Book of Abstracts (2024)

uRos (Use of R in Official Statistics)
27-29 November 2024, Athens, Greece

Contents

A Neural Network Approach to Text Classification for International Standardized Codes

Authors

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

Abstract

International standard classifications such as ISCO (for Occupation), ISCED (for Education), and COICOP (for Consumption) serve as pivotal statistical frameworks, facilitating the organization and categorization of data. In official statistical practices, adherence to these codes is essential for thorough analysis and comparison of findings. Survey respondents typically provide information in an unstructured free textual format, requiring subsequent assignment to standardized codes, a task that is traditionally performed manually, demanding significant time and effort. In this talk, we present our approach that automates the classification of textual data into various standardized codes using simple mathematical techniques combined with neural network-based language models, utilizing the R libraries TensorFlow and Keras. Additionally, we illustrate the development of application programming interfaces (APIs) using plumber, and the deployment of our models, establishing accessibility to a broad user base.

References

No References available

A practical example for embedding R into Eviews-programming

Authors

- Michel Geller (STATEC, Luxembourg)

Abstract

We integrated Eurostat's R package into our EViews data-handling framework, resulting in data retrieval and processing speeds up to 100 times faster than those achieved by the native Eurostat connector in EViews. This integration retains EViews' interface and most of its code, keeping it accessible to users without R programming experience. We adopted this approach because our data-handling environment relies on a legacy system built with the EViews programming environment, and replacing it entirely would require considerable time and resources. Generally, we integrate R code step-wise into our EViews environment, enabling a gradual and controlled transition to an R-based framework in the future.

References

No References available

A process for making imputations in official statistics

Authors

- Claude Lamboray (STATEC, Luxembourg)
- Johann Neumayr (STATEC, Luxembourg)

Abstract

When working with survey data, statisticians often need to address missing or erroneous values by imputing them with plausible data. These imputations can be derived from auxiliary variables that are available within the data set. One widely used tool for this purpose is the R package mice. We propose to embed this package into a generic process that can be adapted to different surveys and (target and auxiliary) variables. The entire process is structured into distinct steps that are conducted one after the other. • Step 1 - Metadata: Involves compiling and storing important information about each variable in the actual data set, providing a foundational framework for subsequent steps. • Step 2 - Data Preparation: Refines the data in line with metadata guidelines, including selecting relevant rows and columns as well as formatting data classes and missing values. • Step 3 - Modeling: Entails selecting variables for imputation models, starting with basic imputation of missing data in predictor variables and progressing to automatic variable selection using Random Forest. • Step 4 - Methodology: Establishes the imputation methodology, defining different techniques for different types of variables. • Step 5 - Imputation: Executes the imputation process using the developed models and methodologies. • Step 6 - Evaluation: Evaluates the imputation process through visualizations and assessments, comparing original and imputed data to identify areas for refinement. • Step 7 - Reporting: Consolidates the outcomes of preceding steps into a comprehensive report. The process can be run iteratively, allowing for the development of an appropriate imputation design. We illustrate the process using data from the Labour Force Survey, focusing on the imputation of the variable INCGROSS (gross income).

References

No References available

Access to official statistics from R: part II

Authors

- Olav ten Bosch (Statistics Netherlands, The Netherlands)
- Edwin de Jonge (Statistics Netherlands, The Netherlands)

Abstract

Providing smooth and easy access to official statistics is of utmost importance for Statistical Institutes and their users. In previous years we analysed the software landscape currently available to end-users to access official statistics data and metadata from R [1] and broader in a FAIR (Findable, Accessible, Interoperable, Reusable) perspective [2]. These analyses were based on the awesome list of official statistics software [3], which contains more than 30 software packages in this category. One of the conclusions was that there is currently no 'one-for-all' software package for access to all official statistics. In addition, a number of generally features were identified from the available software packages that could be used as input for such solution. In this talk we go one step further. Based on an update of the earlier analysis, we design a preliminary setup where the features identified can be offered on as many data providers as possible with a maximum re-use of existing solutions and a minimum of additional maintenance. A smoke test of relevant software packages from the awesome list is part of the work. This will help decide on the feasibility of different design choices that can be made. In this presentation we will present the results, the successes and thoughts on a possible ways forward, with a special focus on FAIRness of official statistics for the R user.

References

- [1] Bosch, O. ten, Jonge, E., Access to official statistics from R: an overview, uRos 2023, Bucharest
- [2] Bosch, O. ten, Jonge, E., Laloli H. To be FAIR, what is missing in Official Statistics?, Conference On Smart Metadata for Official Statistics (COSMOS 2024), April, 11-12, 2024, Paris, France
- [3] http://awesomeofficialstatistics.org

Automatic Classification of Economic Activities with NOGAuto:

Exploring Rules for Use in Statistical Production

Authors

- Athanassia Chalimourda (Swiss Federal Statistical Office, Switzerland)
- Mathias Constantin (Swiss Federal Statistical Office, Switzerland)

Abstract

In previous work presented at the use of R in official statistics 2023 conference, we assessed the integration of NOGAuto, an assistance system for automatic classification of economic activity descriptions in statistical production. NOGAuto is developed by the Business Registers Data section with the methodological support of the Statistical Methods section of the Swiss Federal Statistical Office. It assigns activity descriptions in different languages to categories according to the Swiss NACE version, NOGA, with a certain probability. It performs Natural Language Processing and automatic text classification with a Gradient Boosting Machine (GBM) and follows the hierarchical structure of NOGA with the GBM proceeding to the next NOGA level for the most probable predictions. Connecting the NOGAuto Shiny application to the API of the DeepL automatic translation service allows activity descriptions in other languages, such as German, one of the Swiss national languages, to be translated into French, the language in which the NOGAuto models were trained. While in our previous work we highlighted the challenges of integrating an innovative system into statistical production and discussed possible solutions, in this work we take a further step towards production by evaluating and adjusting the classification process on a large validation set. Using a large validation set allows detailed analysis of aspects such as precision and distributional accuracy of predictions. We are also exploring rules for when to assign a code to an activity description in a fully or semi-automated mode. To this end, we combine the by- class performance measure of precision with the probability of the predicted code. Precision measures the proportion of correctly predicted elements out of all elements predicted to a given NOGA class. Defining lower bounds as thresholds on the prediction probabilities of a given class results in subsets of that class having a higher precision. For example, NOGAuto could automatically advance to the next NOGA level for a predicted code with probability higher than the class threshold, while guaranteeing a certain level of precision. If the threshold value is not reached, the coding expert has to assign the next NOGA level manually, if necessary, by consulting an internal SFSO rule-based system linked to the NOGAuto Shiny application. Throughout the classification process, the expert can review the automatically assigned code, comment on it and, in case of disagreement, register a different code. Either in a fully or semi-automated mode, alone or in combination with other expert systems, this process would allow the expert to code in a timely, convenient and efficient manner, while retaining full control of the coding result.

References

No References available

Automatic validation in R for EU-SILC microdata

Authors

- Margherita Zuppardo (Statistics Iceland)

Abstract

We will discuss in some detail the ongoing development of a validation script designed in R for EU-SILC at Statistics Iceland. The script uses open source R packages in order to access various sources of documentation emitted by Eurostat, that are designed to be read by humans. The EU-SILC survey (European Union Statistics on Income and Living Conditions) is a critical information source on social inclusion and living conditions, both internationally and specifically for Iceland. Despite its importance, meeting Eurostat‘s deadlines and quality standards has been challenging for Statistics Iceland, mainly because of the limited resources and small team. To address these issues, we transitioned our data processing from SQL to R, in order to take advantage of the software's capacity for automation, through extensive use of functions and open-source packages. At this stage, most of the data processing runs automatically every year, although some work is still needed in this direction. One obstacle to the full automatization is that the EU-SILC questionnaire is revised annually, and slight changes to the data format and metadata (flags) make it necessary to review the entire documentation on a yearly basis. This slows down the process and introduces human error into the automatic flow. As a solution, we are developing a flexible data validation tool in R, specific to EU-SILC, and designed to require minimal updates. This tool leverages open-source packages for data mining such as pdftools [1], readxl [2] and uses stringr [3] for text manipulation, reading the different sources of documentation that Eurostat provides. The format and rules, that each variable and its flag are required to follow, are read from the various sources and automatically translated into a R script. This script verifies whether the rules are respected, and counts possible errors. An independent report is printed at the end. This tool is already in use in our team. However, it is not complete, as there are some classes of errors that are not implemented yet. In the future, incorporating large language models (LLM) into the process would help make the tool more flexible and adaptable to small changes in the terminology used in the documentation.

References

- [1] cran.r-project.org/web/packages/pdftools/
- [2] cran.r-project.org/web/packages/readxl/
- [3] cran.r-project.org/web/packages/stringr/

Automating Data Validation on SQL Server Using R and the Machine Learning package

Authors

- Paula Hartung (Statistics Iceland, Iceland)

Abstract

Ensuring data integrity and accuracy is vital for automatically collected data in SQL Server databases. Leveraging Microsoft SQL Server Management Studio's machine learning package with R code and the validation package provides a robust solution for scheduled data validation. At Statistics Iceland, we have developed a workflow to implement automated validation directly within the SQL Server. Our goal is to demonstrate the efficiency, accuracy, and scalability of this approach. The Data Technology team is working project based and is focused on creating adaptable solutions for automating data validation. This initiative aims to standardize validation processes across departments, improving data quality consistently. Previously, measuring the quality of automatically collected data, especially from API sources, was challenging. By fully utilizing R for direct validation on database level and scheduling these tasks, we have found a scalable solution beneficial for all departments at Statistics Iceland. We faced several obstacles, including R version control, package management, resource management, independence from personal computers, and language package issues. We continue to work on resolving these challenges. Our automated data validation framework, developed using SQL Server's machine learning services with R integration, includes: 1. Data Collection: Automatically collecting data from various sources into the SQL Server database. 2. R Integration: Writing R scripts within SQL Server Procedures to perform data validation using a specific validation package (e.g., validate from awesome package). 3. Scheduling: Setting up SQL Server Agent jobs to automate the execution of R scripts at defined intervals. The used R packages include validate [1], xlsx[2] to produce validation files to return to data providers and DBI [3] and odbc[4] to load data from the SQL tables and write reports, respectively. Testing this framework on a high-impact dataset revealed key improvements: 1. Significant reduction in data validation time compared to manual methods. 2. Implementing of additional validation checks based on Eurostat guidelines. 3. Improved accuracy in detecting data anomalies and inconsistencies. 4. Successful scheduling of validation tasks, ensuring continuous data quality without manual intervention. Integrating machine learning with R code in SQL Server for automated data validation offers a robust and scalable solution for real-time data integrity management. This approach not only enhances the efficiency and accuracy of data validation but also provides a reliable mechanism for continuous data quality assurance. Future work includes exploring machine learning models to identify data patterns and anomalies.

References

- [1] cran.r-project.org/web/packages/validate/
- [2] cran.r-project.org/web/packages/xlsx/
- [3] cran.r-project.org/web/packages/DBI/
- [4] cran.r-project.org/web/packages/odbc/

blocking: an R package for blocking of records for probabilistic record linkage based on approximate nearest neighbors algorithms

Authors

- Maciej Beręsewicz (Department of Statistics, Poznań University of Economics and Business, Poland Centre for the Methodology of Population Studies, Statistical Office in Poznań, Poland)

Abstract

Blocking is a crucial aspect of probabilistic record linkage studies in official statistics. The objective of this procedure is to reduce the number of possible comparison pairs (Steorts et al. 2014). This procedure assumes that blocking variables, such as sex, birth dates, or country of origin, are free of errors. However, in practice, such variables are often observed with typos or missing data.

The objective of the blocking package is to enable R users to block records using approximate nearest neighbors (ANN) algorithms and graphs. The blocking package has three primary objectives: (1) to significantly reduce the number of comparison pairs, (2) to account for the possibility that blocking variables may be measured with errors, and (3) to speed up the blocking procedure by employing advanced ANN algorithms.

The blocking package is based on the rnndescent (Melville 2024a), RcppHNSW (Melville 2024b), RcppAnnoy (Eddelbuettel 2024), and mlpack (Curtin et al. 2023) packages, which implement state-of-the-art ANN algorithms. The igraph (Csárdi & Nepusz, 2006, Csárdi et al. 2024) package is used for the creation of blocks. Moreover, the package facilitates straightforward integration with the reclin2 (van der Laan, J. 2024)) package and allows for the assessment of the quality of the blocking procedure.

According to our knowledge this is the only R package that aims at blocking of records using ANN algorithms . The package is currently under development and can be installed from https://github.com/ncn-foreigners/blocking.

Acknowledgements: The blocking package is produced as part of the project ‘Towards census-like statistics for foreign-born populations – quality, data integration and estimation’ funded by the National Science Centre grant OPUS 20 (2020/39/B/HS4/00941).

References

- Curtin, R., Edel, M., Shrit, O., Agrawal, S., Basak, S., Balamuta, J., Birmingham, R., Dutt, K., Eddelbuettel, D., Garg, R., Jaiswal, S., Kaushik, A., Kim, S., Mukherjee, A., Sai, N., Sharma, N., Parihar, Y., Swain, R., & Sanderson, C. (2023). “mlpack 4: a fast, header-only C++ machine learning library.” Journal of Open Source Software, 8(82) [doi: 10.21105/joss.05026]. Retrieved from https://doi.org/10.21105/joss.05026
- Csárdi, G., & Nepusz, T. (2006). “The igraph software package for complex network research.” InterJournal, Complex Systems, 1695. Retrieved from https://igraph.org

- Csárdi, G., Nepusz, T., Traag, V., Horvát Sz, Zanini, F., Noom, D., & Müller, K. (2024). igraph: Network Analysis and Visualization in R [doi: 10.5281/zenodo.7682609]. Retrieved from https://CRAN.R-project.org/package=igraph
- Eddelbuettel, D. (2024). RcppAnnoy: 'Rcpp' Bindings for 'Annoy', a Library for Approximate Nearest Neighbors [R package version 0.0.22]. Retrieved from https://CRAN.R-project.org/package=RcppAnnoy
- van der Laan, J. (2024). reclin2: Record Linkage Toolkit [R package version 0.5.0]. Retrieved from https://CRAN.R-project.org/package=reclin2
- Melville, J. (2024a). rnndescent: Nearest Neighbor Descent Method for Approximate Nearest Neighbors [R package version 0.1.6]. Retrieved from https://CRAN.R-project.org/package=rnndescent
- Melville, J. (2024b). RcppHNSW: 'Rcpp' Bindings for 'hnswlib', a Library for Approximate Nearest Neighbors [R package version 0.6.0]. Retrieved from https://CRAN.R-project.org/package=RcppHNSW
- Parihar, Y. S., Curtin, R., Eddelbuettel, D., & Balamuta, J. (2024). mlpack: 'Rcpp' Integration for the 'mlpack' Library [R package version 4.3.0.1]. Retrieved from https://CRAN.R-project.org/package=mlpack
- Steorts, R. C., Ventura, S. L., Sadinle, M., & Fienberg, S. E. (2014). A comparison of blocking methods for record linkage. In Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings (pp. 253-268). Springer International Publishing.

Detection and visualization of outliers in establishment surveys

Authors

- Valentin Todorov (UNIDO (retired))

Abstract

In a typical establishment survey a large number of variables (300 or more) are recorded. It is common that the data collected may contain outliers and be incomplete (missing values and non-response), usually skewed, and semi-continuous distributions occur often. In general, severe outliers may be relatively rare, but when they occur, they may have a dramatic impact on the estimators in cells defined by the intersection of geographical regions and industries. It is therefore important to identify them during the data editing process, prior to aggregation and analysis of the data. The multivariate aspect of the data collected in business surveys makes the task of outlier identification particularly challenging. The outliers can be completely hidden in one or two dimensional views of the data. Not many studies of outlier detection methods for business survey data based on robust methods are found in the literature, and the two studies that we can mention present two R packages ('rrcovNA' and 'modi') implementing several reliable methods for outlier detection. We study several new outlier detection methods. In cases when the number of variables is larger than the number of observations most of the standard multivariate methods cannot be used and some kind of regularization is necessary. If the survey is periodical and several data sets from past surveys are available, robust time series methods can be applied. The new methods, added to the R package 'rrcovNA' will be described and their performance will be investigated using the real data set included in the 'modi' package as well as a data set based on a survey conducted with the support of UNIDO. In a simulation study, where a subset of this survey is simulated, we compare several approaches. Since all methods are implemented in R, we can use the new package OutliersO3 to comfortably visualize their performance with the Overview Of Outliers (O3) plot.

References

No References available

Encouraging use of R at NSTAC Japan

Authors

- Ichiro Murata (National Statistics Center, Japan)

Abstract

Based on its global community and advantage as an open source software, R has been continuously improved and keeps being a practical and modern tool for statistics production and analysis. Acquiring R would be fruitful and give a good opportunity to broaden one's statistical perspective for any staff engaging in official statistics. National Statistics Center (NSTAC) is an administrative organization in charge of several statistical works such as producing statistics from collected questionnaires or data, managing the system disseminating statistics, and promoting statistical data utilization. Although these can be conducted through an appropriate support of a statistical software, we can find only a few workflows that include a process of using R in our organization and we have only a small number of R users. These years, considering the condition, we have started trying to make R more common in our work by taking some actions. For example, we have prepared a shared Rprofile to change the default R's undesirable behavior and to set the path to the package library at startup process, that are certain best practices while most beginners don't know how to handle them. We also have arranged a dataset converted from existing CSV data from official statistics, that staffs are relatively familiar with compared to the one from `datasets` package. The dataset contains some 2-dimentional data and formatted in Rdata for instant use. It helps our R introductory document showing examples of using some R functions from the view point of national statistics officer. Furthermore, we have planned and held internal lectures with interaction to provide more staffs with concise R experience. This presentation will describe these activities performed at an organization of official statistics.

References

No References available

Estimating equivalized income for school students in Austria

Authors

- Dominik Ernst (Statistics Austria)
- Johannes Gussenbauer (Statistics Austria)

Abstract

The socio-economic background of students is an important indicator for analyses, e.g. pertaining to the educational success of these students. While statistical registers in Austria contain information on individuals' income, this information cannot always be directly linked to parents of school students. This problem predominantly affects poorer demographics and would therefore incur a bias in analyses on school students.

As a remedy we model the equivalized personal income of school students using a combination of sample data from EU-SILC and register data. This way we obtain income estimates for almost every student. Another advantage of the equivalized income is that, for example, household composition and rent prices are incorporated in the estimates.

Lastly we show two applications where these income estimates were used in practice and finish with a short discussion and future developments.

References

No References available

Estimation of population parameters in EL-SILC with the R

packages: the experience of ELSTAT

Authors

- Irene Sarantou (Hellenic Statistical Authority (ELSTAT), Greece)
- Anastasia Mnimatidou (National Technical University of Athens, Greece)

Abstract

The European Union (EU) Statistics on Income and Living Conditions (SILC) is one of the most important data sources focusing on income distribution, poverty and social exclusion. EL-SILC is a regionally stratified two-stage annual survey with rotating panels whose target population consists of all private households and their current members. Regulation (EU) N° 2019/1700 enforced the need to accelerate the collection, processing and dissemination of re- sults and specific indicators while, at the same time, ensuring and increasing their quality and comparability. On the other hand, the use of R software has been widely increased in official statistics (Templ & Todorov, 2016). The use of an open-source programming language and software environment R is expected to provide a comprehensive, simplified and effective pro- cedure using flexible and reusable code. Although various well-established statistical software packages are often available, R seems to be the most economical solution (Templ & Todorov, 2016). In this framework, this study builds on previous studies (EUROSTAT, 2023; Merkouris, 2018; Verma, Betti, & Ghellini, 2017; Templ & Todorov, 2016) and intends to contribute to bridging a gap in estimation of population parameters using entirely R software. In this context, the aim of this presentation is to describe the benefits and lessons learned from the full implementation of the R code at different stages of the estimation procedure used by ELSTAT, at both national and regional levels. Key benefits include improved timeliness, reduced costs, and compliance with the new stringent requirements of the IESS Regulation. Particular emphasis will be placed on the transition from the SAS macro “Calmar” (bounded “logit” method) to the R package “icarus”. This marks the first complete implementation of the EL-SILC procedure using R, utilizing data from the EL-SILC 2023 survey. Further enhancements in efficiency are anticipated through communication and collaboration with other NSIs that also use R for their analyses.

References

- 1] EUROSTAT. (2023). Methodological Guidelines and Description of EU-SILC target variables,
- 2023 operation (Version 6: Draft), Eurostat.
- 1
- [2] Merkouris, P. (2018). Study of the current sampling design of the Survey of Income and Living
- Conditions with the objective to increase/adjust the sample at regional (NUTSII) level (Part I
- & II), Athens: AUEB Research Center.
- [3] Merkouris, T. (2001). Cross-Sectional Estimation in Multiple Panel Household Surveys,

- Canada: Statistics Canada, Vol. 27, No. 2, pp. 171-181.
- [4] Templ, M., & Todorov, V. (2016). The Software Environment R for Official Statistics, Austrian
- Journal of Statistics(45), 97–124.
- [5] Verma, V., Betti, G., & Ghellini, G. (2017). Cross-sectional and longitudinal weighting in a
- rotational household panel: Applications to EU-SILC, Statisticsin transition-new series, 8(1),
- 5-50.

Interactive visualizations of very large datasets using {shiny} and Base R

Authors

- Rahul Sangole (USA)

Abstract

In this talk, I will share my experiences combining Base R graphics and Shiny to create dynamic and responsive data visualizations. I'll present how I take advantage of the simplicity and flexibility of Base R graphics to handle and visualize huge amounts of data, while using Shiny's interactive features to enhance the user experience and explore the data. Drawing from my own projects, I will cover essential techniques for combining Base R graphics with packages like data.table and arrow to efficiently process and render large datasets, ensuring smooth and responsive interactions. While libraries like Plotly offer great interactivity, they do face limitations when plotting a large number of datapoints. Through real-world examples and live demonstrations, I will provide practical insights into building interactive dashboards and applications that can handle the complexities of large data without compromising on performance. This session is designed for data scientists, analysts, and R enthusiasts looking to expand their toolkit for interactive data visualization.

References

No References available

Is statistical matching feasible?

Authors

- Marcello D’Orazio (Italian National Institute of Statistics (Istat), Rome, Italy)

Abstract

Statistical matching (SM) refers to a wide range of statistical methods for integrating data from different sample surveys that refer to the same target population but investigate different phenomena (D’Orazio et al, 2006). The integration aims at analysing the relationship between the different phenomena by exploiting the information shared by the two surveys (common variables). The Italian National Institute of Statistics (Istat) has a long experience in the field of SM, mainly applied to the study of the relationship between income and expenditures of Italian households (see e.g. Donatiello et al. 2022). Recently, in collaboration with experts from the Bank of Italy, the objective has been extended to the estimation of the joint distribution of income, consumption and wealth (ICW) in Italy, also to feed Eurostat’s experimental statistics on this topic (see Balestra and Oehler, 2023)

Integration by statistical matching is based on a number of assumptions. The most problematic is the assumption of independence between the target variables (e.g. income from the income survey and expenditure from the household budget survey), conditional on the subset of common variables shared by both surveys that are being integrated. This assumption is quite strong and rarely valid; it cannot be tested with the available data, but we can get an idea of how it affects the results by exploring the uncertainty in the underlying model caused by the lack of data to estimate the parameters (e.g. correlation/association between the target variables). This paper provides some guidance on how to roughly assess uncertainty in R (D’Orazio, 2024) using “traditional” approaches (e.g. linear regression or multinomial models) or modern approaches such as random forest. This way of working represents a valid tool to evaluate the feasibility of statistical matching; in fact, a high uncertainty indicates that it is preferable to avoid additional effort to perform it; conversely, a low uncertainty indicates that statistical matching methods can be profitably applied.

References

- Balestra, C. and F. Oehler (2023) “Measuring the joint distribution of household income, consumption and wealth at
- the micro level. Methodological issues and experimental results. Edition 2023”, Statistical Working Papers,
- Eurostat/OECD, Luxembourg.
- https://ec.europa.eu/eurostat/web/products-statistical-working-papers/w/ks-tc-22-003
- D’Orazio, M. (2019) “Statistical learning in official statistics: The case of statistical matching”. Statistical Journal of the
- IAOS, 35(3), pp. 435-441. DOI: 10.3233/SJI-190518
- D’Orazio, M. (2024) “StatMatch: Statistical Matching or Data Fusion”. R package version 1.4.2. https://CRAN.R-
- project.org/package=StatMatch

- D'Orazio, M and Di Zio, M and Scanu, M (2006b) Statistical Matching: Theory and Practice. Wiley, Chichester
- Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2014) "Statistical Matching of Income and
- Consumption expenditures". International Journal of Economic Science, Vol. III (No. 3), pp. 50-65.
- Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016b) "The statistical matching of EU-SILC
- and HBS at ISTAT: where do we stand for the production of official statistics". DGINS - Conference of the Directors
- General of the National Statistical Institutes, 26-27 September 2016, Vienna.
- Donatiello G., D'Orazio M., Frattarola D., Spaziani M. (2022) "The joint distribution of income and consumption in Italy:
- an in-depth analysis on statistical matching". Rivista di Statistica Ufficiale - Review of Official Statistics, N. 3/2022,
- pp. 77-109
- Rodgers, W.L. and DeVol E.B. (1982) "An evaluation of statistical matching". Report Submitted to the Income Survey
- Development Program, Dept. of Health and Human Services, Institute for Social Research, University of Michigan

Joint calibration estimators for totals and quantiles for probability and nonprobability samples

Authors

- Maciej Beręsewicz (Poznań University of Economics and Business, Poland)
- Marcin Szymkowiak (Statistical Office in Poznań, Poland)

Abstract

Calibration weighting is a method commonly used in survey sampling to adjust original design weights for sampled elements to reproduce known population totals for all auxiliary variables. Following the calibration paradigm, it can also be used to reproduce known population quantiles for all benchmark variables. This technique is also used in surveys to compensate for nonsampling errors, such as nonresponse or coverage errors. By appropriately adjusting the weights, it is not only possible to ensure consistency with known structures for key variables from other data sources, such as censuses or registers, but also to reduce the bias and improve the precision of final estimates. Calibration weighting is also used in surveys in which the analysed features have asymmetric distributions (in the presence of outliers) to compensate for their negative impact on the final estimates: calibration weights provide robustness while meeting constraints on the calibration variables and the weights. Finally, thanks to the newest trends in the calibration approach, it can be applied to surveys with non-probability samples. In this presentation, we propose a joint calibration approach to estimate the total or any quantile (for instance median) for the variable of interest. Final calibration weights reproduce known population totals and quantiles for all auxiliary variables. The proposed method is based on the classic approach to calibration and simultaneously takes into account calibration equations for totals and quantiles of all auxiliary variables. In this presentation we also consider the use of the proposed approach to extend existing inference methods for non-probability samples, such as inverse probability weighting. Our simulation study has demonstrated that the estimators in question are more robust against model mis-specification and, as a result, help to reduce bias and improve estimation efficiency. The proposed approach has been implemented in a new R package jointCalib, which was used to conduct the simulation study.

References

No References available

Leveraging R for Official Statistics: A Case Study Using Traffic Count Data

Authors

- Dr Nele van der Wielen (Central Statistics Office (CSO), IE)

Abstract

The COVID-19 crisis showed the critical need for timely data to inform policy-makers and the public effectively. To meet the increasing demand for timelier data, the Central Statistics Office (CSO) used traffic counter data from Transport Infrastructure Ireland (TII), to provide up-to-date national transport statistics. This traffic counter dataset encompasses daily traffic counts across all vehicle types, presenting a rich source of information for statistical analysis. By harnessing the capabilities of R, the CSO was able to rapidly exploit this novel big data source. This initiative marked a significant milestone in the integration of innovative data sources, showcasing the power of R for data analysis. The most innovative aspect of this project was the seamless incorporation of big data into the production of official statistics, achieved for the first time on this scale within the CSO. Leveraging Hive (a data warehouse infrastructure) and HDFS (Hadoop Distributed File System) on the Cloudera Data Platform, the project team has successfully automated the end-to-end ETL (Extract, Transform, Load) process for the big data received from Transport Infrastructure Ireland. The daily API-based data ingestion process, combined with a sophisticated data cleaning algorithm and a separate R script for data analysis, ensured the high quality of official statistics despite the challenges of handling big data. The project now features an integrated code pipeline that covers everything from data ingestion to publication. As a result, the CSO now publishes traffic count analysis more rapidly through a newly developed transport dashboard, providing faster indicators to support decision-making. This project highlights the transformative potential of R in enhancing data analysis workflows, from data ingestion to analysis and dissemination. This automation not only improved efficiency and reduced manual intervention but also minimised potential errors. In summary, this traffic counter project marks a significant advancement in utilising innovative data sources and data science within official statistics, setting a new standard for modern data analysis while ensuring the quality of official statistics.

References

No References available

Managing a Central Server for R Programming at Destatis: Practical Insights and Challenges

Authors

- Benedikt Ellinger (Federal Statistical Office of Germany (Destatis), Germany)
- Dr. Andreas Jahn (Federal Statistical Office of Germany (Destatis), Germany)
- Bernhard Fischer (Federal Statistical Office of Germany (Destatis), Germany)

Abstract

This presentation will examine our experiences in managing a central server system with Posit Workbench as a central IDE for Destatis users. We will discuss the practical aspects of managing R-packages in a shared infrastructure, highlight common user-issues and show our take on developing User-Management and monitoring systems using R. Furthermore, we will share our lessons learned from migrating our existing environment to new servers and the benefits of automated server rollout via Ansible. This presentation aims to provide valuable insights for those managing R in official statistics, highlighting the importance of effective system management and automation.

References

No References available

MEDOS: A Shiny Application for Seasonal Adjustment of Short-Term Statistics

Authors

- Muhammed Fatih TÜZEN (Turkish Statistical Institute (TURKSTAT, Türkiye)

Abstract

Seasonal adjustment is a critical process in economic and statistical analysis, enabling the identification of underlying trends by removing seasonal fluctuations from time series data. Recognizing this importance, the Turkish Statistical Institute (TURKSTAT) has developed MEDOS (Seasonal Adjustment Automation System), an innovative R Shiny application that revolutionizes the seasonal adjustment process for short-term statistics. MEDOS addresses the challenges of efficiency, consistency, and standardization in handling complex time series data through a centralized and automated approach. The need for such a system arose from the necessity to manage seasonal adjustments for multiple statistics and users within TURKSTAT. MEDOS performs adjustments from a single, centralized point, establishing a uniform standard across all produced short-term statistics. The application features a user-friendly interface where, upon secure login, subject matter areas can access their specific time series data from a centralized database. MEDOS applies pre-determined seasonal adjustment models, which are annually updated by domain experts using JDemetra+. This expert-driven model management ensures the application of best practices across all adjustments, significantly improving the reliability of results. Key features of MEDOS include sophisticated backend integration with JWSACruncher for rapid processing, automatic aggregation of results according to NACE codes and other classifications, and generation of visual analytics, control files, and revision reports. These features provide a comprehensive toolkit for quality assurance and data interpretation. The automation inherent in MEDOS brings numerous advantages: it reduces processing time, minimizes human error, and allows analysts to focus on interpretation rather than data processing. By streamlining the entire workflow, MEDOS not only enhances operational efficiency but also improves the overall quality and timeliness of seasonally adjusted statistics. One of the system's strengths is its flexibility. As demand grows, new statistics can be easily incorporated into MEDOS, allowing TURKSTAT to respond efficiently to evolving data needs. This adaptability ensures that the system remains relevant and comprehensive in its coverage of short-term statistics. MEDOS represents a significant advancement in the field of economic and statistical research. By centralizing and automating the seasonal adjustment process, TURKSTAT has set a new standard for efficiency, consistency, and accuracy in time series analysis. This innovative approach not only streamlines complex processes but also empowers decision-makers with timely, accurate, and standardized data, contributing to more informed economic policy-making and analysis in TURKSTAT.

References

No References available

Moving away from SAS: an opportunity to modernise the practices of statisticians

Authors

- Romain Lesur (INSEE, France)

Abstract

The rapid evolution of data science requires a reassessment of the tools and practices used by statisticians. This presentation explores the shift from proprietary softwares to open source alternatives, such as R and Python, which are fundamental to modern data science. This transition not only enhances the intrinsic value of statistical coding, but also introduces a more collaborative and interdisciplinary approach within the data community. Through the adoption of tools such as Git, statisticians are increasingly engaging in collaborative practices that align with those of software developers and data engineers. In addition, the adoption of DevOps principles provides an opportunity for statisticians to contribute more effectively to broader project teams, thereby enriching their role and impact in data-driven projects. This shift represents a critical opportunity to modernise statistical practices, thereby increasing project agility and reproducibility. In practice, the increasing use of data science platforms based on cloud technologies means that statisticians can enjoy the greatest possible autonomy. This transition not only modernises statistical practices, but also aligns them with current industry standards for software development and deployment, ensuring that statisticians can fully contribute to the demand of modern data analysis.

References

No References available

New data sources and Official Statistics: An exemplary research workflow in R using

different API wrapper packages

Authors

- Yannik Buhl (Federal Statistical Office of Germany (Destatis), Germany)

Abstract

Destatis continuously investigates new data sources with respect to whether they can improve existing Official Statistics, for instance with regard to their timeliness. This presentation provides an insight into a R research workflow at the German NSI, using the investigation of pedestrian counts as an economic indicator of retail turnover as an example. It focuses on the use of different R API wrappers as applied in the daily production process of creating experimental statistics, particularly the package {restatis} to easily access German Official Statistics as well as the {hystReet} package for automatically accessing pedestrian count data collected via laser scanners. The presentation emphasises the importance of automation using APIs in producing experimental statistics together with a well-defined workflow in R.

References

No References available

nonprobsvy: an R package for modern methods for non-probability surveys

Authors

- Łukasz Chrostowski (Faculty of Mathematics and Informatics, Adam Mickiewicz University, Poznań, Poland)
- Piotr Chlebicki (Faculty of Mathematics and Informatics, Adam Mickiewicz University, Poznań, Poland Centre for the Methodology of Population Studies, Statistical Office in Poznań, Poland)
- Maciej Beręsewicz (Department of Statistics, Poznań University of Economics and Business, Poland Centre for the Methodology of Population Studies, Statistical Office in Poznań, Poland)

Abstract

As the utilization of non-probability samples (e.g., big data, administrative data, online data) in official statistics continues to increase, there is a growing need for software that can seamlessly integrate these data into statistical production.

The goal of nonprobsvy package is to provide R users access to modern methods for non-probability samples when auxiliary information from the population or probability sample is available.

In particular, the package implements the following approaches: 1) inverse probability weighting estimators with possible calibration constraints (Chen, Li, and Wu 2020); 2) mass imputation estimators based in nearest neighbours (Yang, Kim, and Hwang 2021), predictive mean matching (Chlebicki et al. 2024) and regression imputation (Kim et al. 2021); 3) doubly robust estimators with bias minimization (Yang, Kim, and Song (2020)).

The package allows for: 1) variable section in high-dimensional space using SCAD (Yang, Kim, and Song 2020), Lasso and MCP penalty; 2) estimation of variance using analytical and bootstrap approach (see Wu (2023)); 3) integration with the survey package when probability sample is available Lumley (2023); 4) different links for selection (logit, probit and cloglog) and outcome (gaussian, binomial and poisson) variables.

The stable version of the package can be installed from the Comprehensive R Archive Network (CRAN), while the development version can be obtained from GitHub (https://github.com/ncn-foreigners/nonprobsvy).

Acknowledgements: The nonprobsvy package is produced as part of the project 'Towards census-like statistics for foreign-born populations – quality, data integration and estimation' funded by the National Science Centre grant OPUS 20 (2020/39/B/HS4/00941).

References

- Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532), 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

- Piotr, C., Chrostowski, Ł., & Beręsewicz, M. (2024). Data integration of non-probability and probability samples with predictive mean matching. arXiv preprint arXiv:2403.13750. Kim, J. K., Park, S., Chen, Y., & Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. Journal of the Royal Statistical Society Series A: Statistics in Society, 184(3), 941–963. https://doi.org/10.1111/rssa.12696
- Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software, 9(1), 1–19.
- Lumley, T. (2023). Survey: Analysis of Complex Survey Samples.
- Wu, C. (2023). Statistical inference with non-probability survey samples. Survey Methodology, 48(2), 283–311. https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-eng.htm
- Yang, S., Kim, J. K., & Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. Survey Methodology, 47(1), 29–58. https://www150.statcan.gc.ca/n1/pub/12-001-x/2021001/article/00004-eng.htm
- Yang, S., Kim, J. K., & Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high-dimensional data. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(2), 445–465. https://doi.org/10.1111/rssb.12354

nonprobsvy: an R package for modern methods for non-probability surveys

Authors

- Łukasz Chrostowski (Faculty of Mathematics and Informatics, Adam Mickiewicz University, Poznań, Poland)
- Maciej Beręsewicz (Department of Statistics, Poznań University of Economics and Business, Poland Centre for the Methodology of Population Studies, Statistical Office in Poznań, Poland)

Abstract

The nonpobsvy package implements point and variance estimation methods for non-probability samples (e.g., big data, administrative records, or web panels) using additional information from population registers or probability samples. The majority of the methods implemented in the nonprobsvy package can be found in the review article Wu (2022).

The stable version of the package can be installed from the Comprehensive R Archive Network (CRAN), while the development version can be obtained from GitHub (https://github.com/ncn-foreigners/nonprobsvy).

Workshop details

The workshop is designed for users who have a working knowledge of probability and non-probability samples, as well as statistical inference methods.

Time: 1 slot (2 hours): ¼ theory, ¾ practice

Outline:

- Introduction to inference based on non-probability samples
- Prediction approach (mass imputation)
- Pseudo-randomization approach (inverse probability weighting)
- Doubly robust approach (doubly robust estimators)
- Other issues (e.g. variable selection)

Data: - The workshop will focus on the analysis of real-world scenarios related to job vacancies and opt-in surveys.

Acknowledgements

The nonprobsvy package is produced as part of the project 'Towards census-like statistics for foreign-born populations – quality, data integration and estimation' funded by the National Science Centre grant OPUS 20 (2020/39/B/HS4/00941).

References

- Wu, C. (2022). Statistical inference with non-probability survey samples. Surv. Methodol, 48, 283-311.

R packages around JDemetra+: a versatile toolbox for time series analysis

Authors

- Anna Smyk (INSEE, France)
- Tanguy Barthelemy (INSEE, France)

Abstract

JDemetra+ (https://jdemetra-new-documentation.netlify.app/) is an open source time series analysis software that provides algorithms for seasonal adjustment, trend-cycle extraction, outlier detection, nowcasting, and revision analysis. All these algorithms, implemented in Java, are accessible through a graphical user interface and also through an ecosystem of R packages, the rjdverse (https://github.com/rjdverse), which we would like to present at the UROS 2024 conference. JDemetra+ is a key player in official statistics in Europe, it has been officially recommended by Eurostat to the members of the European Statistical System since 2015 and its use is growing in statistical agencies around the world. Its latest version (3.x), first released in May 2023, fills several critical gaps in a time series analyst's toolbox by offering advanced seasonal adjustment capabilities, including high-frequency data and new R production tools, as well as an improved version of all its algorithms. An extended state-space framework offers large modelling capabilities, among which a seasonal adjustment procedure with explicit decomposition and time-varying trading-day correction. JDemetra+ also provides fast and efficient tools for Arima model estimation and UCArima, decomposition, a wide range of (seasonality) tests and routines for calendar regressors generation.

References

No References available

Regional food purchasing behavior characteristics contained

in the Household Income and Expenditure Survey data

Authors

- Atsushi Kimura (National Statistics Center, Japan)

Abstract

The National Statistics Center in Japan is an organization that is in charge of the process of creating clean data based on questionnaires, mainly statistics that are important for national policy decisions implemented by the Statistics Bureau of the Ministry of Internal Affairs and Communications, and of compiling and publishing the clean data in statistical tables. This time, we will report on a topic related to the core statistic called the Household Income and Expenditure Survey. The Statistics Bureau of the Ministry of Internal Affairs and Communications asks about 9,000 households to submit their household accounts every month, and creates and publishes statistics on household income and expenditure. This is called the Household Income and Expenditure Survey. The published statistical tables also include household income and expenditure tables for each prefectural capital city. From this regional statistical table, we can extract only data on food item expenditures and perform cluster analysis using the similarity of food item purchasing trends between regions, which can produce very interesting results. Japan is an island country that stretches about 3,000 km from north to south and about 3,000 km from east to west. It is made up of a total of 47 prefectures, from Hokkaido Prefecture in the north to Okinawa Prefecture in the south. Today, with the spread of social networking sites such as YouTube and Instagram, it is now possible to access the same quality of information in real time from anywhere in Japan using a smartphone. In addition, we are now in a location-free era where you can easily get the products you want anywhere in Japan by using various EC platforms such as Amazon. Nowadays, trends in fashion, music, movies, and more spread across Japan in the blink of an eye. However, even in this era, it is well known from experience that there are regional differences in everyday food preferences in Japan. It is also well known that each region has a diverse and rich food culture. In this presentation, we will report on how we can visualize the regional characteristics of food preferences in Japan by extracting regional food purchase amount data from the published data of the Household Income and Expenditure Survey conducted by the Statistics Bureau of the Ministry of Internal Affairs and Communications and performing Ward's method hierarchical cluster analysis using the hclust function. We will also introduce that this regional characteristic is a common and stable structure that exists in the data of different survey years of the Household Income and Expenditure Survey. The data of the Household Income and Expenditure Survey does not contain any geographical information such as regional adjacent information or distance information between regions, and is a dataset consisting purely of purchase amounts. However, as we will present, it has become clear that regional characteristics are inherent in the purchasing behavior of food items in Japan. We will also touch on the possibility of applying this characteristic to the efficiency of analytical review work (GSBPM: Analyse Phase (6.2 Validate outputs sub-process)). In addition, we will introduce two original methods that are useful for interpreting the results of analysis of data from different survey years. The first is a cluster correspondence method for matching corresponding

clusters when comparing the results of analyzing household Income and Expenditure survey data from different survey years. The second is a method for visually making the characteristics between clusters easier to understand. In cluster analysis using multivariate data, it is useful for understanding the contribution rate of each variable to the differences between clusters that agglomerate.

References

No References available

Rmarkdown report in R Shiny

Authors

- Oļesja Nikoluškina (The Central Statistical Bureau of Latvia, Latvia)

Abstract

In the past years, we have experienced with many regular, repeatable work tasks and production of statistics before the deadline, therefore we made decision to build a package within the integrated shiny framework. This package makes available to employee of The Central Statistical Bureau of Latvia perform data checks, mathematical tasks related to the sampling theory as sampling weights and data quality at any time. It also keeps the previous best practices of non-automated sampling weights adjustment, in a result notifications pops up in case of an errors, warnings and notes. The package provides users with graphical interface and Rmarkdwon interactive reports at once, which automatically are saved for later analysis.
Interaction with users have resulted in supplementing the package with new features, surveys, improvements and functionalities, for example, population frame updating, outliers detection and additional notifications. We provide users with a step by step guide, moreover support from our side is available. The presentation will keep focus on our challenges, benefits and experience of technical side faced when building and publishing the shiny application.

References

No References available

RMLUtils: an Official Statistics oriented common interface for Machine Learning

Authors

- Luis Sanguiao Sande (S.G. for Methodology and Sampling Design, Statistics Spain, Spain)
- Carlos Sáez Calvo (S.G. for Methodology and Sampling Design, Statistics Spain, Spain)
- Sandra Barragán Andrés (S.G. for Methodology and Sampling Design, Statistics Spain, Spain)
- Ester Puerto Sanz (S.G. for Methodology and Sampling Design, Statistics Spain, Spain)
- María Novás Filgueira (-)
- Javier Villaescusa Almagro (-)
- Sergio Pardina Quirós (-)
- Álvaro García Tenorio (-)

Abstract

The use of machine learning methods is becoming increasily common in Official Statistics, and they will be soon an important part of the production pipelines. There are many R packages to provide machine learning methods, but if we would want to replace one method with another one, we might end up changing the script. RMLUtils is aimed to provide a common interface for the machine learning tasks that might take part on the standard processes of a Statistical Office. This way, the algorithm can be changed while the scripts remain unchanged. The package is designed for Official Statistics, so it provides some features that are not so usual in machine learning, specially for aggregate estimation. More OS oriented features will be added in the near future. Moreover, its modular and object oriented design makes it easy to include new methods. For example, there are some compound methods that has been already added: multimodel predictions with benchmarking, modeling with preprocessed features and classification-regression cascade of two models.

References

No References available

Romania's experience in using R in the official statistics

Authors

- Bogdan OANCEA (National Institute of Statistics, Romania University of Bucharest, Romania)
- Marian NECULA (National Institute of Statistics, Romania University of Bucharest, Romania)
- Ana-Maria CIUHU (National Institute of Statistics, Romania Institute of National Economy, Romanian Academy, Romania)
- Ciprian ALEXANDRU (Ecological University of Bucharest, Romania)
- Raluca-Mariana DRĂGOESCU (National Institute of Statistics, Romania)

Abstract

This presentation addresses two main aspects. The first focuses on the organizational and technical challenges of introducing R into the National Institute of Statistics (NIS) of Romania, while the second concentrates on training users within the office to proficiently use R. At NIS, R is predominantly used in two main domains: innovative tools in statistics and social statistics. Current developments in the use of R for innovative tools include, but are not limited to: web scrapping, package development for extracting data from the Romanian INS Tempo database, use of remote sensing data, text classification and MNO data. In social statistics, R is actively used for sampling, calibration, and calculation of quality indicators in household surveys, data validation, and generation of data tables for dissemination. Additionally, R is employed for analyzing paradata, accessing databases from statistical and administrative sources, and conducting Small Area Estimations. The implementation of R programming language in Romanian INS has started in 2014 and covered various levels and applications of R in official statistics such as: basics of R, advanced R for creating R packages, automated reports, and parallelization, model-based estimates and small area estimation. Training sessions ranged from one-week courses to six-month on-the-job training, tailored to practical applications. We conclude with a series of proposals on future research opportunities and other potential analytical procedures using R for innovative tools and social statistics.

References

No References available

Semantic address matching using Keras for R

Authors

- Paula Cruz (Statistics Portugal NOVA IMS)
- Leonardo Vanneschi (NOVA IMS)
- Marco Painho (NOVA IMS)
- Filipa Ribeiro (Statistics Portugal)

Abstract

Statistics Portugal is presently facing two major challenges: the production of annual census statistics based on administrative data (involving the mapping of individuals to housing units) and the move into a more regular and efficient production of essential indicators, with a reduced burden on respondents (including enterprises). Current household statistical operations rely on the National Dwellings Registry, which comprises approximately six million address records. Statistics Portugal is also responsible for managing non-residential statistical units corresponding to farms and establishments. These files are updated based on data from internal and external sources, with varying degrees of quality. However, current parsing and matching procedures are based on a time-consuming process that mostly relies on deterministic or rule-based techniques and is highly dependent on address quality. To overcome these limitations, the adoption of deep learning architectures for semantic address matching is being evaluated by using synthetic labelled data in combination with current accumulated real labelled data. Initial results using Keras and Tensorflow for R packages appear to be promising, but the developed algorithms require further improvement in terms of accuracy and performance compared to those implemented in Python.

References

No References available

Statistical Business Registers updates monitoring and dashboards

Authors

- Pedro Carrasqueira (Statistics Portugal)
- António Portugal (Statistics Portugal)
- Isabel Farinha (Statistics Portugal)

Abstract

National Statistical OÜices receive regularly a huge amount of legal and economic information about enterprises and other institutions. The collected information is used to update business registers, which serve several statistical purposes. It is of utmost importance to analyse and monitor regularly the operations performed on registers. This task may be achieved by stablishing a sequence of steps to be executed in a regular basis. R is a powerful tool to accomplish this work in an eÜective way, since it enables to automate all process from data collection to dashboard release. Thus, an algorithm was programmed in R to obtain, in an automatic fashion, the data required, from the data warehouse, and further produce a monthly dashboard just by providing the month and year of analysis. The data collected enabled us to produce multiple indicators such as number of new business' raised for each NACE Level 1, variables' update counting or homologous comparison on the referred operations. R flexdashboard tool was used to design the dashboard to make the information easily understandable. This automated workflow makes it easier to release a monthly report strongly reducing human time consumption.

References

No References available

Statistical disclosure control for Census data with sdcMicro package: the computational challenges

Authors

- Vasiliki Spiliopoulou (Hellenic Statistical Authority)
- Dionysios Fragkopoulos (Hellenic Statistical Authority)

Abstract

Confidentiality must be respected for any dataset containing sensitive information about individuals. In this context, the implementation of methods ensuring adherence to ethical and legal commitments to protect statistical data confidentiality is necessary. The cell key perturbation method has been widely tested and used by National Statistical Offices (NSOs) for this purpose, particularly through the R package sdcTools. The Hellenic Statistical Authority has used this method on Census cubes submitted to Eurostat, as well as for the dissemination of national census tables. This paper presents the computational challenges encountered during this process, providing details on the hierarchies and the computational times required for complex cubes. An exponential impact on the time and system capacity needs is introduced for the production of protected cubes with more than six variables and the effect of the variable hierarchies. Additionally, the technical and statistical solutions employed to efficiently automate and speed up the production of protected datasets, overcoming computational resource constraints, are discussed.

References

- Rodrigues, I., Paulino, P., Campos, P., & Fragoso, T. (2019). A framework for assessing perturbative methods for protection of Census 2021 data at Statistics Portugal.

Streamlining Web Paradata Retrieval and Analysis at Statistics Sweden: A Semi-Automated Approach for Enhanced Efficiency and Consistency

Authors

- Gustaf Andersson (Statistics Sweden)
- Emma Stavås (Statistics Sweden)

Abstract

With the growing popularity of web surveys, we present a novel standard report designed to streamline the retrieval and analysis of web paradata collected at Statistics Sweden. Previously, extensive manual cleaning and inconsistent paradata variable definitions led to significant delays and inefficiencies. In addition, the lack of a structured approach to utilize paradata rendered potential applications unfeasible. Our solution employs GUI programming with RShiny and RMarkdown to create a user-friendly point-and-click interface and parameterized reports. This automation reduces the time needed for analysis and ensures consistency across different reports. The standard report enables users across various departments, regardless of programming skills, to effortlessly access ready-to-analyze web paradata, fostering data-driven decision-making and enhancing operational efficiency. We also present some feedback which indicates substantial improvements in workflow efficiency and data accessibility, demonstrating the report's value in increasing the use of web paradata at Statistics Sweden.

References

No References available

Supervised statistical (machine) learning for domain estimation

with business survey data

Authors

- Vasilis Chasiotis (Department of Statistics, Athens University of Economics and Business, Athens, Greece)
- Nikos Tzavidis (Department of Social Statistics and Demography, and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom)
- Chiara Bocci (Department of Statistics, Computer Science, Applications “G. Parenti”, University of Florence, Florence, Italy)
- Paul Smith (Department of Social Statistics and Demography, and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom)

Abstract

We discuss current research on supervised statistical (machine) learning methods (ran- dom forests) and extensions to mixed effects random forests (Krennmair & Schmid, 2022), as a flexible method for domain estimation with business survey data. Random forests excel in terms of predictive performance. Automated model-selection and detecting covariate in- teractions make their use appealing for prediction problems. Mixed effects random forests, however, appear to be going against the algorithmic modelling culture (Breiman, 2001), that treats the prediction mechanism as unknown, and are more in line with the data modelling culture (Efron, 2020). Model-based estimation with business survey data requires careful handling and may include outlier robust estimation, complex modelling of the model variance and use of data-driven transformations. We explore the use of random forest-type algorithms for estimation of finite population parameters. We focus on critically evaluating (a) the role of random effects in machine learning algorithms, (b) the role of data transformations, and (c) whether machine learning algorithms offer protection under misspecification of linear-type models. Small area predictors are derived by using a smearing-type estimator that has been explored in small area and survey estimation before in the context of outlier-robust estimation (Chambers et al., 2014). A non-parametric bootstrap MSE estimator is evaluated. We compare machine learning-based predictors to empirical best predictors and outlier robust predictors under a linear mixed model (Smith et al., 2021) using real business survey data form Italy. This work aims to inform the discussion on the use of machine learning methods in the production of official statistics

References

- Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5–32.
- Chambers, R., Chandra, H., Salvati, N., & Tzavidis, N. (2014). Outlier robust small
- area estimation. Journal of the Royal Statistical Society: Series B , 76 (1), 47-69.
- 1
- Efron, B. (2020). Prediction, estimation, and attribution. Journal of the American

- Statistical Association, 115 (530), 636-655.
- Krennmair, P., & Schmid, T. (2022). Flexible domain prediction using mixed effects
- random forests. Journal of Royal Statistical Society: Series C (Applied Statistics),
- 71 (5), 1865–1894.
- Smith, P., Bocci, C., Tzavidis, N., Krieg, S., & Smeets, M. J. E. (2021). Robust
- estimation for small domains in business surveys. Journal of Royal Statistical
- Society: Series C (Applied Statistics), 70 (2), 312—-334.

Systematic Data Validation with the validate Package

Authors

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

Abstract

Testing whether data is fit-for-purpose is one of the cornerstones of producing official statistics. Ideally this is done such that it is clear for all stakeholders what is being tested, yielding testing results that are unambiguous. The R package validate allows domain experts to define, document, and maintain sets of data validation rules. Using validate, the rules can be applied to a data set, yielding clear and unambiguous results that can be processed, analyzed and studied as data. In this short interactive tutorial we will touch upon the main theoretical principles underlying data validation and the validate package. Subsequently, participants will work with the package and get to know its main functionalities and workflows.

Participants are expected to bring a laptop with a recent version of R and the validate package installed.

References

- • MPJ van der Loo, E de Jonge (2021). Data Validation Infrastructure for R. Journal of Statistical Software 1–22 97 paper.
- • MPJ van der Loo, E de Jonge (2020). Data Validation. In Wiley StatsRef: Statistics Reference Online, pages 1-7. American Cancer Society. pdf
- • M Di Zio, N Fursova, T Gelsema, S Giessing, U Guarnera, J Ptrauskiene, L Quensel-von Kalben, M Scanu, K ten Bosch, M van der Loo, K Walsdorfe (2015). Methodology for data validation. ESSNet on validation deliverable No. 2 . pdf

TEAM: an R package for time series model identification

Authors

- C. Sáez Calvo (S.G. for Methodology and Sampling Design, Statistics Spain, Spain)
- L. Sanguiao Sande (S.G. for Methodology and Sampling Design, Statistics Spain, Spain)
- Félix Aparicio Pérez (S.G. for Methodology and Sampling Design, Statistics Spain, Spain)
- María Teresa Vázquez Gutiérrez (S.G. for Information Technologies and Communications, Statistics Spain, Spain)
- José Fernando Arranz Arauzo (S.G. for Information Technologies and Communications, Statistics Spain, Spain)

Abstract

Seasonal adjustment of time series plays a pivotal role in modern official statistics, ensuring accurate and reliable data analysis. To obtain a seasonal adjusted time series, it is necessary to identify an adequate seasonal RegARIMA model for the series. The quality of the identified RegARIMA model is crucial to ensure the quality of the obtained seasonal adjusted series. However, due to resource constraints and time limitations, the models identified in an automatic way using the current software may not be optimal. This leads to a worse performance of seasonal adjustment, since these models must be maintained for a year. We present a new R package, Time-Series Exhaustive Automatic Modeling (TEAM), building on the JDemetra+ R ecosystem, which aims to automate and enhance the yearly model identification phase. The goal is to provide in an automated way a list of optimal models, where the optimality criteria can be specified by the users to meet their specific needs. The methodology employed in TEAM is characterized by an exhaustive search and ranking of models. Initially, an exhaustive search of specifications is conducted for each time series, testing all possibilities for parameters such as data transformations (logarithms or levels), the order of the ARIMA model, inclusion of outliers, and calendar regressors. Subsequently, each specification is processed using the JDemetra+ software in a parallelized way, yielding diagnostic information to construct several indicators assessing the quality of the specifications in several areas of quality. To rank the models, a final score is computed by appropriately combining the obtained indicators. Importantly, users retain the flexibility to adjust the weights assigned to each area according to their specific requirements. For instance, users could prioritize models with minimal revisions based on their preferences. Finally, TEAM presents the user with a selection of the best models based on the final score, enabling them to choose the most suitable model according to their needs.

References

No References available

Testing for the bias in the estimation of business structure indexes from different data sources

Authors

- Michaela Balkoudi (Aristotle University of Thessaloniki, Department of Mathematics, Greece)

Abstract

A key question that concerns Statistical Authorities using data collected from surveys and administrative sources for the compilation of specific statistical indexes is whether the data source (survey, administrative files) affects the indexes. In this paper, aggregated data from theGreek Structural Business Statistics of the reference years 2014-2018 for 8 statistical indexes of 140 business branches, originated from two sources, survey and administrative files, have been analyzed to assess whether the computed statistical indexes from the two data sources differ. The same analysis is repeated for each of the eight statistical indexes. The structure of the data is that of repeated measures (5 years) with random effects, where the within-subjects factors of the experimental design are the data source (of prime interest) and the reference year, and the random effects regards the business branches, as the level of a statistical index differs distinctly across the 140 business branches. The statistical testing was performed using two parametric statistical tests, the two-way repeated measures ANOVA and the linear mixed models, as well as the respective bootstrap tests (using wild bootstrap for the linear mixed models). The consistency of the parametric and bootstrap tests was first assessed on simulated data, using the data setting of the real data but determining different scenarios for the dependence of the statistical index on each factor. The simulation study concluded that even for strong deviations of the data from normality the parametric and bootstrap tests agreed to the correct test decisions. The results of the real data analysis confirmed the overall agreement of the parametric and bootstrap tests, and for two statistical indexes statistically significant effect of the data source was found. The results of the study suggest that the data source may have an impact on the derived statistical indexes.

References

No References available

The R Package "Survey Data Quality"

Authors

- Ioannis Andreadis (University Thessaloniki Vasileos Irakleiou, Greece)

Abstract

In this paper, the R Package "Survey Data Quality" is introduced, a library of functions that can be used for the assessment of the quality of survey data. Survey methodology scholars have used various methods to measure the attentiveness of the respondents and the quality of the collected data: item-nonresponse, mid-point responses in Likert-type scale items, straight-lining, the time spent on questionnaire items (speeding), etc. Using the aforementioned response quality indicators we can create an innovative multidimensional estimation of response quality for each completed questionnaire. Using this estimation, we can identify questionnaires that have been submitted by less attentive web survey respondents.

References

No References available

The Synergy of R and Generative AI in Statistics

Authors

- Vytas Vaiciulis (Central Statistics Office, Ireland)

Abstract

The integration of generative AI (GenAI) with R software offers improvements and advancements for statisticians and data scientists in National Statistical Offices (NSIs). This paper explores the application of GenAI to address challenges in statistical code development and maintenance. We emphasize the implementation of prompt engineering and guardrails to ensure the generation of high-quality, reliable outputs and detail our continuous efforts in testing and validating the results produced by GenAI. Importantly, GenAI functions as an assistant with a focus on facilitating code translation between programming languages. Users are required to verify and validate all outcomes. This paper outlines the ongoing iterative process of refining and enhancing GenAI models, emphasizing their evolving capabilities and the continuous improvements being made. By leveraging the strengths of both R and generative AI, we aim to enhance statistical production systems, making them more adaptable and efficient. Our study also opens up discussions on the potential future innovations that this synergy can bring to statistical analysis and data science. Overall, this paper sheds light on the transformative potential of integrating generative AI with R, laying the way for future advancements in the field and offering new methodologies to tackle the dynamic challenges faced by NSIs.

References

No References available

Time Series Analysis of Well-being and Poverty Using R

Authors

- Tudor IRIMIA (The Bucharest University of Economic Studies, Romania)
- Iliana CARAGEA (The Bucharest University of Economic Studies, Romania)
- Emilia ȚIȚAN (The Bucharest University of Economic Studies, Romania)

Abstract

Accurate measurement of poverty and well-being is essential for informed policy formulation and socio-economic planning. Reliable data on these indicators enable governments and organisations to design effective policies, allocate resources efficiently and monitor the impact of interventions over time. By understanding the trends and dynamics of poverty and well-being, policymakers can address root causes, anticipate future challenges and implement sustainable solutions. This study uses advanced time series analysis techniques in R to model and forecast key indicators of poverty and well-being. Using R's comprehensive time series packages, including forecast, tseries, TSstudio, and prophet, we provide a detailed overview of multiple indicators through the lens of time series capabilities.

References

- Time Series Analysis of Poverty and Income Inequality: Crespo, N., Proença, I. and Fontoura,
- M.P., 2010. The spatial dimension in FDI spillovers: Evidence at the regional level from
- Portugal. Working Papers Department of Economics, 2010/17. ISEG - Lisbon School of
- Economics and Management, Department of Economics, Universidade de Lisboa. Available at:
- RePEc.
- Bibliometric Analysis of Poverty Reduction Studies: Frontiers, 2021. Poverty reduction of
- sustainable development goals in the 21st century: A bibliometric analysis. Frontiers in
- Sustainable Development. Available at: Frontiers.
- Sustainability, Financial Inclusion, and Poverty Reduction: Li, Z. and Qamruzzaman, M.,
- 2023. Nexus between Environmental Degradation, Clean Energy, Financial Inclusion, and
- Poverty: Evidence with DSUR, CUP-FM, and CUP-BC Estimation. Sustainability, 15(19),
- 14161. Available at: MDPI.

Transforming Excel reports into SMDX compatible datasets with R

Authors

- Athanassios Stavrakoudis (Applied Informatics and Computaional Economics Lab Department of Economics Univesity of Ioannina, Ioannina, Greece)

Abstract

Many official data providers distribute datasets using Excel files. This is the case for example for the Hellenic Statistical Authority (ELSTAT) and the Bank of Greece (BoG). Datasets are mostly in untidy format and it is very difficult to use them in analysis without prior heavy data-wrangling work. In this article I proposed the use of R and its tidyverse rerated toolkits in order to transform untidy data reports (inflation, GDP, unemployment, prices, etc) into SDMX (Statistical Data and Metadata Exchange) compatible format. After transformation everything is included into an R package. The main problems with untidy excel files that this package solves are: 1. Different date formats. The same date (for example 1st quarter of 2024) can be found as 2024-I, 2024-1, 1/2024, 2024/1, 2024-Q1, Q1-2024, or values in 2 or 3 cells. This is rather confusing and of course such dates formats can not easily handled with econometric software packages, like eviews, stata, matlab, etc. The proposed package transforms every single date into ISO format YYYY-MM-DD and solves this problem. This also has a benefit that facilitates the correct join of different datasets, for example GDP growth and unemployment rate. With current excel files this is possible, just because files use different structure and date notation. 2. Geo reference. Neither ELSTAT nor BoG provide a geo code for administrative regions of Greece, like NUTS1, NUTS2, etc. The proposed package uses a standard way to represent the regions that makes comparison, join, merge, etc, of different datasets rather easy. Moreover, for NUTS1, NUTS2 regions the EUROSTAT's geo dictionary is also used for compatibility. 3. Separates dataset from metadata and report. Every time a single and clean orthogonal dataset is provided, ready to be used by any software. Dataset can be exported as SMDX flavors (csv/json) in .xlsx, .csv, .txt, .dta, .json, etc , file formats. Metadata and possible reporting/visualization are separated. The package depends heavily on tidyr, dplyr, readxl, lubridate, stringr, validate, and some other packages. Currently approximately 100 datasets are downloaded and packaged on a monthly basis. Package release is scheduled for mid of September 2024.

References

No References available

Unemployment, Net Income and Emigration. An Econometric Study

Authors

- Nicolae-Marius JULA (Faculty of Business and Administration, University of Bucharest, Romania)
- Dorin JULA (Institute of Economic Forecasting, Romanian Academy, Romania Faculty of Financial Management, Ecological University of Bucharest, Romania)

Abstract

In the paper, we built an econometric model to analyse the impact of unemployment rates and the dynamics of mean equivalised annual net income on emigration from Romania to Spain. Emigration from Romania to Spain has become a significant phenomenon in recent decades, generating social, economic, and cultural changes both in the country of origin and in the host country. The main explanations for emigration could be economic factors (poverty, lack of employment opportunities, and low salary levels in Romania), social factors (higher quality of life in Spain - social, medical, educational services), institutional, legal, demographic factors, cultural and political. As a methodology, we used an ARDL-type model with automatic specification selection (lag dimension). As a technique for solving the model, we used the R environment. The model suggests that, in the long run (cointegration relationship), the unemployment rate in Romania is positively associated (as a sign) with emigration from Romania to Spain while the unemployment rates in Spain for foreign country citizenship moderate emigration (the association is negative), as does the increase in mean equivalised annual net income in Romania. Conversely, the increase in mean equivalised annual net income in Spain encourages emigration from Romania.

References

No References available

Use of R to automate Eurostat statistical publications:

Insights from the third edition of the EMOS Coding Lab

Authors

- Matyas Meszaros (Eurostat, Luxembourg)
- Andrea Gallelli (Eurostat, Luxembourg)
- Tina Steenvoorden (GOPA Worldwide Consultants, Luxembourg)
- Maja Islam (Eurostat, Luxembourg)

Abstract

The presentation will delve into the outcomes and experiences of the third edition of the EMOS Coding Lab, held between April and June 2024, focusing on the intersection of statistical education, computing skills and use of R to enhance organisational dissemination processes. The Coding Lab, an initiative by Eurostat, aims to foster a bottom-up approach to statistical dissemination, aligning with principles of transparency and accessibility outlined in the European Statistics Code of Practice. The presentation will highlight the significance of reproducibility in statistical workflows and the role of literate programming in achieving this goal. In the third edition of the Coding lab, through the replication of Eurostat's Statistics Explained articles using R programming, participants engaged in data retrieval, analysis, visualization, and text integration. This approach helped automatise the data dissemination processes at Eurostat while also facilitating challenge-based learning and promoting collaboration among students enrolled in EMOS-labelled master's programmes across Europe. The key elements of the presentation will include the organisational set-up, description of the tasks developed for students, results and learning outcomes achieved by the participants. Through their presentation, the Coding lab organisers will offer valuable insights into the role of collaborative coding labs in developing statistical education, promoting reproducibility, and fostering a culture of open science in official statistics. Through practical examples and firsthand experiences, attendees will gain actionable insights into implementing similar initiatives in their respective domains.

References

No References available

Use of R tools in the statistical disclosure control of census data in Poland

Authors

- Kamil Wilak (Statistical Office in Poznań, Centre for Small Area Estimation, Poland Poznań University of Economics and Business, Chair of Statistics, Poland)
- Tomasz Józefowski (Statistical Office in Poznań, Centre for Small Area Estimation, Poland Poznań University of Economics and Business, Chair of Statistics, Poland)
- Andrzej Młodak (Statistical Office in Poznań, Centre for Small Area Estimation, Poland The University of Kalisz, Inter-faculty Department of Mathematics and Statistics, Poland)
- Tomasz Klimanek (Statistical Office in Poznań, Deputy Director, Poland Poznań University of Economics and Business, Chair of Statistics, Poland)

Abstract

The preparation of data from the censuses meets the needs of potential users of statistical data concerning detailed and versatile information but also must ensure efficient protection of statistical confidentiality. Therefore, it is necessary to use the Statistical Disclosure Control (SDC) methods aimed at effectively protecting statistical confidentiality by minimizing the risk of identifying an entity while maximizing the usefulness of the shared data. Due to a number of variables of various type and very large number of records application of SDC in the case of census data is special challenge. In our presentation we will show two cases of application of SDC using the relevant R tools. First of them is the protection and methodological solutions used to carry out the data disclosure control process for hypercubes, as well as the effects obtained in this regard. We will discuss the 119 hypercubes subjected to the SDC process. Most of them concern individuals, but some also concern households, families and dwellings. The key element of the SDC process is the cell key method – a post-table interference tool for protecting statistical confidentiality, recommended by Eurostat. The usefulness of the data prepared in this way was verified using similarity measures based on the comparison of relevant cells before and after the disturbance. When discussing the process and indicate how to use the cellKey R package in this case. Especially, the our original complex algorithm enabling simultaneous perturbation of all hypercubes of the same type will be presented. The second issue will be a trial to create an efficient algorithm of the Statistical Disclosure Control process for microdata in statistical portal developed for dissemination of versatile and detailed statistical information in spatial dimension. The proposed algorithm is based on perturbative methods, such as microaggregation using the Gower distance for categorical variables and addition of correlated noise for the continuous ones, but allows also for usage of several alternative options in this context. It ensures also an assessment of the information loss using the measures of distribution disturbance and measures of impact on the power of connections between variables (the latter – for continuous variables). The specific algorithm – using possibilities of the sdcMicro R package – was tested using the microdata on farms and farm animals collected during the Agricultural Census conducted in Poland in 2020. We present obtained results in both cases and formulate the main problems and challenges connected with application of such tools in practice.

References

No References available

Using R and SAS Viya for Statistical Production – a Multiple-Tool Approach

Authors

- Dr. Andreas Jahn (Federal Statistical Office of Germany (Destatis), Germany)
- Benedikt Ellinger (Federal Statistical Office of Germany (Destatis), Germany)
- Bernhard Fischer (Federal Statistical Office of Germany (Destatis), Germany)

Abstract

Official statistics agencies are currently facing the issue of modernizing its production environments. On the one hand, cloud technology becomes more and more accepted and use-cases provide promising results. This trend favors the use of commercial software, such as SAS. On the other hands, budget limitations and the trend towards using open-source products leads to a broader use of R. Due to the federal structure of the German statistical offices, taking an either – or decision was difficult to achieve and our agency uses a multi-product approach, fostering the use of SAS Viya and R at the same time. The installed base of more than 10,000 code snippets made an immediate transition to R impossible to achieve. On the other hand, new staff entering the agency is more fluent with R code and calls for an appropriate production environment. The use of SAS Viya enables the parallel use of R and Python within the SAS environment. Our approach consists of analyzing existing SAS software code, creating a timeline for SAS Viya migration, building a powerful R-server and addressing the issue of creating a data-management architecture. In the area of people management, we created an extensive education program. To make the parallel use of SAS and R easier, we tested the use of Large Language Models (ChatGPT4) for automated translations of code between the two technologies. Statistical disclosure control was implemented in R and SAS using similar web-services. Results of this paper can be used for benchmarking future multiple-tool approaches within statistical agencies and to learn from prior experiences.

References

No References available

Using R for efficient content production

The case of a monthly Statistics Explained article

Authors

- Bogdan Alexandru Micu (European Commission – DG TAXUD, Belgium)

Abstract

I present a content production helper, produced with R software and R Markdown, used for rapidly generating the elements of the monthly update of a Statistics Explained article. The presentation will not focus on the capabilities of R software in the areas of data processing, analysis, or visualization - which are well established - but on its use for more efficient generation of number-embedding narrative (text that presents numbers, and includes descriptions of those numbers and their relationships) and for document generation and file manipulation.

References

No References available

Utilizing R for Simulating Studies in International Large-Scale Assessments

Authors

- Umut Atasever (IEA, Hamburg, Germany)
- Francis L. Huang (University of Missouri, USA)
- Leslie Rutkowski (Indiana University, USA)

Abstract

International large-scale assessments (ILSAs) like TIMSS and ICILS, conducted by the IEA, provide invaluable insights into student performance and educational system trends over time. These assessments employ complex methodologies, including multi-stage cluster sampling designs and the generation of sampling weights to accurately represent populations. Simulation studies play a critical role in evaluating these methodological choices; however, they frequently fall short in transparency, with code often unavailable, and many rely on infinite population models that can be challenging to understand and replicate. To address this issue, we propose an alternative approach: generating a finite population for simulations with known population values. This method is inspired by techniques from our working paper, which focuses on the use of weights in multilevel models within ILSAs. This topic has been widely debated, particularly regarding the optimal way to apply weights at different model levels. Using repeated sampling from this finite population, we generate weights based on a stratified, two-stage cluster design used in ILSAs, with schools as primary sampling units and classes or students as secondary sampling units. For each sample drawn, we then assess bias and coverage rates for fixed and random effects across models with varying weight applications: weights at both levels, only at level 2, only at level 1, and without weights. Our findings indicate that applying weights only at level 2 results in the most accurate estimates, while models without weights or with rescaled level-1 weights alone show the highest bias. While we illustrate these results using a simulation approach tailored for multilevel models, the framework can be adopted for other research projects using R, enabling a comparison of various methodologies in the context of ILSAs. In this short paper, we demonstrate how to apply this simulation technique in R, providing a practical guide for researchers to adopt this method in their studies. Our goal is to enhance the transparency and reproducibility of simulation studies in educational assessments, ultimately contributing to more robust and reliable research outcomes. We also demonstrate how to use R's parallel processing capabilities to enable faster computations of repeated samples.

References

No References available

Why and for what purpose R in Official Statistics?

Authors

- Sandra Barragán (INE, Spain)

Abstract

In the discipline of Official Statistics there is an absolute necessity of evolving towards programming languages oriented to modern data science to face new challenges such as fast prototyping innovative ideas, standardizing production processes and automatizing manual tasks. The implementation of statistical techniques, methods and methodologies can be done with R functions, scripts and packages in a proper manner following the principles of modularity. Moreover, some essential characteristics such as open source, user-friendly, functional programming, object-oriented can be met by taking advantage of the high degree of development of R packages related to the discipline of Official Statistics. In this talk, we will overview the most important challenges that can be undertaken by the R language implementing an innovative end-to-end statistical production process through a real-life use case. This presentation will showcase a novel process that combines machine learning techniques, time series filtering, and benchmarking. Utilizing existing R packages, this innovative approach achieves a significant improvement in time granularity.

References

No References available