). Additionally, any non-confidential cell that could lead to the disclosure of a confidential value at the same level of disaggregation is also suppressed to ensure full protection (secondary suppression). Such confidential cells may only be disclosed after being combined with other cells in a way that prevents the identification of the underlying confidential data. This study introduces an R Shiny-based application developed to automatically apply these confidentiality rules to both database-extracted and user-uploaded (e.g., Excel) datasets. The application dynamically identifies and suppresses both direct and indirect confidential cells during tabulation processes, fully adhering to the national legal framework. The application enhances transparency, standardization, and efficiency in the production of confidentialized statistical tables, significantly reducing the manual workload and error risk. Moreover, it ensures that official statistical products are aligned with legal confidentiality obligations before dissemination. This case study highlights the use of R and Shiny in the modernization of statistical production processes and demonstrates how R can serve as an effective tool for ensuring data security and legal compliance in official statistics.

References

No References available

Checklist improves collaboration, quality and visibility of your code

Authors

- Thierry Onkelinx (Research Institute for Nature and Forest (INBO), Belgium)

Abstract

The checklist package is a set of rules for R packages and R source code projects. The ruleset covers several topics: folder structure, filename conventions, spelling, code style, citation metadata, licence, contribution guidelines. Adherence to a common set of rules within an organisation facilitates collaboration between its members. Enforcing citation metadata and an open source licence improves the visibility of projects. Automated checks via GitHub Actions detect problems as soon as possible. Checklist helps you getting started by providing functions to create a new project of package from scratch. These functions guide you interactively through the process and allow you to reuse information. Checklist is based on the rcmdcheck, lintr, pkgdown, codemetar and hunspell packages. Where applicable, we use the same rules for projects and packages. The maintainer can choose which parts of the ruleset apply to a project. In the case of an R package, the entire ruleset is mandatory. The ruleset is partly hard coded in the checklist package, partly set via an organisation level repository. Publishing code on Zenodo is easy if you link to your GitHub repository. Each release on GitHub triggers a new version on Zenodo with a specific DOI. A GitHub action creates a new release for each new version of the package. Checklist is available through https://inbo.r-universe.dev/checklist.

References

No References available

Clusterwise Tensor GAM–NB for NUTS-3 Regional Air-Pollution Mortality in Italy

Authors

- Francesca Del Duca (University of the Study of Campania Luigi Vanvitelli, Italy)
- Rosanna Verde (University of the Study of Campania Luigi Vanvitelli, Italy)
- Antonio Balzanella (University of the Study of Campania Luigi Vanvitelli, Italy)

Abstract

Based on the Burden of Disease of Air Pollution dataset by the European Environment Agency (https://doi.org/10.2909/258fa83c-dec6-4d88-b1fb-959f2d90008f), this study introduces a statistical learning framework for modelling regional mortality effects linked to air pollution in Italy at the NUTS-3 level. The dataset provides comprehensive estimates of mortality attributable to key air pollutants, including PM . , NO , and O , enabling detailed epidemiological modelling at fine spatial resolution. To accommodate the complex, nonlinear associations between multiple pollutants and mortality counts, a hybrid Generalized Additive Model (GAM) with a Negative Binomial response is employed to address overdispersion common in count data. The model incorporates tensor-product smooths to capture pollutant interactions, spatial and temporal smooth terms to represent geographic and temporal trends, and random effects for administrative units. An autoregressive (AR(1)) correlation structure accounts for serial dependence within regions, while population exposure is integrated as an offset to facilitate standardized rate estimation. The core methodological contribution lies in the integration of the hybrid GAM into a cluster-wise estimation framework that jointly infers latent subgroup structures and exposure–response functions. An iterative algorithm alternates between fitting GAMs within clusters and reassigning observations based on prediction error minimization, thereby unifying clustering and model estimation. This approach contrasts with traditional methods that separate clustering from modelling or rely exclusively on covariate similarity. By aligning cluster membership with the epidemiological model's objective, the framework identifies latent heterogeneity in pollutant–mortality relationships beyond predefined stratifications, enhancing interpretability and epidemiological insight. The method's flexibility and scalability render it well suited for environmental health research involving spatially structured populations and complex exposure profiles. Overall, the proposed framework advances the use of statistical learning and spatial modelling within the field of official statistics applied to public health, offering a novel tool to detect subtle subgroup-specific effects in large-scale environmental epidemiology. All statistical analyses were conducted in R (version 4.4.3), an open-source software environment widely adopted in epidemiology and official statistics. The use of open-source tools ensures transparency and reproducibility, facilitates the implementation of advanced modelling techniques and enables researchers and institutions to replicate, validate, and extend the analysis without licensing constraints. This open-access dimension is particularly relevant in the context of official statistics, where methodological accessibility, comparability across studies, and societal impact are essential.

References

No References available

Confidentiality of SBS data for SDMX transmission, case of Albania

Authors

- Elsa Dhuli (INSTAT)
- Elona Dushku (Bank of Albania)
- Vanesa Celaj (INSTAT)
- Anxhela Petriti (INSTAT)

Abstract

Structural Business Survey (SBS) is an essential component of economic statistics in Albania. In compliance with Eurostat regulations and Albanian Law No.17/2018 On Official Statistics, the results of SBS 2021 were prepared to be transmitted in SDMX (Statistical Data and Metadata eXchange) format. Controlling disclosure risk is highly important to protecting confidential and sensitive information before it is transmitted. NSI have developed some 'safety rules' or 'sensitivity rules' as measures to assess disclosure risks. INSTAT aims to determine optimal SDC method and solution that minimize disclosure risk while maximizing the utility of the data for transmission purpose.This paper presents an experimental study on Structural Business Statistics (SBS) magnitude tables by applying Statistical Disclosure Control techniques.Tau-Argus was initially employed as the primary tool for identifying confidential cells prior to data transmission. However, INSTAT sought to transit to a more efficient solution for the adjustment of sensitive cells available in the R-package sdcTable. Special attention was given to the implementation of this method due to its ability to mix harmoniously with R-based workflows, following the Open Source Strategy 2022–2030 of INSTAT, which promotes the adoption of open-source software in order to increase transparency, reproducibility and reduce cost. Suppression consists of sensitive cells identification based on dominance, frequency rules, etc, followed by secondary suppression in order to prevent re-identification of protected data. Application of sdcTable followed a step-by-step approach to defining hierarchical dimensions, the implementation of safety rules and the validation of the final output for SDMX transmission. The final tables were generated automatically in the format of the SDMX template as part of the process, making preparation of transmission-ready data simpler and faster. This paper exemplifies the practical use of open-source software to improve secure and standardized statistical production. The solution provides a partially reproducible template for national statistical offices in EU candidate countries, seeking to modernize and automatize confidentiality protection of tabular outputs for SDMX transmission.

References

No References available

Designing Effective Visualizations: Choosing Colors in Practice

Authors

- Madoyan Habet (American University of Armenia)

Abstract

This presentation explores how color perception, theory, and modeling intersect in effective data visualization. Through examples in R and ggplot2, it demonstrates the practical use of color models and color pallets and teaches how to chose colors based on the type of the variable.

References

No References available

Developing a Composite Index Prototype with R: An Application for SDG Monitoring

Authors

- Eduard Luca (Romanian National Institute of Statistics (INS))
- Ana-Maria Ciuhu (Romanian National Institute of Statistics; Institute of National Economy, Romanian Academy)
- Ioan-Silviu Vîrva (Romanian National Institute of Statistics (INS))

Abstract

The measurement of sustainable development in regional contexts requires access to simple, accessible tools that have the capacity to integrate multisource data into usable conclusions. This presentation introduces a prototype composite index, created with R, aimed at quantifying and comparing regional sustainability performance according to predefined proxy indicators for the United Nations SDGs. The index delivers an annual score ranging from 0 to 1, facilitating comparison at the national, NUTS2, and NUTS3 levels. The method, coded in R, encompasses data cleaning, normalization, weighting, and aggregation steps. The prototype integrates both official statistics and satellite-derived indicators, including forest coverage, the Normalized Difference Vegetation Index (NDVI), and the Normalized Difference Built-Up Index (NDBI), derived from Sentinel-2 data. Additionally, gridded population data were used to provide demographic context. By using R as an open-source and multi-purpose environment, the project demonstrates that statistical innovations are possible without relying on proprietary software packages. The composite index serves as an interface between sophisticated statistical data and stakeholder comprehension, thereby facilitating communication, policy assessment, and strategic planning at various administrative levels. Acknowledgement: The project was developed by the NSI Romania team at the European Big Data Hackathon 2025: Earth Observation: from Space to European Statistics.

References

No References available

dpmaccount: Fast and Flexible Bayesian Demographic Accounting in R with Template Model Builder

Authors

- Daniel Ward (Office for National Statistics (ONS), UK)
- Duncan Elliott (Office for National Statistics (ONS), UK)

Abstract

National Statistical Institutes (NSIs) increasingly require timely and coherent demographic estimates from diverse and often noisy data sources. The R package dpmaccount offers a novel solution by implementing a robust Bayesian demographic accounting framework. It empowers users to integrate multiple datasets by defining system models for underlying demographic rates and flexible data models that specify the link between observed datasets and the true unobserved values we want to calculate. The core innovation of dpmaccount lies in its use of Template Model Builder (TMB) for computationally efficient inference. By leveraging Laplace approximations and automatic differentiation via a pre-compiled C++ template, dpmaccount achieves significant speed gains over traditional MCMC methods, making complex, high-dimensional demographic modeling practical for routine production and research. This presentation will introduce the dpmaccount package, highlighting its user-friendly object system for model specification, the underlying statistical methodology, and its modular architecture which facilitates the addition of custom data models. We will demonstrate its capabilities through a case study using real-world demographic data, showcasing the workflow from model definition to results extraction and analysis. The dpmaccount package sits at the heart of the Office for National Statistic's (ONS) Dynamic Population Model (DPM) which itself is central to the transformed population and social statistics system that ONS is building, representing a significant advancement in statistical software for official statistics and offering a powerful tool for producing consistent, uncertainty-quantified demographic accounts. This work underscores the potential of combining Bayesian methods with high-performance computing tools like TMB to meet modern statistical challenges.

References

No References available

finitization: A New R Package for Efficient Approximation of Discrete Distributions

Authors

- Bogdan Oancea (University of Bucharest; National Institute of Research and Development for Biological Sciences, Romania)

Abstract

We introduce finitization, a new R package designed to approximate discrete probability distributions with high accuracy and exceptional computational efficiency. A finitized distribution of order n exactly preserves the first n moments of a target distribution such as the Poisson, Binomial, Negative Binomial, or Logarithmic distributions, while allowing faster random variate generation. This approach enables a powerful trade-off between statistical fidelity and computational speed, ideal for applications in large-scale simulation, stochastic modeling, and machine learning. The core of finitization is implemented in optimized C++ code using the Rcpp interface, ensuring seamless integration with the R environment. Symbolic manipulation is handled via the GiNaC library, enabling the exact algebraic derivation of finitized probability mass functions based on truncated probability generating functions. The package employs efficient alias sampling techniques for random generation, achieving near constant-time performance regardless of distribution complexity. finitization is modular and extensible. Its architecture separates symbolic modeling, numerical approximation, and sampling logic, making it easy to extend to new families of distributions. The package provides human-readable outputs of finitized density functions both in standard R expressions and LaTeX format, supporting documentation, teaching, and publication needs. Key application areas include: Large-scale Monte Carlo simulations where runtime is critical, Approximate Bayesian computation where moment preservation matters more than exact distributional fidelity, Synthetic data generation for AI and data science experiments, Educational environments needing controlled complexity for statistical teaching. The package is thoroughly validated using statistically robust unit tests that account for the randomness of simulation-based methods. It is built to be cross-platform (Linux, macOS, Windows) and compliant with CRAN policies, dynamically linking to standard external libraries (GMP, CLN, GiNaC). In summary, finitization combines symbolic computation, high-performance algorithms, and solid statistical guarantees to provide a practical and elegant solution for modern simulation challenges in statistics, data science, and applied probability.

References

No References available

Hands-on Bayesian Demographic Accounting: A Practical Tutorial on the R Packages Behind the UK's Dynamic Population Model (DPM)

Authors

- Daniel Ward (Office for National Statistics)
- Duncan Elliott (Office for National Statistics)

Abstract

Producing consistent and reliable demographic estimates from multiple, imperfect, and noisy data sources is a critical challenge for official statistics and demographic research. This tutorial offers a hands-on introduction to the packages that form the backbone of the UK Office for National Statistics (ONS) Dynamic Population Model (DPM), guiding participants through the entire workflow of building, fitting, and interpreting Bayesian demographic accounts. It will introduce participants to the R packages dpmaccount, which provides a powerful and computationally efficient framework for Bayesian demographic accounting by leveraging Template Model Builder (TMB) for fast inference, as well as dpmsplit to split migration estimates into their separate components. Participants will learn to: Understand the principles of Bayesian demographic accounting and the role of dpmaccount. Specify system models for demographic rates (births, deaths, migration) using prior information. Define various data models to link observed data (e.g., from surveys, administrative records etc) to true underlying counts, accounting for different error structures. Run the estimation process using functionality from the dpmaccount package, understanding how TMB is used for efficient computation. Extract, interpret, and visualise key results, including estimated population stocks, flows, demographic rates, and associated uncertainty measures, using interactive dashboards from the dpmdash package. Gain an introductory understanding of the package's C++ back-end and the principles for extending dpmaccount with custom data models, highlighting its flexibility for advanced users. Split the combined migration estimates produced by the dpmaccount demographic accounting process into separate components (internal, cross-border, and international) using functionality from the dpmsplit package. This tutorial is aimed at statisticians, demographers, and data scientists working with population data in National Statistical Institutes (NSIs), academia, or research institutions. Participants should have a basic understanding of R and demographic concepts, have RStudio installed (or an alternative R IDE), as well as have installed the dpmaccount, dpmsplit, and dpmdash packages. By the end of the session, attendees will be equipped to apply the various DPM methods to their own demographic accounting problems, harnessing its speed and flexibility.

References

No References available

Harmonizing Greek Census Data Across Time

Authors

- Athanasios Stavrakoudis (University of Ioannina, Greece)

Abstract

Greece has raw data from national censuses every ten years since 1971, resulting in six rounds of rich demographic data. However, a major obstacle to longitudinal analysis lies in the frequent and complex changes to the country's administrative boundaries over time. Communities, municipalities, prefectures, and other units have undergone repeated mergers, splits, and reclassifications—making direct comparisons between censuses difficult, if not impossible. Adding to the challenge, geospatial data and administrative shapefiles are only available for the two most recent census years, 2001 and 2011. Earlier years lack official spatial representations, further hindering historical comparisons. To address this issue, I developed a comprehensive solution using R and key packages such as tidyverse, sf, and googlesheets4. I systematically matched every settlement in Greece (approximately 13,500 in total) to all relevant administrative boundaries from 1971 to 2021. This harmonized framework allows researchers to compare census statistics consistently over time—at the level of municipalities, NUTS3 regions, and other administrative units—despite boundary changes. As a proof of concept, I built a Shiny application that visualizes intergenerational mobility in Greece. Using census data on educational attainment within nuclear families, the app demonstrates how spatial harmonization enables meaningful longitudinal analysis of social mobility across regions and time periods. This framework provides a robust foundation for a wide range of historical and socioeconomic studies in Greece, ensuring both spatial consistency and analytical comparability across five decades of census data. Intergenerational mobility map can be accessed here: http://stavrakoudis.econ.uoi.gr:9988/greece/igm/

References

No References available

How to build a statistics bot in almost no time?

Authors

- Bernhard Meindl (Statistics Austria)
- Alexander Kowarik (Statistics Austria)

Abstract

This paper presents a prototype chatbot designed as an intelligent search assistant rather than an independent knowledge source. It exclusively processes content from statistik.at (or your institute's website), ensuring factual correctness, validity, and consistency with official publications of Statistics Austria. Strict system prompts enforce objectivity: only queries directly related to statistik.at are answered, while irrelevant requests are rejected. The architecture combines a large language model (gpt-oss 120b) with a Google Custom Search Engine restricted to statistik.at. Implemented initially in R and Shiny for pilot testing, the system provides transparent, reliable responses and a foundation for future production-level applications. The R code is available on Github for everyone to test themselves.

References

No References available

Impact of Globalization on Macroeconomic Statistics

Authors

- Fahad Fayaz (EMOS Student, Czech Republic)

Abstract

The paper critically assesses whether traditional economic measures such as GDP are still relevant to their users in today's fast-evolving economic environment. It explicitly aims to understand it within the context of globalization's significant impact on macroeconomic indicators. In other words, we need to know whether macroeconomic indicators are relevant across the globe; if not, what should we do as statisticians to stay relevant? Furthermore, this study addresses how updates to the System of National Accounts (SNA) affect the operations and output produced by statistical offices. An innovative aspect of this paper is also to propose methodological improvements in existing macroeconomic measurements to better reflect the realities of today's interconnected world. Lastly, the paper uses unsupervised machine learning algorithms - cluster analysis using R programming to evaluate the effects of globalization on different countries. Specifically, it identifies which OECD countries are substantially affected by globalization and which are less affected. Similarly, it examines non-OECD countries to determine those substantially impacted by globalization and those less affected by globalization.

References

No References available

Introduction to Network Analyses for Official Statistics

Authors

- Mark van der Loo (Statistics Netherlands and University of Leiden)

Abstract

In traditional statistics a population is regarded as a set of independent entities. Such an approach precludes a deep understanding of phenomena that may arise from the interaction between persons, households, or businesses. Recently, the use of Network Science to study society and economics has been gaining attention in official statistics. For example, Statistics Netherlands is now publishing segregation measures, based on a population-scale network[1]. In this workshop, I will first introduce networks, and some use cases that are already published or investigated in official statistics. Next, we will dive into basic network theory using R, including visualization, measures for node importance, and clustering. Participants are expected to have a sound command of English, and a working knowledge of R. The R package igraph[2] will be used for exercises and demonstrations.

References

- [1] VAN DER LAAN, Jan, et al. A whole population network and its application for the social sciences. European sociological review, 2023, 39.1: 145-160. [2] Csardi G, Nepusz T (2006). "The igraph software package for complex network research." InterJournal, Complex Systems, 1695. https://igraph.org.

Mapping poverty at the level of subregions using a univariate Fay-Herriot model

Authors

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

Abstract

The European Survey on Income and Living Conditions (EU-SILC) is the basic source of information that Statistics Poland use to produce national and regional poverty indicators. This is also true in the case of other countries that are facing a growing demand for good poverty maps. In order to implement appropriate social strategy measures, which are consistent with the EU's cohesion policy, it is necessary to measure poverty and provide information about this phenomenon at lower levels of spatial aggregation. Poverty maps are used to inform decision making with important political implications, such as the allocation of development funds by governments, ministries of infrastructure and development or international organizations, such as the World Bank. Such decisions should be based on the most accurate poverty indicators, estimates or figures and should be delivered at the lowest level of spatial aggregation. However, given small sample sizes in relevant cross classifications of the EU-SILC survey, one needs to apply the latest techniques of indirect estimation and rely on alternative data sources (censuses or administrative registers) to estimate the parameters of interest at low levels of spatial aggregation with acceptable precision. Since the EU-SILC survey does not cover adequately all specific areas or population subgroups, the required information can only be obtained using small area estimation (SAE) techniques, which are based on the idea of "borrowing strength". In Poland, for instance, EU-SILC data are only sufficient to publish poverty indicators at the level of the whole country and at NUTS 1 level. Owing to small sample sizes, reliable estimates at lower levels of spatial aggregation cannot be delivered. The main aim of the presentation is to report selected results of a poverty mapping study undertaken by Statistics Poland and the World Bank involving the use of a univariate Fay-Herriot model, R software and data from different statistical sources at NUTS 3 level, at which no official poverty statistics have been published in Poland to date. The univariate Fay-Herriot model and poverty maps for Poland will be created using R statistical software, which provides a comprehensive environment for small area estimation methods through specialised packages such as SAE, HBASE and EMDI. These packages offer robust implementations of various techniques, including the Fay-Herriot model.

References

No References available

NMAR: An R Package for Estimation under Non-ignorable Non-response in Sample Surveys

Authors

- Igor Kołodziej (Warsaw University of Technology, Poland)
- Mateusz Iwaniuk (Warsaw University of Technology, Poland)
- Maciej Beręsewicz (Poznań University of Economics and Business, Poland)

Abstract

Non-ignorable (NMAR) non-response presents a persistent challenge in official survey statistics, where missingness mechanisms depending on unobserved variables can significantly bias key national indicators. The NMAR package addresses this issue by providing a comprehensive suite of modern estimation methods within a unified API, specifically designed for the complex reality of NMAR scenarios in official statistics. The package implements following techniques: Generalised calibration with more variables in cali- bration than response model [1]; Generalised method of moments [2]; Empirical likelihood-based ap- proaches [3],[4]; Fractional imputation and non-parametric methods (Minsun Kim Riddles, Jae Kwang Kim, Jongho Im 2016)[5] Currently, the package is under development with a planned submission to the Comprehensive R Archive Network (CRAN) prior to the conference. The package enables researchers to produce defensible estimates from survey data affected by non- ignorable missingness, while fostering methodological transparency in official statistics. Acknowledgements The NMAR package is developed under the project Towards census-like statistics for foreign-born populations funded by the National Science Centre, Poland (OPUS 20 grant no. 2020/39/B/HS4/00941).

References

- [1] Kott, P. S., Liao, D. (2017). Calibration Weighting for Nonresponse that is Not Missing at Ran-
- dom: Allowing More Calibration than Response-Model Variables. Journal of Survey Statistics and
- Methodology, 5(2), 159–174.
- [2] Kosuke Morikawa, Jae Kwang Kim (2016) A note on the equivalence of two semiparametric estimation
- methods for nonignorable nonresponse, S tatistics & Probability Letters
- [3] Qin, J., et al. (2002). Estimation with Survey Data under Nonignorable Nonresponse or Informative
- Sampling. Journal of the American Statistical Association, 97(457), 193–200.
- [4] Fang, F., et al. (2010). Empirical Likelihood Estimation for Samples with Nonignorable Nonresponse.
- Statistica Sinica, 20(1), 263–280.
- [5] Riddles, M. K., Kim, J. K., Im, J. (2016). A Propensity-score-adjustment Method for Nonignorable
- Nonresponse. Journal of Survey Statistics and Methodology, 4(2), 215–245.

Painlessly Improve Your Git History

Authors

- Maëlle Salmon (rOpenSci)

Abstract

A confident Git practice can change the working life of anyone writing code or prose with R, resulting in a useful history to browse or inspect, the ability to work in parallel on different aspects, and so on. In particular, the best Git practice is to create small atomic commits with informative messages. But why? And how? Learn three reasons why small Git commits are worthwhile. Find out how to create them realistically, without too much hassle. Practice safely with the saperlipopette R package, and share your own Git tips and questions! I will use playgrounds created by saperlipopette to illustrate Git techniques and strategies. The participants can then re-create the exact same playgrounds and practice before we move on to the next topic. This tutorial will therefore be active and engaging! Prerequisites: - Basic knowledge of Git (git add, git commit, git push/pull, branch creation, merging branches on a platform like GitHub). - Local installation of Git. https://happygitwithr.com/install-git + https://happygitwithr.com/hello-git - Install saperlipopette. https://docs.ropensci.org/saperlipopette/ - I'll do two demos using Positron (the other ones in RStudio IDE) https://positron.posit.co/, if you want to test too you can install it and install the GitLens extension. But the exercises can be done no matter in which interface you use R and Git! I have delivered part of the content (minus the participants' time for practice) at several meetups: - https://masalmon.eu/talks/2025-04-11-meilleur-historique-git/ - https://masalmon.eu/talks/2025-05-13-hack-your-way-git-history/ I will deliver the same workshop (including participants' time for practice) in French in June: https://stateofther.netlify.app/#upcoming_workshops And in Spanish for the rOpenSci champions program. The saperlipopette package has messages in English, French or Spanish. I can answer questions in English, French, Spanish, German, Swedish, Catalan.

References

No References available

Partial Correlation Network Analysis of Well-being Indicators in Italian Provinces Using EBICglasso

Authors

- Francesca Tamburrino (University of the Study of Campania Luigi Vanvitelli, Italy)
- Antonio Irpino (University of the Study of Campania Luigi Vanvitelli, Italy)

Abstract

This study investigates the complex interrelationships among official well-being dimensions in Italy, utilizing the Equitable and Sustainable Well-being (BesT) indicators collected by ISTAT (the Italian National Statistical Institute) at the provincial level (NUTS-3). The analysis focuses on 2021 data, covering 106 Italian provinces and 11 thematic domains of well-being, including Health, Environment, Economic well-being, Education, and others. To explore the conditional dependencies among these multidimensional indicators, we estimate a partial correlation network using the Extended Bayesian Information Criterion Graphical Lasso (EBICglasso) algorithm. This method enforces sparsity to produce a more interpretable network, highlighting direct associations while controlling for indirect effects simultaneously. The resulting network provides a data-driven representation of the underlying structure of well-being across provinces, surpassing the limitations of simple correlation analyses. Centrality metrics, adapted for partial correlation networks, were computed to identify the most influential indicators within the network, providing insights into which dimensions play pivotal roles in the structure of well-being. Furthermore, community detection algorithms were applied to identify clusters of closely related indicators, enabling an assessment of the extent to which these clusters align with the official BES thematic domains. This approach also reveals potential cross-domain linkages and latent groupings not captured by the original classification. All analyses were performed in R, mainly using the qgraph, igraph, and bootnet packages, leveraging their capabilities for network estimation, visualization, and stability assessment. This network-based approach offers a novel perspective on the multidimensional nature of well-being, with implications for enhancing multidomain monitoring frameworks in official statistics. Our findings contribute to a deeper understanding of how well-being dimensions interrelate with fine geographic resolution. This methodological framework can support policymakers and statisticians in designing integrated indicators and targeted interventions, highlighting the value of advanced network analysis techniques in the field of official statistics.

References

No References available

procR: An R Package for Fast and Intuitive Custom Cross-Tabulations of Microdata

Authors

- Younes Saidani (Federal Statistical Office of Germany (Destatis))

Abstract

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. However, their creation is not trivial, due to requirements such as: non-exclusive or non-exhaustive summaries of variables (e.g. to display sub-totals, or only selected sub-items), custom grouping of variables (e.g. transforming a continuous age variable into age groups for aggregation), multi-dimensional cross-tabulations (with variable combinations in rows and/or columns), custom cell suppression in accordance with data confidentiality or data quality protocols (e.g. bracketing or replacing sensitive values), and the generation of standardised output formats that include metadata and comply with corporate design guidelines. Although numerous R packages address individual aspects of these needs, there is currently no comprehensive tool that facilitates the entire process of creating these tables from start to finish. In contrast, SAS procedures like PROC FORMAT, PROC TABULATE, and PROC REPORT are well-established for producing custom cross-tabulations; however, they tend to be less flexible, performant, and extensible compared to modern R solutions. To fill this gap, we developed procR, an R package that combines the flexibility and power of R with a streamlined, end-to-end workflow for creating custom tabulations of microdata, supporting all the requirements outlined above. In particular, procR offers (1) an intuitive and R-native syntax for specifying custom variable groupings, allowing for non-exclusive and non-exhaustive categories as well as combinations of variables, thus aiding reproducibility and error reduction. (2) Custom cross-tables are specified using a streamlined and familiar syntax inspired by PROC TABULATE. (3) Common aggregation functions (e.g., weighted counts, means, and sums) are computed efficiently via matrix algebra and sparse matrices, allowing the package to scale exceptionally well to large datasets and complex tables. The package also supports user-defined aggregation functions, which maintain high performance through the use of the efficient collapse package. (4) Tables can be exported to .xlsx-format while implementing custom statistical disclosure control. The presentation begins by outlining the motivation behind the development of the package, highlighting how it improves existing workflows in the German Microcensus and lowers the barrier to transitioning to R. It then provides an overview of the package's current features and includes a live demonstration of a typical use case. The talk concludes with a discussion of planned extensions that are currently under development.

References

No References available

R packages, good vibes only

Authors

- Maëlle Salmon (rOpenSci)

Abstract

Organizations using R often maintain packages answering their specific needs for ingesting data from particular formats, performing specialized statistical analyses, or producing branded reports. Such Institutional toolboxes are time savers and help with continuity of organizations' activities.These toolboxes can be made even more useful and delightful to work with through well-designed curation that cultivates quality at both the package and collection level. In this talk, I will explain how your organization can apply lessons, guidance and tooling from the rOpenSci package review system to strengthen code and interface quality, and improve statistical robustness; how your organization can rely on R-universe to officially onboard your organization's packages, to then easily distribute and showcase them. All of this in a kind, accessible and constructive atmosphere like the one we strive to curate at rOpenSci: good vibes only!

References

No References available

Research Software Engineering in Official Statistics

Authors

- Reijer Idema (Statistics Netherlands (CBS))

Abstract

In official statistics, a lot of code is produced testing methodologies on available data. Research software is primarily written to generate insight at the time of development, but has value beyond that. Other projects may benefit from reproducibility of the results or from reusable portions of the code. The researchers writing the code are experts in methodology, but are generally not trained in software engineering. They are experienced in translating methodology to code and verifying correctness using statistical measures. They are often less experienced in writing easy-to-use, easy-to-read, reusable, maintainable code. As a result, not all research software reaches its full potential. As Research Software Engineer at Statistics Netherlands, it is my mission to improve the value of our research software. We are implementing a program based on a standard git workflow with code reviews, that focuses on coaching and learning from each other on the job. By promoting open communication and collaboration with more experienced developers, instead of enforcing courses and coding rules, the program maximizes the engagement of the researchers while minimizing the disruption to their work.

References

No References available

Review of (some) R tools for seasonal adjustment of high frequency data

Authors

- Anna Smyk (Statistics France (INSEE))

Abstract

High frequency data have been widely used in official statistics for several years, especially after the Covid-19 pandemic, which fueled the demand for infra-monthly indicators. Since then, it has moved from an experimental status towards standard production. Many infra-monthly indicators show (multiple) seasonal patterns and need to be seasonally adjusted. Therefore, well-established seasonal adjustment methods for monthly and quarterly data have been extended to meet the methodological requirements of high frequency data and additional ad hoc methods have been developed. With a wide range of algorithms and tools available today, the user is faced with a selection problem. We provide an empirical review of the available R tools and underlying algorithms for seasonal adjustment of high-frequency data, comparing them from different perspectives: quality of the seasonal adjustment process (residual seasonality, revisions, forecasting), possibility to implement methodological refinements: automatic outlier detection and parameter selection, time-varying calendar correction. We also examine the accessibility, usability and performance of the tools in a mass production approach.

References

No References available

Robust and fast data synthesis with the synthesizer package

Authors

- Mark Van Der Loo (Leiden University Medical Center (LUMC), Netherlands)
- Marije Sluiskes (Leiden University Medical Center (LUMC), Netherlands)
- Mishca Jacobs (Leiden University Medical Center (LUMC), Netherlands)
- Maria Anthoulaki (Leiden University Medical Center (LUMC), Netherlands)

Abstract

Synthetic data holds the promise of allowing one to share useful data while preserving the privacy of entities represented in the original dataset. For this reason, both the creation of synthetic data as well as methods for measuring and balancing privacy and utility for specific applications is currently of great interest in multiple research communities. Data sets in official statistics offer a unique set of challenges, both due to the nature of the data and due to the methods of observation. For example, economic datasets often demonstrate mixed (zero-inflated) distributions, and variables with strongly skewed distributions while linear correlations are high. Moreover, economic variables are often connected by mathematical and logical restrictions, while errors, outliers, and missing values are common. These complexities add up to multivariate distributions that are hard to reproduce. In this contribution we present a robust synthetic data methodology that is fast, easy to understand, and is capable of handling categorical, mixed, or numerical data with complex multivariate distribution. The method combines sampling from individual empirical distributions of the variables with rank order matching to reproduce (non)linear correlations. Missing value patterns are automatically taken into account. Finally, the method is implemented with the option to decrease rank correlation, optionally per variable, allowing users to smoothly move between high utility-low privacy and low utility-high privacy synthesis.

References

- [1] MPJ van der Loo M (2025). synthesizer: Synthesize Data Based on Empirical Quantile Functions and Rank Order Matching. R package version 0.4.0, https://doi.org/10.32614/CRAN.package.synthesizer

Robust estimation with survey data

Authors

- Tobias Schoch (University of Applied Sciences Northwestern Switzerland)

Abstract

Outlier detection and handling is a non-trivial task, even when the data are regarded as a random sample from an infinite population. In this context (i.e., classical statistics), outliers are typically considered to be generated by a model other than the one under study. Compared to classical statistics, outliers are a very different concept in finite population sampling. In the context of sampling (design-based inference), where no statistical model is assumed, outliers are extreme values that deviate from the bulk of the data. In addition, unlike in classical statistics, we also have to consider the sampling weights. Observations that are not considered outliers (i.e., that are in the bulk of the data) can still strongly influence an estimator due to their large sampling weight (influential values). An estimator or procedure is called (qualitatively) robust if it is resistant or insensitive to the presence of outliers and influential values. In principle, robust estimation can be implemented in two ways: i) detection and treatment of outliers, or ii) direct application of robust estimation techniques. We limit our attention to the latter approach. The robsurvey package implements: i) basic robust estimators of the mean and total (e.g., robust Horvitz-Thompson estimator), robust survey regression, and model-assisted estimation (e.g., robust generalized regression estimator, GREG). In the talk, we will take a look at some of the methods and illustrate them with examples from business surveys.

References

No References available

scimetr: An R Package for Bibliometric Assessment of Countries, Institutions, and Researchers Using WoS Data

Authors

- Borja Lafuente-Rego (MODES group, Universidade da Coruña, Spain)
- Rubén Fernández-Casal (MODES group, CITIC, Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, Spain)
- María José Lombardía-Cortiña (MODES group, CITIC, Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, Spain)
- Julián Costa (MODES group, Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, Spain)
- Javier Tarrío Saavedra (MODES group, CITIC, INDITEX-UDC Chair in Artificial Intelligence for Green Algorithms, Department of Mathematics, Ferrol Engineering Polytechnic University College, Universidade da Coruña, Spain)

Abstract

scimetr is a R package designed to support exploratory, comparative, and bibliometric analysis from scientific publications indexed in Web of Science (WoS). The package provides a set of tools for processing, analyzing, and visualizing bibliometric metadata, with a focus on accessibility, reproducibility, and integration into academic workflows. It facilitates the computation of key indicators such as production, visibility, impact, collaboration, and thematic structure. It includes functions to import WoS document-level metadata and, additionally, to incorporate JCR (Journal Citation Reports) impact measures. A relational database is generated and used to perform the different analyses and generate reports. The package is particularly suited for researchers and evaluators aiming to understand scientific trends, collaborative structures, or policy impacts through transparent, open-source analytics in R. It provides support for decision making in the framework of such as important institutions as universities and relative government organization. For example, to establish the funding of universities and institutions, this package allows to obtain indicators of their research activity in an open and transparent way. The development version of this package is currently available for download at https://rubenfcasal.github.io/scimetr/index.html, and we expect a stable version to be available soon on CRAN.

References

No References available

Seeing Red: Using R to Detect Survey Anomalies Before It's Too Late

Authors

- Thomas Delclite (Statistics Belgium (Statbel))

Abstract

Monitoring and quality control of field-collected data remain key challenges for national statistical institutes. While data cleaning is typically conducted after fieldwork is completed, by that time, it is often too late to correct anomalies that originate from the data collection process itself. However, many issues can be identified and potentially mitigated during the fieldwork phase. On a day-to-day basis, this involves ensuring that survey data are successfully transmitted to our databases. It is also essential to rapidly detect any unintended changes in questionnaires or data collection programs that could compromise the comparability of statistics over time. Furthermore, for face-to-face surveys, it is crucial to identify patterns of anomalies that may stem from the behavior or practices of interviewers themselves. We present an R Shiny dashboard developed to monitor and control data collection in real time. Drawing inspiration from exploratory data analysis (EDA) tools, this dashboard is tailored to the operational requirements of statistical institutes while remaining adaptable to a wide variety of survey formats. At the core of the application is a heatmap that highlights, for each survey variable, the discrepancies between the responses collected by individual interviewers and the overall distribution across all interviewers. This visual tool enables survey managers to quickly identify potential issues: the greater the deviation (such as high missing value rates, abnormal Chi-square test results for categorical variables, or significant differences in medians for continuous variables) the more intensely the corresponding cell is colored in red. This instant feedback allows users to spot unusual results at a glance and investigate them further. By clicking on a cell, the dashboard displays additional insights for both the selected interviewer and the overall survey population, enabling a deeper investigation into the potential anomaly. The dashboard also includes several complementary tabs to facilitate quality checks: automated comparisons of distributions across survey waves (e.g., year N vs. N–1), dynamic handling of missing values and exceptional cases, all computed on the fly using raw data files in CSV or any format supported by R. The dashboard has already been used successfully by Statbel for over six months on several major European surveys, including ICT Households, the Labour Force Survey, EU-SILC, and the Travel Survey. We are plan to release it as an open-source package, aiming to support quality assurance across European surveys, promote best practices in statistical monitoring, and foster the wider use of R in official statistics.

References

No References available

Small Area Estimation of Unemployment Rates Using Big Data: Integration of Labor Force Surveys with Geospatial Data

Authors

- Joseph Nyajuoga (Örebro University School of Business, Sweden)

Abstract

This paper explores the integration of satellite-derived geospatial data and administrative registers into a Small Area Estimation (SAE) framework to generate precise unemployment rate estimates at the county level in Sweden. Using the Fay–Herriot (FH) model combined with Empirical Best Linear Unbiased Prediction (EBLUP), the study assesses how effectively selected geospatial covariates such as nighttime lights, elevation, urban cover, NDVI, and temperature enhance the precision of unemployment estimates compared to traditional survey based estimates alone. Empirical results show that incorporating geospatial indicators reduces the mean squared error of estimates, confirming that satellite-derived data significantly enhances estimation accuracy. Model diagnostics support the validity of the FH approach in integrating geospatial covariates. The study also critically evaluates the practical considerations, including computational complexity and the potential limitations inherent to Big Data integration.

References

No References available

Spatial Interdependence of Firm Profitability in Romania: A Regional Econometric Analysis

Authors

- Horatiu-Gabriel Tibrea-Marcu (The Bucharest University of Economic Studies, Romania)
- Ioana-Diana Petre (The Bucharest University of Economic Studies, Romania)
- Marian Dardala (The Bucharest University of Economic Studies, Romania)
- Titus Felix Furtuna (The Bucharest University of Economic Studies, Romania)

Abstract

This study explores the spatial interdependence of firm profitability across Romanian administrative units, using a detailed economic dataset spanning 2008 to 2021, which includes the number of firms, number of employees, turnover, net profit, and net loss, aggregated by CAEN activity codes and linked to official SIRUTA territorial identifiers. Our main objective is to test whether the net profit in a locality depends not only on its own firm structure and turnover but also on the profit levels of neighboring localities; for example, whether profit levels in Cluj might be influenced by the economic performance of adjacent counties such as Alba or Mureș. The dependent variable is primarily the net profit, with net loss also explored for comparison. Explanatory variables include the number of firms, number of employees, and the turnover. The methodological steps involve: constructing a clean shapefile with valid SIRUTA codes; creating a spatial weights matrix to define neighboring relationships; testing for spatial autocorrelation in profit levels using Moran's I; fitting Spatial Autoregressive models to estimate spillover effects; and visualizing results through maps. All analyses and visualizations are performed using the open-source R ecosystem (sf, spdep, spatialreg, tmap), demonstrating a practical pipeline for spatial econometric analysis on real Romanian firm data. These results highlight the importance of accounting for spatial relationships in regional economic analyses and support evidence-based policies that encourage resilient and cohesive local development.

References

No References available

Standardizing variance estimation in statistical production processes

Authors

- Claude Lamboray (Statistics Luxembourg (STATEC))
- Lisa Borsi (Statistics Luxembourg (STATEC))
- Guillaume Osier (Statistics Luxembourg (STATEC))

Abstract

We present an internally developed R package aimed at standardizing variance estimation across statistical production workflows. Built on top of the gustave package, our solution enables the construction and application of modular variance wrappers that implement specific analytical variance estimation methodologies. The package streamlines the estimation process through a structured five-step workflow: (1) data preparation and validation, (2) specification of linearization formulas for complex estimators, (3) specification of the variance estimation methodology, (4) application of the methodology to compute estimates, and (5) automated generation of a Quarto report summarizing the results. Users input survey data, technical metadata (e.g., weights, calibration variables), and configuration parameters. The output includes a comprehensive set of uncertainty measures for target variables and estimators, optionally disaggregated by domain. A generic variance wrapper is included, suitable for a wide range of surveys conducted at STATEC. The package is designed to be extensible, allowing us to integrate additional wrappers to accommodate additional survey designs.

References

No References available

surveysd - Overview with a Focus on New Features

Authors

- Johannes Gussenbauer (Statistics Austria)
- Alexander Kowarik (Statistics Austria)
- Eileen Vattheuer (Statistics Austria)

Abstract

The R package surveysd offers functionality for estimating statistics from survey data with complex sample designs. It supports calibration of sample weights using iterative proportional fitting and enables the calculation of calibrated and appropriate bootstrap weights. The package provides a simple yet powerful interface for computing estimates based on these bootstrap weights. The highlight of the presentation will be recent enhancements to the package, particularly the addition of a new bootstrap method: Rao-Wu (Rao and Wu, 1988). This method offers a robust alternative to the default bootstrap approach proposed by Preston (2009).

References

- [1]Preston, J. 2009. "Rescaled Bootstrap for Stratified Multistage Sampling." Survey Methodology 35 (December): 227–34. https://www.researchgate.net/publication/281735659.
- [2]Rao, J. N. K., and C. F. J. Wu. 1988. "Resampling Inference with Complex Survey Data." Journal of the American Statistical Association 83 (401). Taylor & Francis: 231–41. https://doi.org/10.2307/2288945.

The awesome official statistics software landscape: state of play and future directions

Authors

- Olav Ten Bosch (Statistics Netherlands (CBS))
- Mark Van Der Loo (Statistics Netherlands (CBS))

Abstract

National Statistical Offices (NSOs) are increasingly adopting open-source tools to automate and operationalize the production of official statistics. This shift is driven by the potential to enhance transparency, improve efficiency, and foster reproducibility, and by the fact that many young professionals in statistics and data science enter the labor market with strong skills in these tools. The Official Statistics R community has played a key role in this change and will continue to do so, and this conference is an important driving force in this. The presentation gives an overview of the latest developments on the "Awesome List of Official Statistics Software," a valuable resource for the official statistics open-source community. This curated list, initiated in 2017, facilitates knowledge exchange among statistical organizations and lists mature and popular open-source tools available. The continued growth and popularity of the list has led the authors of this abstract to consider future directions that guarantee mature solutions, core features such as true independence of modules, propagation of uncertainty in process chains, a healthy software landscape, and other ideas. The presentation also outlines the principles guiding open-source adoption, currently under consideration for endorsement by the UNECE Conference of European Statisticians (CES), derived from best practices around software on the list. Finally, it discusses the need for advanced metrics to assess the maturity of open-source software modules for official statistics.

References

No References available

The use of R tools to synthetize microdata using classical and differential privacy methods

Authors

- Andrzej Młodak (Statistics Poland, Statistical Office in Poznań, Centre of Small Area Estimation; University of Kalisz, Inter-Faculty Department of Mathematics and Statistics, Poland)
- Tomasz Józefowski (Statistics Poland, Statistical Office in Poznań, Centre of Small Area Estimation; Poznań University of Economics and Business, Chair of Statistics, Poznań)
- Grzegorz Grygiel (Statistics Poland, Statistical Office in Poznań, Centre of Small Area Estimation)

Abstract

In this paper we will describe some applications of R tools to construct synthetic data. The possibility of efficient synthetization will be examined taking into account the risk of disclosure, i.e. possible re-identification of respondents, and the prospective use of synthetic data. We will compare results of a number of synthetization methods available in the synthpop R package (e.g. CART, regression models preserving statistical relationships) and a method of generating synthetic data based on differential privacy, which is implemented in the DPpack package. The efficiency of synthesis will be assessed using various measures of information loss, including those available in the sdcMicro R package. Our analysis is based, among others, on publicly available microdata from the 2019 Survey of Human Capital in Poland conducted biannually since 2009 by the Polish Agency for Enterprise Development (PARP) with the support of the Jagiellonian University in Kraków. The survey provides information for the Balance of Human Capital project, which monitors the demand for specific competences and skills gaps in different sectors. PARP publishes anonymized microdata from the survey.

References

No References available

Transitioning to R for Official Statistics: Lessons, Challenges, and Open Questions from the Flemish Statistical Authority

Authors

- Jorre Vannieuwenhuyze (Flemish Statistical Authority (Statistics Flanders), Belgium)

Abstract

The Flemish Statistical Authority (VSA) is responsible for producing the Official Statistics of Flanders. This responsibility calls for highly automated and standardized processes in statistical production. However, due to the organization's historical development, the VSA currently employs many staff members who have limited experience with programming languages and the statistical production process. Last year, the decision was made to adopt R as the core software for statistical production, complemented by SQL for database management and Python for data science applications. While this transition holds great promise, the implementation of R presents several challenges. Along the way, we have encountered many important questions, such as: - How should we structure our data model? Should we adopt complex existing European frameworks like SDMX, or would it be better to start from a blank slate and gradually develop our own model so that all colleagues can follow the logic and contribute to translating it into R? - How do we standardize production processes? To what extent can we reasonably expect colleagues to build R expertise—for instance, in using functions, mappings, or loops? - How do we define and demarcate what counts as Official Statistics, and how do we translate these distinctions into R-based workflows? - Which version or dialect of the R language should we adopt as our standard—should we work primarily with the tidyverse ecosystem, or opt for data.table? - Would it be worthwhile to train colleagues in R Markdown as well, so they can document their processes more clearly and professionally? - How can we design an effective human resources policy that builds a diversified set of R-related competencies within the team? In this presentation, I will share the choices and steps we have taken so far at the VSA. More importantly, I hope to spark a broader discussion with the audience that can help us clarify the path ahead. Their insights will be valuable as we work toward a smoother and more inclusive transition to R across the organization.

References

No References available

tRansparent deep/machine learners for official statistics

Authors

- Violeta Calian (Statistics Iceland)

Abstract

In this paper we show that deep/machine learning algorithms can be made more transparent and accountable by measuring, tuning and reporting on the uncertainty of their results and of their performance metrics. This is a particularly important objective for official statistics production where black box strategy is not recommendable and where errors as well as biases and limitations should be always described and controlled. Interpretability of deep/machine learning results has become more systematically reported in recent years. In contrast, the uncertainty around the performance of the algorithms or around the output results is rarely measured and declared. For official statistics though, one should avoid overly confident predictions and pointwise results or evaluations. To achieve this goal, we employ methods which are either analytical [1] and therefore not computationally but only theoretically demanding, data-driven [2] thus rather widely applicable or methods based on the Bayesian paradigm [3] thus reporting on uncertainty by construction. We argue that these methods can be used to deliver the transparency we seek, and we illustrate the idea with two concrete applications which are implemented in (open code) R [4], at Statistics Iceland. In the process, we use highly performant R-packages [5-7] in order to: (i) perform time series forecasting based on hierarchical GAMs with stochastic process priors and (ii) produce synthetic datasets for confidentiality protection and research with risk/utility tuning and disclosure scenarios evaluation.

References

- [1] Deep Neural Networks as Gaussian Processes, Lee, J. et al., https://arxiv.org/abs/1711.00165
- [2] Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models, Huang, Y. et al., https://arxiv.org/abs/2307.10236
- [3] Bayesian Deep Learning is Needed in the Age of Large-Scale AI, Papamarkou, Th. Et al., https://arxiv.org/abs/2402.00809
- [4] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- [5] https://cran.r-project.org/web/packages/brms/
- [6] https://cran.r-project.org/web/packages/mgcv/
- [7] https://cran.r-project.org/web/packages/mvgam/index.html

UnitMix: An R Package for Detecting and Correcting Unity Measure Errors via Gaussian Mixture Modeling

Authors

- Cristina Faricelli (Italian National Institute of Statistics (ISTAT))
- Renato Magistro (Italian National Institute of Statistics (ISTAT))

Abstract

Unity Measure Errors are a frequent and well-known issue in statistical data processing, but they remain challenging to detect and correct in a systematic way. These errors typically arise when data is reported using incorrect or inconsistent scales—for example, monetary values entered in euros instead of thousands, or quantities given in kilograms instead of grams. If left unaddressed, these anomalies can lead to misleading aggregates, bias in model estimates, and misinterpretation of key indicators. To address this challenge, we developed UnitMix, an R package designed for the detection, diagnosis, and correction of unity measure errors in multivariate numeric datasets. The approach is based on model-based clustering via Gaussian Mixture Models (GMMs), building on the method proposed by Di Zio et al. (2005), where users define a set of plausible error patterns expressed as translation vectors or scaling factors. The algorithm estimates the probability that each observation belongs to one of these patterns, and assigns it to a cluster only if two uncertainty conditions are met: a minimum posterior probability and a maximum normalized entropy. UnitMix offers a set of tools to assist analysts in each step of the workflow. These include digit-based comparisons to identify magnitude discrepancies and graphical tools for visualizing cluster separations and uncertainty levels. Once patterns are reviewed—optionally but preferably with domain expert input—the package can apply consistent row-wise corrections based on the cluster structure. To facilitate use in production environments, a Shiny web interface is also available. This allows users with no R programming experience to explore, validate, and correct measurement errors interactively. We illustrate the method through a case study on the Italian survey on energy consumption in enterprises. The approach is general-purpose and has been applied to other contexts, including agricultural and business statistics. While UnitMix offers a transparent and reproducible solution, it is not intended as a fully automatic tool: defining realistic error patterns is the most delicate part of the process and requires both analytical expertise and domain knowledge.

References

- [1]Di Zio, M., Guarnera, U., & Luzi, O. (2005). Estimating the impact of measurement errors on the distribution of economic data. Survey Methodology, 31(1), 53–63. https://www.istat.it/it/files/2014/05/Survey-Methodology-311-53-63.pdf

Using FH model for estimating AROP rate in small areas in Croatia

Authors

- Ivana Levačić (Croatian Bureau of Statistics)
- Lidija Gligorova (Croatian Bureau of Statistics)

Abstract

Sample surveys are known as effective means of obtaining information on the population of interest in terms of reduced cost and greater speed. However, the sample sizes may not be large enough to support the production of estimates of acceptable precision for subpopulations (referred to domains or areas). To overcome this problem, small area estimation methods are used. These methods "borrow strength" by using values of the variable of interest from related areas and/or time periods and thus increase the "effective" sample size. Small area models are classified into two broad types: area-level models and unit-level models. In off-census years, the typical small area approach applied has been an area-level model, such as a Fay-Herriot. As poverty reduction is one of the main goals and longstanding commitment, accurate and timely data at the subnational level is needed in order to help policymakers in adopting right policies. For that purpose, the Fay-Herriot (FH) model is used to estimate at-risk-of-poverty (AROP) rate in Croatia using 2023 SILC data and auxiliary variables from different sources at the level of municipality.

References

No References available

Using the EurostatRTool for generating Turkiye's Human Development dashboard

Authors

- Antonio Grosso (Eurostat, Unit C.1 – Macroeconomic Indicators, Luxembourg)
- Rosa Ruggeri Cannata (Eurostat, Unit C.1 – Macroeconomic Indicators, Luxembourg)
- Fethi Saban Ozbek (Turkish Statistical Institute)

Abstract

This study presents an application of the EurostatRTool, an open-source R package developed under the European Statistical system Innovation Agenda, to generate an interactive dashboard for the Türkiye's Human Development Index (HDI). Originally designed for disseminating short-term European macroeconomic indicators, the tool proved to be a reproducible, low-code solution for creating interactive visualisations tailored to national and regional contexts. Developed through a collaboration between Eurostat and the Turkish Statistical Institute (TurkStat), the application displays the HDI and its key sub-indicators - education, income, and life expectancy, across Türkiye's NUTS-3 regions for the period 2018 to 2022. The tool's modular design allowed for an easy integration of Turkish datasets, language labels, institutional logos, and colour schemes, requiring only minimal adjustments to the source code. The tool is easily updatable to incorporate the most recent data, without requiring any expertise in R. The dashboard development started with the uploading of structured input files (data, layout, labels), followed by the configuration of the core presentation elements (titles, metadata, colour themes), and ending with the generation of the final HTML dashboard by running a single function only. The selected visualisation modes are timeline plots, comparative bar charts and tables. The tool's output is a set of html pages not requiring access to a database, minimizing then risks of security infringement. This implementation demonstrates the tool's ability to simplify the creation and maintenance of complex interactive dashboards without significant resource requirements. It enables TurkStat to independently update the dashboard for future HDI releases and shows how the EurostatRTool can support sustainable, open-source statistical dissemination at national level, or even in more general contexts. The source code is freely available on GitHub; giving access to shape your own dashboard, fitted to users' needs.

References

No References available

Validating data on education finance using R: a user-friendly tool for national statistical offices

Authors

- Viktoria Kis (Directorate for Education and Skills (OECD))
- Simon Normandeau (Directorate for Education and Skills (OECD))
- Erika Lee (Directorate for Education and Skills (OECD))

Abstract

Each year over 50 countries submit data on education expenditure as part of the Unesco-OECD-Eurostat (UOE) finance data collection, based on Excel questionnaires. The validation process traditionally involves many e-mails between national statistical offices and international organisations to resolve quality issues. To streamline this process, we have developed an R package that allows national experts to run our checks and correct problems before submission. In our talk, we will speak about the design choices we made to make our package easy to use for people with varying levels of R expertise, and easy to adapt and maintain by our team. Our solution offers a simple user interface, that requires minimal coding, with concise messages guiding users through each step. Feedback is written to an Excel file, alongside key indicators based on current and historical data, helping to identify unusual patterns. From a development perspective, we focused on ensuring that validation rules can be easily updated. Given the diversity of rules (from basic arithmetic checks to more complex relational validations) the rules are defined in a configurable .csv file. This file includes both the mathematical logic and domain-specific descriptions, allowing rules to be added or modified without changing the core code.

References

No References available

vimpute(): An extension of the imputation methods in the VIM package

Authors

- Eileen Vattheuer (Statistics Austria)
- Nina Niederhametner (Statistics Austria)
- Alexander Kowarik (Statistics Austria)
- Johannes Gussenbauer (Statistics Austria)

Abstract

The vimpute() function enhances the VIM package by introducing a solution for handling missing values through a sequential imputation approach powered by machine learning. This method leverages advanced algorithms such as random forests and XGBoost from the mlr3 ecosystem to accurately estimate missing values. Imputation is performed iteratively, addressing each variable with missing data step by step while incorporating all available information from the dataset. A feature of this approach is the integrated hyperparameter optimization, which selects the optimal model settings for each variable, enhancing prediction accuracy. For numerical variables, predictive mean matching ensures not only accurate imputation but also the preservation of the original data distribution. For categorical variables, a stochastic imputation technique based on predicted class probabilities generates particularly realistic results. The entire process is iterative, refining imputations over multiple cycles until a stable solution is reached. The method is presented in realistic and simulation setups.

References

No References available

XML Generator R-Shiny Web App: Using Case in Central Bank of Malta

Authors

- Ibrahim Cagan Kaya (Central Bank of Malta)

Abstract

The Central Bank of Malta currently employs a tool for the collection of statistical data from reporting entities. In parallel, the Bank also utilizes structured CSV files for data revision and direct ingestion into its internal systems. Given that the collection platform accepts only XML files conforming to a predefined schema, there is a critical need for a reliable and user-friendly solution to convert CSV data into compliant XML format. This study presents the development of a web-based XML Generator application using the R programming language and the Shiny framework. The application enables users to upload structured CSV files, preview the data, and generate XML documents that adhere strictly to the required schema. The backend leverages R's utils package for efficient CSV parsing and the XML and xml2 packages for constructing well-formed XML. The user interface, built with bslib for enhanced usability, supports real-time validation, customizable output file naming, and instant XML download. By automating the conversion process, the application eliminates the need for manual XML coding, reduces the risk of formatting errors, and significantly improves the efficiency of data integration workflows. This tool is particularly valuable in official statistics contexts, where data accuracy and schema compliance are paramount.

References

No References available

Author Index

Affiliation Index