uRos (Use of R in Official Statistics)

Book of Abstracts (2023)

Contents

Access to official statistics from R: an overview

Authors

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

Abstract

Providing access to output data is an essential task for Statistical Institutes. This is reflected in the awesome list of official statistics software [1] and the CRAN Task View: Official Statistics & Survey Statistics [2]. The awesome list was created in 2017 and grew over time [3] with many contributions from the official statistical community, partly from uRos workshops. Currently the vast majority of packages on the list is R software. The list is organised according to the Generic Statistical Business Process Model (GSBPM) [4], see Figure 1. The largest category on the list is "Access to official statistics" which contains over 30 software packages that help users access official statistics data or metadata from International or National organisations. In this presentation we take a closer look at the R-packages on this list to describe the current state of access to official statistics from R and we suggest potential improvements to this software landscape. There are 28 R-packages in this category. Some are more generic and targeted at standardised access to multiple data providers, others contain detailed and dedicated functionality to access just one national or international organisation. Figure 2 provides an overview. It shows the dependencies between the packages on the list, the statistical data providers and the standards being used. These relationships were derived from the R-packages documentation, the pages they link to and running some of the packages, such as rdsmx which offers a list of pre-configured data-providers. Only main data providers on (inter)national level were considered. We note that this is an abstraction of reality as for example different versions of standards and endpoints are not taken into consideration, which would complicate the figure considerably. From this network we can identify standards-based packages such as rsdmx, readsdmx using the SDMX standard [5], rjstat offering access to the JSON-STAT data format [6] and various px* packages targeted at the PC-Axis format [7]. The ODATA [8] standard is used by one organisation: statistics Netherlands. Others packages such as inegiR, readabs and statcanR provide dedicated access to just one specific data provider. Note that it could be that a package does use a standard internally but this is not mentioned in the documentation as the package writer tries to hide these details from the user. In that case the use of standards is underestimated here. This work in progress can be followed here: https://observablehq.com/@olavtenbosch/access-to-official-statistics-from-r Looking from the data provider side we can see that certain data providers can be accessed via multiple R packages. Eurostat (ESTAT) and supporting SDMX as well as JSON-STAT and the existence of two dedicated packages restatapi and eurostat is a clear example. For the rest, the set of providers offering JSON-STAT data is mostly disjoint from the set of SDMX providers. All PX providers provide JSON-STAT as well. Some organisations, such as Eurostat and the World Bank, provide multiple endpoints for specific domains. Some endpoints provide harmonised data on one specific domain via an dedicated R-package. Examples are rdbnomics offering access to economic data from many institutions and ipumsr providing access to census and survey data integrated across time and space. Although a special category it is useful to note the existence such official statistics aggregator sites and their dedicated R-packages in the official statistics open data landscape. The list R-packages on the awesome list also provide us some insight into the functionality that is usually offered. We can see certain features reoccurring, such as:endpoint hiding: wrapping the preconfigured endpoint(s) in a R function within the package;

catalogue retrieval: the ability to list the availability datasets on the endpoint(s); search: the ability to search for datasets or within datasets on the endpoint(s); endpoint queries: the ability to query for subsets on the endpoint(s) side; local queries: the ability to easily slice or filter the retrieved data on the client; caching: preventing unnecessary roundtrips to the endpoint(s) by caching results; cartographic queries: retrieve a (cartographic) map to be used with the data. A category, not covered so far, is access to statistical metadata. Many organisations, mostly international, offer access to definitions, classifications and code lists in metadata registries. These predominantly use SDMX. Access from R to SDMX metadata has proven to be useful for statistical operations, such as checking data against internationally harmonised code list in the validation process [9]. Some organisations offer metadata in the form of linked data. Examples are Eurostat and Statistics Netherlands [10]. Linked data has the promise to make it easier to link and re-use statistical content with other open data sources, aligning to the FAIR principles [11]. Linked data can be accessed from R via generic software, such as rdflib, jsonld, or more experimental packages such as glitter if the endpoint provides for a queryable linked data interface in SPARQL. All in all, we expect that the growing use of linked (meta)data in official statistics will positively influence the official statistics open data landscape in the near future. From the above, we see that the official statistics open data landscape grows towards standardisation, but also that in many cases there is a need to develop dedicated software targeted at specific functionality or specific data providers. At this point the R user can choose from at least 28 packages to access official statistics, each offering different functionality. There is no 'one-for-all' R-package that provides access to all official statistics data providers. This is understandable from an organisational viewpoint, but from an end-users viewpoint a generic package would be convenient. The situation will improve with ongoing standardisation on the data providers side, however as an uRos community it might be useful to also work towards creating a generic interface to all official statistics data providers, notwithstanding the value of the dedicated packages. The analysis presented in this paper could serve as a start.

References

- [1] Awesome list of official statistics software, https://github.com/SNStatComp/awesome-official-statistics-software; [2] CRAN Task View: Official Statistics & Survey Statistics, https://cran.r-project.org/web/views/OfficialStatistics.html; [3] Olav ten Bosch, Mark van der Loo, Alexander Kowarik, The awesome list of official statistical software: 100 … and counting, The Use of R in Official Statistics - uRos202 (virtual); [4] Generic Statistical Process Business Model GSBPM, version 5.1, UNECE, https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1; [5] Statistical Data and Metadata eXchange (SDMX), https://sdmx.org; [6] JSON-stat: A simple lightweight standard for data dissemination, https://json-stat.org/; [7] PC-Axis software family, https://www.scb.se/en/services/statistical-programs-for-px-files/; [8] OData (Open Data Protocol), https://www.odata.org/; [9] Olav ten Bosch, Mark van der Loo, (2021), Validation in R Using Metadata from SDMX Registries, 8th SDMX Global Conference, INEGI, Mexico (virtual); [10] ten Bosch O, de Jonge E, Laloli H, Laaboudi-Spoiden C (2022) FAIR Digital Objects in Official Statistics. Research Ideas and Outcomes 8: e94485. https://doi.org/10.3897/rio.8.e94485; [11] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

An automated machine-learning pipeline for statistical matching

Authors

- Theresa Küntzler (Destatis)

Abstract

Statistical matching is a technique to add variables to a data set that are initially only available in a second data source. Thereby, it becomes possible to investigate new research questions without the need for additional data collection. Instead of gathering new data, however, statistical matching requires building a well-fitting model to estimate the variables that should be matched, which is no trivial task and itself requires time and effort. This talk presents an automated data processing pipeline developed at Destatis to support and facilitate statistical matching. The new pipeline enables data scientists to efficiently carry out the complete matching procedure within short time frames, while not compromising on high methodological standards. The pipeline uses a donor data set as input and automatically performs the following steps: (1) Data preprocessing, including simple imputations and more. (2) A variety of possible models is trained and tuned using default hyperparameter search spaces1, wrapped in a nested resampling to then (3) evaluate each model on its performance. (4) The best performing model is retrained and -tuned on the full data set and finally (5) used to estimate the variables of interest in the receiving data set. (6) In addition, the pipeline automatically produces a report containing information about the data sets in use, the distributions of the predictors, the performance of all tested models and the relationships between the variables of interest. The pipeline is implemented using the mlr32-package in R. The presentation will describe the architecture of the pipeline in detail, highlight and explain methodological decisions and illustrate the technical implementation.

References

- 1 Becker M (2023). mlr3tuningspaces: Search Spaces for 'mlr3'. https://mlr3tuningspaces.mlr-org.com, https://github.com/mlr-org/mlr3tuningspaces.; 2 Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B (2019). "mlr3: A modern object-oriented machine learning framework in R." Journal of Open Source Software. doi:10.21105/joss.01903, https://joss.theoj.org/papers/10.21105/joss.01903.

Assessing coherence between estimated distributions in R

Authors

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

Abstract

The quality of statistics produced and disseminated by a National Statistical Institute is anchored to the "fitness-for-use" paradigm, i.e. whether and how statistics meet the users' needs. Quality of statistics is a multidimensional concept that encompasses a set of interrelated criteria; the European Statistical System (ESS) sets five dimensions: relevance; accuracy and reliability; timeliness and punctuality; coherence and comparability; and, accessibility and clarity (these dimensions correspond to the principles 11-15 of the European Statistics Code of Practice). This note aims at investigating the coherence of statistics. In fact nowadays, assessing coherence has becomes crucial not only at the end of a statistical process, to measure coherence with outputs produced by a different division within the same statistical agency (cross-domain coherence) or by another agency being part of the National Statistical System (NSS) (cross-agency coherence), but also at the beginning of a modern process statistical production process involving integration of data stemming from different sources (integration of sample survey data, integration of survey and administrative data, use of big data for production of official statistics, …). In this latter case coherence helps to ascertain whether variables available in two or more of the available data sources (having the same or a slightly different definition) share the same marginal distribution and consequently can be directly compared (for matching purposes, for using them as a predictor in statistical models, etc.). This work shows how to assess coherence when dealing with a categorical nominal variable that can admit two or more categories (for instance: marital status, education level, professional status, level of satisfaction, etc.) by using very simple indicators already available in the R environment. In particular the paper considers measures of distance between estimated distributions of categorical nominal variable, like the total variation distance or the Hellinger distance. The work tries to address also the problem of evaluating coherence between estimated marginal distributions of the same continuous variable. The R environment provides different facilities spread over a number of packages that however require a careful application as often they are not suited to manage data from complex sample surveys with unequal weighing system. In this case development of ad-hoc R codes may be needed.

References

No References available

At the coRe of UNIDO's revamped statistical publications

Authors

- Martin Haitzmann (UNIDO)

Abstract

The International Yearbook of Industrial Statistics is the statistical flagship publication of the United Nations Industrial Development Organization (UNIDO). It provides a snapshot of industrial development trends across the world. Following a publication style that for a long time contained similar types of tables, the publication has been completely redesigned with the aim of making statistical information more accessible and practical. The electronic publication now consists of short analyses, infographics, charts, and other data visualizations. Given limited budgetary resources, but also given the powerful and versatile open-source software available, we opted for a process based on R(markdown) and LaTeX for the pdf publication (the pdf publication was given higher priority for several reasons, but the option of embedding an html version of the publication in our website is on the to-do list). Team collaboration and version control should be ensured with the use of git. In case of success and positive feedback, we wanted to use the basic template with its styles and elements for all our publications and reports, as this would allow for efficiency gains, foster synergies between different statistical products and a uniform, common "look" for all our publications. When this vision became clear enough, we decided to take up the challenge without any further external help (e.g., from graphic designers or publication experts). With this presentation, we want to share some of our experiences in implementing this process. In addition to careful planning and development, flexibility and competence in problem solving were essential, as a variety of problems arose. These were not only R- or Latex-specific coding problems, but also, for example, restrictions on the use of open-source software due to (changing) IT group policies. Currently, all our regular quarterly reports, topic-specific reports (SDG 9 progress report) and methodology/working papers have already been converted to this template or will be. Despite continuous adaptations, we can call it already an enormous success given the positive feedback we have received. Moreover, some by-products of the process, such as a customized ggplot theme package and useful plot functions, have emerged and are now a central part of our daily work.

References

No References available

data.table and Rcpp, two possibilities to speed up your R processing

Authors

- Alexander Kowarik (Statistics Austria)

Abstract

Data transformation tasks are often a big part of any data science / statistical production process. Depending on the size of the data sets, this part can also use a significant chunk of computation time. With the R package data.table it is possible to speed up classical ETL processes a lot. After getting used to the data.table syntax, it will also speed up the implementation time of new process steps. Another part of processing are often adhoc implementation of iterative processes, e.g. benchmarking of independently estimated aggregates against each other. If such process steps are implemented as base R for-loops, they can take a lot of computation time. In this tutorial we will look at the Rcpp package as any easy and comfortable way to transfer some tasks in C++ and speed up the processing by it. Sound knowledge of base R and basic knowledge of C++ is recommended for this tutorial.

References

No References available

Difference on Evaluation Scores Considering Image Descriptions

Authors

- Yukako Toko (National Statistics Center, Japan)
- Mika Sato-Ilic (Institute of Systems and Information Engineering, University of Tsukuba, Japan)

Abstract

In official statistics, text response fields are often found in survey forms. Coding tasks, translating text descriptions into corresponding classes, are usually performed on those respondents' text descriptions for data processing. Coding tasks are originally performed manually, whereas the studies of autocoding have made progress with the improvement of computer technology in recent years. As a typical computer technology, many techniques in natural language processing have been proposed in artificial intelligence. One of the typical surveys in Japanese official statistics is called the Family Income and Expenditure survey. The data are used in various fields, such as policy planning, econometric analysis, and market research. It also supplies basic data for the calculation of such macroeconomic figures as the Gross Domestic Product and the Consumer Price Index. In this survey, households are asked to keep their daily incomes and expenditures. The data obtained from households includes purchased items' names in short text descriptions, including text descriptions extracted from images of shopping receipts. Recently, obtaining images of shopping receipts has increased, and the ratio of receipts images has become larger than that of text descriptions. In addition, the importance of investigating various criteria for evaluating the results of classification has been emphasized in various areas, including official statistics. Therefore, evaluating results on various criteria considering dynamical trends of difference in data forms such as receipts images or short text descriptions is essential for obtaining correct data. Hence, this paper presents results of evaluation criteria for autocoding with respect to the trends of different forms of data collection.

References

No References available

Elevating Data Documentation Proficiency: Harnessing the Power of DDI for Official Statistics

Authors

- Adrian Dusa (University of Bucharest)

Abstract

In the current era defined by data-driven decision-making, the significance of robust data documentation cannot be overstated. This training session will equip the audience with the skills and knowledge necessary for effective data documentation, preparing for long term preservation and dissemination. We will delve into the growing importance of the Data Documentation Initiative (DDI) standard, focusing on the DDI Codebook to document individual data sets. This training will introduce two key R packages called DDIwR and declared, to achieve interoperability with other statistical formats and generate DDI Codebook templates, offering a roadmap for improving data transparency, accessibility, and quality in the evolving landscape of official statistics.

References

No References available

From SAS to R: moving to open source solutions in Romanian NSI

Authors

- Bogdan Oancea (Romanian National Institute of Statistics)
- Marian Necula (Romanian National Institute of Statistics)

Abstract

Official statistics agencies, ours included, play a crucial role in collecting, analyzing, and disseminating data that informs policy-making, research, and public understanding. Since this year we have relied on proprietary software SAS (Statistical Analysis System) for several needs such as data processing for social statistics or implementing sampling methodologies. However, over the last years we are witnessing a growing trend toward transitioning to open-source solutions like R in official statistics. More than that, we are also facing budget cuts for buying software licenses. Therefore, we started an internal project to translate the existing SAS solutions used in our NSI to R. Nevertheless, this is not an easy task; in fact, it proved to be a complex and resource-intensive process. Among the difficulties we encountered we can mention the poor knowledge of SAS programing language of the members of our team, the need for assuring the same quality requirements for the new R solutions, the skills gap regarding R software system for the statisticians in charge with running these programs. In order to facilitate a smooth transition from SAS to R we started with a comprehensive planning. Our plan includes a thorough assessment of existing SAS software pieces, a timeline for migration, and strategies for addressing data compatibility issues. We also envisaged a period of training and capacity building in order to build R proficiency among staff. We adopted an incremental implementation strategy by starting with smaller projects or specific statistical domains when transitioning, allowing for iterative improvements and minimizing disruption to ongoing data collection and analysis. A side effect of the project is the opportunity to fit the current statistical production process chain, based on legacy SAS, to GSPBM (Generic Statistical Business Process Model) and to design the software according to the well-known and well-test Waterfall methodology. Besides manual re-writing of the SAS scripts, we also tested the usage of LLMs such as Chatgpt or Google Bard to automate the translation but the results were poor: while these AI tools correctly translate small code snippets, they fail to translate larger scripts.

References

No References available

Improvements on sdcSpatial: privacy protected maps

Authors

- Edwin de Jonge (Statistics Netherlands)

Abstract

Plotting data on a map is a popular and helpful tool to analyze spatial data. R makes it easy to plot spatial data with packages such as ggplot2, tmap, mapview or leaflet. When plotting the spatial distribution of a sensitive variable, e.g. income or unemployment, you may accidentally reveal a sensitive value of an individual observation. sdcSpatial provides Statistical disclosure control (SDC) methods to create and assess disclosure risk of spatial distribution maps and methods to suppress or adjust locations with revealing sensitive values [1, 2, 3]. The presentation will introduce the package and will discuss the latest improvements on sdcSpatial (version 0.6), including performance and information loss metrics. It will also discuss some of the results of the user experiments with sdcSpatial [4] that have been performed to compare the method for different countries. It will end with an outline on the further developments.

References

- [1] Suñé, E., Rovira, C., Ibáñez, D., Farré, M. (2017). Statistical disclosure control on visualising geocoded population data using a structure in quadtrees, NTTS 2017; [2] de Jonge, E., & de Wolf, P. P. (2016, September). Spatial smoothing and statistical disclosure control. In International Conference on Privacy in Statistical Databases (pp. 107-117). Springer, Cham.; [3] de Wolf, P. P., & de Jonge, E. (2018, September). Safely Plotting Continuous Variables on a Map. In International Conference on Privacy in Statistical Databases (pp. 347-359). Springer, Cham.; [4] Spatial SDC experiments and evaluations – multiple countries comparison, UNECE Expert Meeting on Statistical Data Confidentiality 2023

Integrating a System for Automatic Classification of Economic Activities into Statistical Production: Challenges and Solutions

Authors

- Athanassia Chalimourda (Swiss Federal Statistical Office (SFSO))
- Lorenz Helbling (Swiss Federal Statistical Office (SFSO))

Abstract

As part of the Swiss Federal Statistical Office's (SFSO) innovation project with the same name, NOGAuto is an automatic assistance system for classification of economic activity descriptions, which come primarily from the Register of Commerce among other sources. It assigns activity descriptions in different languages to categories of the General Classification of Economic Activities (Nomenclature générale des activités économiques – NOGA, the Swiss "Nomenclature statistique des activités économiques" - NACE) with a certain probability. In order to take account of Switzerland's multilingualism, in a first step, NOGAuto detects the language out of a possibly very short activity description. It continues performing Natural Language Processing and text classification with a Gradient Boosting Machine. NOGAuto and the associated shiny application were developed in R by the Business Registers Data section of the SFSO, with the aim of providing an assistance tool for coding economic activities. In the present joint work between the Statistical Methods and the Business Registers Data sections, the integration of NOGAuto into the statistical production has been assessed. We highlight the challenges that arose and discuss solution approaches. Classes of economic activities are inherently imbalanced, with some having few activities' descriptions comparing to others. This can lead to inaccurate predicted probabilities, which can in turn result in over- and under-representations of economic activity classes. We describe the applied measures to assess quality and how these can help to identify areas in the classification process that can further be improved or where it seems advisable that experts complete the classification task. Furthermore, possible sources of bias in the NOGAuto pipeline could arise during the interaction with coding experts, or other expert systems, each of which has its own strengths and weaknesses. Connecting NOGAuto and the shiny application to an automatic translation service like DeepL not only saves time in the modeling of language detection but also allows coding experts to work in the language they feel most comfortable in. The combination of NOGAuto with an SFSO-internal rule-based classification system can further improve convenience, efficiency and user-friendliness. A revision of the NOGA classification, in line with the NACE revision, reveals one kind of intrinsic activities that can lead to model drift in assistance systems like NOGAuto. The risk of model drift because of the evolution of the economic activities in time combined with the fact that the area of natural language processing itself undergoes rapid change show the need for the definition of regular revision circles. The coding of economic activities directly affects business surveys, thus making the need of a sound quality assurance monitoring even more evident.

References

No References available

Managing our R infrastructure with 400+ users

Authors

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

Abstract

Managing an R infrastructure for 400+ R users can be a challenging task. In addition to providing software and resources, it is also important to provide users with support. It is also important to monitor the R infrastructure to ensure that it is performing well and that users are not experiencing any problems. Another important aspect of managing an R infrastructure is to provide users with enough resources to run their workloads while preventing single users to exhaust the limited resources on their own. We will present attempts to automate user generation, monitoring and other crucial steps as much as possible and how the R infrastructure itself can be used for these tasks with shiny apps and scheduled reports.

References

No References available

Modernizing Statistical Production Systems Using R

Authors

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

Abstract

It is widely known that National Statistical Offices have are facing challenges both in the area of data collection and output production. On the input side, NSIs are increasingly relying on volatile data sources including administrative data, sensor data, web data and other big data sources, and (in the future) data from private sources. On the output side NSIs are increasingly expected to quickly respond to current events and needs in society. These trends put considerable strain on the statistical production chains of NSIs. Considering the increasing volatility, NSIs need to move from a situation where production chains are static over time, to a situation where change and innovation is the norm. Production systems will be permanently under construction. One way to cope with this situation is to build production systems using a modular approach. Although this idea is hardly new, the question of module granularity, which functionality a module should offer (and which not) remains largely undiscussed in the official statistics community. In this talk I will discuss what I have learned about this approach over 15 years of R package development. It turns out that figuring out where a module starts and where it ends is surprisingly difficult. I will highlight technical issues as well as issues related to design, module development and the user perspective. Finally, I will highlight properties of the R language that make it especially suitable for a modular approach.

References

No References available

Official statistics and report automation and replication using R

Authors

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

Abstract

Many reporting tasks are either highly repetitive, for example when one has to write a report about inflation that changes every month, or a report about monitoring food prices that change every week. Similarly, there many parts of a report that can be applied to others reports as well. R can help users in automate these repetitive tasks in many ways and at different stages of the workflow: 1. API interfaces to almost all official statistics providers (Eurostat, OECD, UN, etc) to facilitate data download; 2. A large set of tools and packages that provides an efficient and easy way to data wrangling, manipulation, plotting, etc.; 3. Integration with Quarto, a new but extremely efficient tool to produce high quality .html, .pdf, .docx or .pptx documents. The current proposal deals with the combination of above three features applied to report automation using monthly data of inflation and agricultural products from Eurostat database.

References

No References available

Programming and collaborations: roles and tools when using R.

Authors

- Isabella Gollini (UCD School of Mathematics and Statistics, Ireland)

Abstract

Programming is a crucial part of a data scientist's job and it is generally a collaborative endeavour. We may collaborate for many different reasons such as new applications, or to develop new methods, improve efficiency in existing methods, apply or create some user-friendly package software or interface. In this talk, I will go through some of the main aspects associated to the different roles that we can assume when collaborating in programming. I will then focus on the tools we can use to organise and share our projects in the context of R programming.

References

No References available

R & Jenkins - Open Source for Open Government

Authors

- Philipp Bosch (Statistical Office of the Canton of Zurich)

Abstract

Inside the Canton of Zurich, the Statistical Office is the competence center for public statistics. It documents and analyzes significant social and economic developments in the canton and economic region of Zurich. In addition, the Statistical Office is responsible for the operations and continued development of the Open-Government Data-Catalog of the canton, which provides the general public with high-quality and relevant data. More than ten years ago, the Statistical Office made the strategic decision to shift its data related pipelines and processes to R and open source tools in general. In the meantime, R serves as the backbone of every "Data Product" the Statistical Office delivers. Therefore, a whole ecosystem of packages and dedicated workflows in R has been created in order to ensure reproducibility and scale productivity concerning recurring tasks. In the presentation, I want to zoom in on one particular part of our R setup which showcases how R is brought into production using the Open-Source tools gitea and Jenkins. Gitea is a self-hosted git service and almost a one-to-one replacement of Github. Jenkins on the other hand is an automation server, a CI/CD tool. Combined with R, both not only offer great value for classical CI/CD tasks but are also leveraged as workflow tools inside the Statistical Office. By using examples from the daily work in the Statistical Office I want to demonstrate the variety of tasks in which the setup of gitea, Jenkins and R can reap the power of R in production. This includes simple recurring maintenance tasks and checks of pipelines as well as the daily harvesting, cleaning and publishing of Open-Government-Data. However, the talk is not only intended as an example of a successful implementation of R in production, but also to point out stumbling blocks and caveats. Ultimately, we as the Statistical Office constantly try to improve our processes and learn from the experience of other members of the community.

References

No References available

R Shiny apps for asymmetry analysis via selective editing

Authors

- Mauro Bruno (Italian National Institute of Statistics (ISTAT))
- Maria Serena Causo (Italian National Institute of Statistics (ISTAT))
- Giulio Massacci (Italian National Institute of Statistics (ISTAT))
- Francesco Ortame (Italian National Institute of Statistics (ISTAT))
- Giuseppina Ruocco (Italian National Institute of Statistics (ISTAT))
- Simona Toti (Italian National Institute of Statistics (ISTAT))

Abstract

The introduction of intra-EU export Micro-Data Exchange (MDE) provides National Statistical Institutes with a new data source to compile intra-EU import statistics. This approach tackles two key challenges: diminishing the overall response burden on data providers and meeting user expectations regarding the quality of the produced statistics. However, before transitioning to a data production system based on MDE data, it is essential to evaluate the coherence and comparability issues between the MDE and National import data. To standardize and enhance the identification of potential asymmetries between the two data sources, Istat developed an innovative set of open-source interactive tools designed to foster cooperation among Member States. These tools were developed using R through the Shiny package. They include exploratory analysis, systematic error detection, and selective editing (Generic Statistical Data Editing Model, UNECE, 2019). The core of this approach is the dynamic computation of a relative contribution index and an asymmetry suspicion index (Jäder A., Norberg A., 2005). The most relevant asymmetries are identified through user-defined numerical thresholds. Although a more in-depth study of the discrepancy between the two sources is needed, the achieved results are encouraging. Sharing the developed solution within the European Statistical System enhances interoperability, promotes method harmonization, and encourages the adoption of official statistical standards.

References

- 1. Generic Statistical Data Editing Model (GSDEM - Version 2.0, June 2019). Available from: https://statswiki.unece.org/display/sde/GSDEM; 2. Jäder A., Norberg A. (2005), A selective editing method considering both suspicion and potential impact, developed and applied to the Swedish foreign trade statistics, paper presented at the Work Session on Statistical Data Editing, Ottawa, Canada, 16-18 May 2005.

Renewable energy and GHG emissions in European Union countries

Authors

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

Abstract

Goal 13 of the UN 2030 Agenda refers to “climate action”, i.e. “urgent action to combat climate change and its impacts”. The term “climate change” usually refers to those significant long-term changes in the patterns and characteristics of Earth’s global or regional climate that affect society and various vulnerable ecosystems. Electricity production by burning fossil fuels is one of the most important greenhouse gas (GHG) emitters. In the paper, using causality analyses (Granger, Dumitrescu-Hurlin, Toda-Yamamoto) and through econometric models such as VAR, VEC and ARDL, we estimate the impact of increasing the share of renewable energy in production and final consumption on the GHG intensity of economic activities. The models are estimated with panel data for EU-27 and detailed for Romania. We identified (and measured) positive feedback relationships between the share of renewable energy in the energy mix and the intensity of total energy consumption in GHC emissions, relationships mediated by economic growth and energy intensity (as control variables).

References

No References available

Split-Apply-Combine with Dynamic Grouping

Authors

- Mark van der Loo (Statistics Netherlands)

Abstract

Groupwise aggregation is one of the most common operations in data analyses. There are use cases where the grouping is determined dynamically by collapsing smaller subsets into larger ones to ensure sufficient support for the target aggregate. Examples include cases where some of the target groups suffer from (unit or item) nonresponse, or cases where the quality of target group data is judged to be too low. Often, hierarchical classifications serve as a basis for forming larger groups, but custom 'collapsing schemes' are in use as well. In this presentation we demonstrate the R package 'accumulate'[1] that offers interfaces for defining grouped aggregation, where the grouping may be dynamically determined, based on user-defined aggregations, user-defined decision rules, and user-defined collapsing schemes. The package offers several ways to define collapsing schemes, including tabular definitions that can be maintained separately from the aggregation code. It also includes facilities to use hierarchical classifications and for testing the (possibly complex) decision rules that user can create.

References

- [1] Van der Loo M (2023). accumulate: Split-Apply-Combine with Dynamic Groups. R package version 0.9.0, https://cran.r-project.org/package=accumulate.

Survey Monitoring with Paradata: Leveraging the Survey Solutions Paradata Viewer Application in R

Authors

- Michael Wild (World Bank, Austria)
- Ciprian Alexandru (Ecological University of Bucharest, Romania)

Abstract

Paradata, the invaluable data collected alongside survey responses, plays an essential role in ensuring data quality and survey performance. In this article, we explore the application of paradata for survey monitoring, shedding light on the Paradata Viewer Application and its integration with the R environment. This innovative tool empowers survey researchers and practitioners with a comprehensive framework for leveraging paradata in survey monitoring, even in a real-time. Using case studies and empirical findings, we illustrate the significant impact of paradata analysis on survey data quality and operational efficiency. We start presenting the methodology behind paradata analysis and discuss the practical features of the 'susoparaviewer' Application. Real-world case studies highlight its application, demonstrating substantial improvements in survey data quality, including census data, and revealing actionable insights for survey administrators. This article showcases the pivotal role of paradata in survey research and presents a practical solution to harness its potential for improving data quality. Our discussion explores the implications of implementing such tools in survey research, emphasizes the importance of paradata in the data collection process, and addresses potential challenges and limitations. By presenting the Paradata Viewer Application as a valuable resource for survey monitoring, we aim to contribute to the growing body of knowledge in the field and encourage further research in this area. We conclude with a forward-looking perspective on the future directions and possibilities for enhancing survey quality through paradata analysis.

References

No References available

The bbkplot package: Deutsche Bundesbank Corporate Design in R – Motivation, Challenges, Benefits

Authors

- Hendrik Christian Doll (Deutsche Bundesbank, Research Data and Service Centre)
- Daniel Ollech (Deutsche Bundesbank, Research Data and Service Centre)

Abstract

We introduce the bbkplot package, in order to facilitate user-friendly and reliable production of data-driven graphics in accordance with Deutsche Bundesbank's corporate design. The presented package allows conversion of figures and tables into graphics precisely adhering to corporate design standards. These standards specify formatting requirements for graphics in great detail in order to ensure consistency in public brand appearance. bbkplot automates the exact application of these requirements and enables quick and easy production of high-quality graphics by the user. Challenges we need to handle in order to create a fully flexible package include detailed design requirements making it necessary to rely on the low-level grid package to construct graphics. Furthermore, consistent import of specific proprietary fonts across all (virtual) machines in the IT infrastructure was a challenge. We see two value propositions from the package. (1) By advancing from pre-specified graphic types that traditional graphics suites implement to flexible plot creation, bbkplot efficiently allows creation of a wider range of graphics for official statistics to publish in corporate design; (2) by facilitating low-threshold user-driven creation of simple graphics for presentations, we support a more comprehensive external brand appearance, which is important for official statistics to convey the reliable origin of information.

References

No References available

Transitioning Sample Surveys to R: insights from Statistics Lithuania

Authors

- Donatas Šlevinskas (State Data Agency. Statistics Lithuania)

Abstract

Statistics Lithuania is currently in the process of migrating sample surveys to R. This report aims to provide a detailed overview of the strategic migration from legacy software to R for the official sample surveys. The paradigm shift reveals the compelling advantages of R's robust analytical capabilities and automation skills. Statistics Lithuania aims to optimise survey workflows, enable automation and leverage parallel computing capabilities by using the recently introduced SaaS platform. The main differences and challenges associated with this transition will be highlighted: data management and preparation, sample design and weighting, analysis and visualisations. In addition, practical insights are provided by presenting specific surveys that have been successfully migrated to R, thus emphasising the practicality and impact of this transition.

References

No References available

Use of Extreme Gradient Boosting to predict

census variables

Authors

- Lorenzo Asti (Statistics Italy (ISTAT))
- Loredana Di Consiglio (Statistics Italy (ISTAT))
- Tiziana Pichiorri (Statistics Italy (ISTAT))

Abstract

Starting from 2018, the traditional Population census paradigm has changed. Sample data are collected yearly and statistical information is produced by the joint use of samples and administrative data. At this scope extensive use of modelling has been applied. In particular here we discuss the method and the tools applied for the estimation of the place of work of the employed population. Information on place of work is partly available in different statistical sources where work locations are recorded. However, the administrative information cannot replace the statistical survey because it often contains the location of the enterprise only and not those of the local units. Moreover the statistical register does not cover freelance workers, cannot distinguish situation for which there is not a fixed place of work and it is under-covered for employed in foreign countries. The use of such data is extremely valuable and can be integrated with surveydata though statistical modelling. Previous similar estimation problems where solved by the use of multinomial logistic regression modelling (Ciccaglioni et al.2023). In this case traditional multinomial logistic regression was not enough satisfactory, in terms of bias, that we measure in comparison with higher aggregation of the direct estimates. We applied the eXtreme Gradient Boosting to predict probabilities of the region of work for each employed person in the population register. The whole procedure was done in R. The package caret was used to call the xgboost package implementing eXtreme Gradient Boosting and to learn models by selecting the approrpiate configurations of hyperparameters through cross validation. In particular caret permits to easilly define custom objective functions involving single individuals inferred propabilities; this revealed to be crucial in order to obtain good results in terms of aggregated data.

References

- Carolina Ciccaglioni, Loredana Di Consiglio, Tiziana Pichiorri, Fabrizio Solari Masurement of daily commuting in the Italian permanent population census, SIEDS 2023

Use of R at NIS Romania: Focus on Calibration in Household Surveys

Authors

- Ana-Maria Ciuhu (Romanian National Institute of Statistics)
- Monica Rafu (Romanian National Institute of Statistics)

Abstract

Effective and accurate calibration of survey data remains a challenging aspect that can significantly impact the quality of statistical outputs. This presentation outlines the pioneering use of the R programming language at the National Institute of Statistics (NIS) in Romania, specifically employing the ReGenesees package, to tackle the calibration challenges. A comprehensive case study detailing our migration from traditional calibration software (i.e., SAS) to an R-based workflow will be presented. The focus is on the benefits, challenges, and practical insights gained through the implementation of the ReGenesees package, an R package tailored for generalized calibration.

References

No References available

Variance estimation with the R package gustave: the experience of STATEC

Authors

- Claude Lamboray (STATEC)
- Dmitri Lebrun (STATEC)
- Guillaume Osier (STATEC)

Abstract

The variance is an important quality measure commonly used in survey sampling. Within official statistics, there is an increasing need for variance estimation in the context of quality reports but also for monitoring compliance with statistical legislation. The R package gustave is a toolbox for analytical variance estimation. It mainly consists of constructing a “variance wrapper” that incorporates all the methodological elements required for analytical variance estimation. For example, the variance may depend on the sampling design, on the way that non-response is modelled or on the calibration that is applied. Once finalized, the variance wrapper can be easily used by the subject matter expert for variance estimation during production. The R package gustave already contains a pre-defined variance wrapper that can be applied to many of the surveys that are conducted at STATEC. In addition, customized variance wrappers can be constructed for more complex situations. Beyond some common parameters such as the ratio, the package makes it possible to implement additional linearization formulas for other non-linear parameters. Moreover, domain estimation is conveniently handled when calling the variance wrapper. The package includes a number of checks to verify both data formats and methodological consistency. This feature adds robustness to the analysis. In this presentation, we will describe the variance estimation methodology for the EU-SILC survey, explain how we implemented this methodology within gustave and discuss the main results. Finally we will highlight the lessons learned when applying this R package in the context of STATEC.

References

No References available