uRos (Use of R in Official Statistics)

Book of Abstracts (2022)

Contents

A hybrid estimation of distribution algorithm for joint stratification and sample allocation

Authors

- Mervyn O'Luing (Insight Centre for Data Analytics, Department of Computer Science, University College Cork, Ireland)
- Steven Prestwich (Insight Centre for Data Analytics, Department of Computer Science, University College Cork, Ireland)
- S. Armagan Tarim (Cork University Business School, University College Cork)

Abstract

In this study we propose a hybrid estimation of distribution algorithm (hEDA) to solve the joint stratification and sample allocation problem. EDAs are stochastic blackbox optimization algorithms which can be used to estimate, build and sample probability models in the search for an optimal stratification. We enhance the exploitation properties of the EDA by adding a simulated annealing algorithm to make it a hybrid EDA. Results of empirical comparisons for atomic and continuous strata show that the HEDA attains the best results found so far, when compared to benchmark tests on the same data using a grouping genetic algorithm, simulated annealing or hill-climbing. The R package hEDA presents this algorithm for use in joint stratification and sample allocation designs, enabling comparison with existing algorithmic approaches.

References

No References available

A shiny application for processing life tables in Czechia

Authors

- Jiří Novák (Statistics (Census) Unit for Coordination of Census Preparation and Processing, Czech Statistical Office)
- David Morávek (Czech Statistical Office)

Abstract

The contribution introduces a shiny user-friendly web application in the environment for the calculation of life tables, visualization, and their export to the structure of a publication intended for the researchers and the general public. It contains an algorithm for calculating life tables that allows various settings of input parameters, including the selection of a suitable statistical model for smoothing and mortality modelling. The application was developed within the International Data Science Accelerator program in 2022 with the main goal of simplifying the process of processing and publishing life table data at the Czech Statistical Office and will be freely available to the public within the R package of CZSOLifeTables

References

No References available

About planning a Revolution - Lessons Learned

Authors

- Elisa Oertel (Statistical Office of Lower Saxony, Germany)

Abstract

The work at the Statistical Office of Lower Saxony includes the collection, processing and publication of data from almost 300 statistics. These tasks are increasingly being taken over by specialized software products from external providers. However, depending on commercial software comes with risks and limitations. Useful open source software such as Keycloak, Docker, Python, R and Shiny is largely unknown or unused in official statistics, not only in Lower Saxony, but apparently throughout Germany. In order to show the untapped potential, our group tested the use of R and Shiny in the context of one of the currently largest legacy IT replacement problems in official statistics in Germany and came to surprising results. We have found that R works much more efficiently than all the software used so far and that using Shiny also increases user-friendliness to a level that users nowadays expect from an application or program. In addition, IT tasks such as versioning, maintenance and updating of the application were significantly improved. Nonetheless, people in decision-making positions seem to have many reservations about the use of R and its introduction in Lower Saxony has not been without difficulties. Rather, the publication of our results sparked an old debate about R itself and also about the question of how much in-house development official statistics actually need. In my presentation, I will focus on the lessons learned in the process of establishing R as an accepted standard in Germany's official statistics and on what our group has achieved so far.

References

No References available

Advantages of using R to enhance the analysis of women homicide in Mexico

Authors

- Dafne Gissel Viramontes Ornelas (INEGI-Mexico)

Abstract

Recently, the use of R as a tool for data analysis has gained relevance within the social sciences. Its influence lies in the fact that it has shown to improve the processing of large amounts of data and has enabled operating databases more efficiently for the study of social phenomena. In this sense, this paper aims to present a practical application of R software in data analysis from a gender perspective. Based on a conceptual proposal and using official statistics of deaths from homicide published by the National Institute of Statistics and Geography of Mexico (INEGI, by its acronym in Spanish), we classify women homicides to determine in which circumstances they occur. The results allow us to have an approach to identifying the main factors involved in women's homicide in Mexico. This seeks to highlight the importance of open-source software in social phenomena analysis and how it can contribute to closing the gaps in the generation of statistics.

References

No References available

Anonymization of Census data with sdcMicro package – the Bulgarian case

Authors

- Lyubomira Dimitrova (Bulgarian National Statistical Institute)

Abstract

The Bulgarian National Statistical Institute conducted the Population Census in 2021. Variables such as religion, nationality and ethnicity are considered sensitive and thus the dataset needs to go through an anonymization process. In this way the privacy of those respondents for which personal information can be extracted through the data would be protected. The used method is record swapping through the sdcMirco package in RStudio. The method has been applied for the whole dataset and the results and difference in distribution will be discussed in the following presentation.

References

No References available

Automated monthly procedure with R: the e-invoice case

Authors

- Bruno Lima (Statistics Portugal)
- João Poças (Statistics Portugal)
- Sofia Rodrigues (Statistics Portugal)

Abstract

Monthly, Statistics Portugal receives data from Tax Authority regarding a mandatory invoice declaration system. This e-invoice system is part of implemented anti-fraud measures, although the data is not always thorough. So that we can use this administrative data for statistical production, monthly we analyse data from more than 2000 issuers (the most relevant ones) looking for possible missing data. The urgency on the availability of this data requires a significant effort in its treatment. Likewise, we implemented an automated R procedure to identify and impute values on potential missing data. Main R packages used in the procedure: {targets} to define a reproducible workflow; {tidyverse} and {tsibble} for data manipulation; {ROracle} to connect to our data warehouse; {isotree} for the implementation of isolation forest on anomalies detection; {imputeTS} for imputation of missing data; and we also use an 'inhouse' package developed to aid data analysis. For each one of the more than 2000 relevant issuers, we first identify those that remain in business and had a reported value in the previous month. For each one of these issuers, we have the monthly taxable amount aggregated by issuer and acquirer and the number of registers reported according to acquirers' class. We distinguish two types of missing data: a total missing when no value is reported, by issuer; and a partial missing when reported value is way lower than would be expected. For the identification of partial missing, we use the isolation forest algorithm both for taxable amount and for number of records aggregated by issuer. Imputation of taxable amount is done with Kalman Smoothing in structural time series models at level issuer / acquirers' class. In the case of partial missing, we subtract the current lower value to Kalman Soothing imputations. This procedure allows to identify and correct obvious reporting errors and provides better quality administrative data to our users.

References

No References available

Automatic product classification for price statistics

Authors

- Bogdan Oancea (University of Bucharest and National Statistics Institute of Romania)
- Ana Tiru (National Statistics Institute of Romania)
- Iulia Toma (National Statistics Institute of Romania)
- Marian Necula (National Statistics Institute of Romania)

Abstract

One of the directions followed in the last years to modernize the official statistics production was to augment the classic data collection process for price statistics with data collected from the major e-commerce sites. Thus we developed a set of R scripts for web scrapping and collected a large number (~10^5) of product's records on a weekly basis. In order to use these records, products must be classified according to the categories used to compute the CPI. We piloted an automatic classification process, starting with a reduced set a product categories, choosing 15 categories from ECOICOP international classification, from food and appliances categories. Product names being text data, our process begun with building different embeddings, i.e. transforming text data into numeric vectors which are suited to be used by ML classification methods. We used the count vectorization, TF-IDF, Word2Vec, Fasttext and Glove methods to build these embeddings. For Word2Vec and Fasttext we built the embedding of a product name in two different ways: by adding the embeddings of each word from the product name and by averaging the embedding of each word. Having the numerical representation of product names, we proceeded with a series of ML techniques: logistic regression, Naive Bayes, Decision trees, SVM, Random forests, Neural networks. The results showed an impressive accuracy for LR and Naïve Bayes (0.98). Regarding the vectorization methods, the best results were obtained with count vectorization, TF-IDF and Fasttext for all classification methods used in our study. All data processing was performed using the R software system.

References

No References available

Calculating multilateral prices indices with parallel computing

Authors

- Tobias Brünner (Federal Statistical Office of Germany (Destatis))

Abstract

The Federal Statistical Office of Germany (Destatis) aims to incorporate scanner data –transaction data provided by supermarkets and other retail outlets– into the monthly calculation of the consumer price index and the harmonized index of consumer prices. To that end, Destatis built a Proof of Concept that, after some preprocessing, aggregates scanner data into elementary price indices. As the data to be processed is very large, computation is distributed across a Spark cluster using sparklyr as an interface. Elementary price indices are compiled using the IndexNumR package for multilateral methods. The R package arrow is used to speed up the data transfer between R and Apache Spark.

References

No References available

cchsflow: An open science approach to transform & combine population health datasets

Authors

- Kitty Chen (University of Ottawa, Ottawa Hospital Research Institute)
- Warsame Yusuf (Public Health Agency of Canada)
- Carol Bennett (Ottawa Hospital Research Institute, ICES)
- Yulric Sequeira (Ottawa Hospital Research Institute)
- Douglas G. Manuel (Ottawa Hospital Research Institute, University of Ottawa)

Abstract

The Canadian Community Health Survey (CCHS) is one of the world's robust ongoing cross-sectional population health surveys, with over 130,000 respondents every two years (currently over 1.5 million respondents since its inception in 2001). While the survey remains relatively consistent over the years, differences between cycles, including a major redesign in 2015, present a challenge when conducting longitudinal analyses. cchsflow is an R package that transforms and harmonizes variables into consistent formats across multiple CCHS cycles (currently, 2001 to 2018). By implementing open science practices, cchsflow aims to minimize the amount of time spent data cleaning for the wide number of CCHS users. Worksheets were used to guide the variable transformation process and to generate harmonized metadata. A GitHub repository was created for the package (Manuel et al. 2020) where the broader community of CCHS users collaborate to improve the consistent use of the surveys. Currently, cchsflow package is available for installation through the Comprehensive R Archive Network (CRAN) and contains support for over 300 CCHS variables for 1.5 million respondents of the CCHS Public Use Microdata File (PUMF) from 2001 to 2018. There are over 400 package downloads per month. The functions and approach from cchsflow now forms the basis of a new package, recodeflow (Sequeira, Bailey, and Vyuha 2021), which is used to transform and harmonize an increasing number of databases. The worksheets of cchsflow and recodeflow are comma separated values (CSV) tables can be used in any software systems.

References

- Manuel, Doug, Warsame Yusuf, Rostyslav Vyuha, and Carol Bennett. 2020. Cchsflow: Transforming and Harmonizing CCHS Variables. https://github.com/Big-Life-Lab/cchsflow; Sequeira, Yulric, Luke Bailey, and Rostyslav Vyuha. 2021. Recodeflow: Interface Functions for PMML Creation, and Data Recoding. https://CRAN.R-project.org/package=recodeflow.

Circular Letter Generation using R

Authors

- Bernhard Meindl (Statistics Austria)

Abstract

In this contribution we present and outline the ideas behind a project that may become the default way at Statistics Austria to produce and manage standardized circular letters using the existing R environment. The idea behind this new development is that any circular letter to be produced is considered as a bundle. A bundle consists of a LATEXtemplate, a number of assets (e.g. images) that are used within the template as well as a finite number of data fields. These fields (e.g addresses and names) are then replaced with individual information for each letter from a data source. In order to make the approach as flexible as possible, the data source containing the individual data entries can be either a flat text file or an R binary file which can be easily produced e.g from database queries. To facilitate the process of document generation, an (internal) R package (schriftverkehr) was developed which allows to generate a R6 object which allows to create, upload, start, stop and monitor jobs on on Rstudio Connect instance and finally download (individual) letters. Furthermore, the package allows to manage assets (upload, download, move) that can be used when defining jobs. The functionality of the package heavily depends on two different APIs (one for asset management and one for job management) that are deployed using Rstudio Connect. Any request to those APIs needs to be authenticated and we also show how these APIs were used to build an interactive shiny-based app to manage jobs and assets interactively. To make sure, the circular letter generation is reproducible, jobs can also be specified using a git-based repository and starting jobs can be triggered using hooks or using the scheduling options from Rstudio Connect. We discuss the reasons for this setup and identify strengths and possible weaknesses of this approach. Furthermore, we discuss how the setup could easily be extended to not only be able to produce circular letters as pdf-files but also - for example - emails.

References

No References available

Comparing Noise Mechanisms for Differential Privacy

Authors

- Mustafa Salamh (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)
- Ehssan Ghashim (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)

Abstract

We examined two privacy-preserving mechanisms for contingency table release. The two mechanisms achieve -differential privacy (DP) by adding to cell counts noise generated from a double-geometric (DG) distribution or a Laplace (LAP) distribution, respectively. The performance of the two mechanisms was investigated under a variety of conditions using simulated data. In a Monte Carlo experiment, we controlled for the sample size of the original data, level of correlation between attributes, and level of privacy ( ). We assessed DP contingency tables produced using the LAP and DG mechanisms by comparing Cramer's V statistics and estimates of the marginal distributions for each attribute to those of the original data. We also used the Jensen-Shannon divergence to measure the distributional similarity between the original contingency table and the DP ones. All of our results suggest the LAP and DG mechanisms are effectively equivalent for contingency table output and show that sample size is the dominating factor affecting performance. All three measures of performance (Cramer's V comparison, bias in estimating the marginal distributions, and the Jensen-Shannon divergence) improve significantly with larger sample size (above 10,000). From that we developed an R script which can be used to apply the two mechanisms on tables stored in a local R environment or on a SQL server.

References

No References available

compindexR: An R package for calculating composite indicators

Authors

- Michał Bernard Pietrzak (Gdańsk University of Technology, Department of Statistics and Econometrics)
- Marta Kuc-Czarnecka (Gdańsk University of Technology, Department of Statistics and Econometrics)
- Olgun Aydin (Gdańsk University of Technology, Department of Statistics and Econometrics)

Abstract

Composite indicators are widely used for evaluating and ranking of economic objects, including countries and institutions in terms non-measurable, multidimensional economic phenomena. It is common practice to use composite indicators for a variety of tasks, such as monitoring policies and creating recommendations, providing easily understandable information, and establishing rankings. They are straightforward, offering a snapshot of multidimensional phenomena in a way that makes evaluation and comparison of economic objects much easier. Although composite indicators are an important and easy way of interpreting complex of economic phenomena in one dimension, it requires expertise in terms of calculating them. There are some softwares which help researchers calculate composite indicators. However, none of them offer possibilities such as automatic variable selection and end to end calculation. This is a road block for domain experts who are willing to calculate composite indicators. They are very well equipped when it comes to economic knowledge of the analyzed economic phenomena but don't have enough knowledge about calculating indicators step by step or they don't have experience with any type of softwares or programming language. This paper covers an R package called compindexR. The package aims to provide an interface for researchers to calculate composite indicators without requiring additional knowledge regarding calculation, programming and taxonomometry. Authors of the package are about to release the first version of the package on CRAN and they are also willing to build an user friendly Shiny application on the top of the package for the future studies.

References

No References available

Cross-data sources analysis with Dynamic reports using {flexdashboard} and {crosstalk}

Authors

- Alexandre Cunha (Statistics Portugal)
- João Poças (Statistics Portugal)
- Sofia Rodrigues (Statistics Portugal)

Abstract

At Statistics Portugal, a large volume of data, regarding business activities, is collected monthly both from surveys and administrative sources. To compare data from those two sources, it is produced a Dashboard that summarize key financial metrics on a group of selected enterprises. The Dashboard was built in R and relies mainly on {flexdashboard} and {crosstalk} packages allowing for a dynamic and interactive interface easy to the user. R packages used to build the Dashboard: {readxl} and {fs} help in file control and getting company data into R; {ROracle} creates a fast connection to multiple Oracle data sources; An html file is generated using {flexdashboard}, creating an interactive dashboard; {crosstalk} enables filters to be applied to different components, keeping plots and tables within the same filter; {DT} is used for markdown tables and {plotly} fordynamic plots; {taskscheduler} creates a daily routine that keeps the report updated. This set of packages gives the users multiple exploratory analysis tools and the ability to direct its own analysis interactively. This approach allows developers to avoid shiny servers, keeping their data in an html file without losing the shiny features of dashboard customization and interaction. The output is an html file composed of three tabs: Main view: summary of the last available month for N companies; View by company: select a company and a period. Plots, tables and custom insights are dynamically changed by the given selection; View by month: select a time range and companies will be filtered for that period. This Dashboard improves communication across different statistical-domain teams without the need for any third-party servers and easy access to up-to-date data, allowing users to keep track of differences between sources, potential missing values, outlier detection or data trends.

References

No References available

Data visualization for machine learning practitioners

Authors

- Julia Silge (RStudio PBC, USA)

Abstract

Visual representations of data inform how machine learning practitioners think, understand, and decide. Before charts are ever used for outward communication about a ML system, they are used by the system designers and operators themselves as a tool to make better modeling choices. Practitioners use visualization, from very familiar statistical graphics to creative and less standard plots, at the points of most important human decisions when other ways to validate those decisions can be difficult. Visualization approaches are used to understand both the data that serves as input for machine learning and the models that practitioners create. In this talk, learn about the process of building a ML model in the real world, how and when practitioners use visualization to make more effective choices, and considerations for ML visualization tooling.

References

No References available

Data-driven data validation: the R-package 'validatesuggest'

Authors

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

Abstract

Data validation is a cornerstone in data intense industries, such as the art of making official statistics. To create and maintain high quality statistical output, data needs to be checked before being used in statistical processes. A number of projects have been executed to streamline and optimize data validation processes, both within organisations as well as among organisations. An important success factor for effective data validation is the design and maintenance of validation rules that cover the dynamics of the data to be checked. This has led to the definition of standardised validation rules that cover the most common use-cases in official statistics. Examples are the definition of internationally agreed 'main types of validation rules' by Eurostat [1], and the set of recipes and standard functions as offered in the well-known R-package 'validate'[2], which are documented in the online cookbook [3]. In this presentation we take another approach to rule maintenance. In addition to the knowledge of the domain specialist we let the data speak. Properties of the data, such as type, range, distribution, correlation can be used to derive rules that catch the essentials of the data. Since the number of rules that could potentially be derived from data in general could be endless, we use the existing international and national standardised validation rule systems to know what type of rules make sense. A refinement of the concept is to also take the time dimension of time series data into consideration. That way time-dependent validation rules com into reach. The suggested rules are expressed in a human-readable form, so that the domain specialist / rule maintainer can inspect and understand them, as the data-driven concept is intended as a suggestion to the rule maintainers. They should always be checked and interpreted before putting into production. The type of rules currently implemented in the experimental R-package 'validatesuggest' [4] are: Positivity checks; Range checks; Checks on whether a variable may contain NAs; Checks on uniqueness of a variable; Type checks; Ratio checks; Discovery of conditional rules. Rules are expressed in the R-validate syntax so that they can be applied directly in an R context. A high level function 'suggest_all' is available to infer all supported validation rules in one pass. The ratio check and discovery of conditional rules are not straight forward. The ratio check uses a correlation threshold: only variables that are (enough) correlated are considered. The discovery of conditional rules implements a unsupervised machine learning technique (association rules), that checks the co-occurrence frequency of values. e.g. if the values "job: retired", "income_type: pension" co-occur, the rule "if (income_type == pension) job == retired" will be derived. The direction of the causality relation is derived from the occurrence of other values for the respective variable. In this example for "job: retired" there is only one income_type, where for "income_type: pension" there are multiple job values. If no direction can be derived two rules are generated. Up to now the result of the discovery of conditional rules are if then else statements. More complex conditions can be expressed in decision trees (Classification and Regression Trees or CART). Such constructs have the big advantages over other ML techniques that they are still human interpretable for a rule specialist, which is a precondition for this work, since rules have to be interpreted by rule maintainers. The implementation of the suggestion of decision trees from data is experimental and ongoing work. As an example of the current functionality of the validatesuggest package Figure 1 shows the set of suggested rules from the (fictitious) retailers dataset available in R-validate. In this presentation we explain about the data-driven data validation concept, the implementation in 'validatesuggest' and we present our ideas for the use of this approach and the software in practice. We look forward to possibilities to extend the concept.

References

- [1] V. Tronet (2018), Main types of validation rules for ESS data (version 1.0.3). Eurostat Working document.; [2] Mark P. J. van der Loo, Edwin de Jonge (2021). Data Validation Infrastructure for R. Journal of Statistical Software, 97(10), 1-31. doi:10.18637/jss.v097.i10, validate: https://cran.r-project.org/package=validate; [3] M. van der Loo, O. ten Bosch, The Data Validation Cookbook: http:

//data-cleaning.github.io/validate/; [4] E. de Jonge, O. ten Bosch (2022), R-package validatesuggest: https://github.com/data-cleaning/validatesuggest

Demonstration of an Exploratory Method for Categorical Data

Imputing Inventories Zero or Non-zero Values

Authors

- Anri Mutoh (Rissho University, Tokyo, Japan)
- Ichiro Murata (National Statistics Center, Tokyo, Japan)

Abstract

In the course of the data processing of the Unincorporated Enterprise Survey, National Statistics Center has been developing imputation methods for missing values in respective enterprise records, namely, the amounts of Sales, Purchases, Salaries, etc. Among others, the amount of Inventories is found to be zero in many cases, which could undermine the accuracy of imputation. While it should be noted that such enterprises in the service industry may not necessarily keep stocks, in those cases prediction values are small but non-zero, then impairs the accuracy. There are some methods for data with many zero values including a tobit model or a zero-inflated Poisson model in predicting the amount of Inventories as an objective variable. Taking categorical attributes of respective enterprises such as Industrial Class, Sales Range Group, etc. as explanatory variables, we predict whether Inventories are zero or non-zero, followed by predicting the missing values themselves. To select categorical variables related to Inventories zero or non-zero, we adopted the summary statistics S1*() proposed as a method of exploratory data analysis because of easy calculation, stability, and interpretable modelling. We also used well-known variable selection methods for the regression model for categorical variables such as Lasso regression or the AIC stepwise method. We found that the latter two methods are less interpretable compared to S1 method because the results are obtained at each level of the categorical variables and the selection of variables is unstable with cross-validation. Further, we conducted simulations about predicting Inventories zero or non-zero by the models made by each method. The result showed that the accuracy of S1 model was as good as that of other methods which is originally intended for accuracy. Therefore, the result partially demonstrates the advantage of S1. (* The paper in which the author proposed the summary statistics S1 is currently under peer review. The review is scheduled to finish by this November. Its detail will be cited in the presentation slides and discussed.)

References

No References available

Eurostatistics - from PDF format to interactive web visualisation using R

Authors

- Rosa Ruggeri-Cannata (Eurostat)
- Piotr Ronkowski (Eurostat)
- Anette Sundstroem (Eurostat)
- Johannes Buck (Eurostat)

Abstract

Making statistics accessible and meaningful for the public is an important task for statistical institutions around the globe. One of the best techniques for communicating data is to visualize the numbers in a graph. Interactive visualization tools are becoming more and more common. Data storytelling based on good visualization and narrative is of great interest for the user. In order to meet these expectations, we have transformed "Eurostatistics", a monthly publication of Eurostat traditionally published as a PDF, to an interactive web visualisation using R. This new tool is primarily based on the Flexdashboard package of R, which allowed us to quickly set up a flexible content structure, and on the Plotly package, which made us able to create interactive graphs. The storyboard layout of Flexdashboard presents the data visualizations in a sequence together with a written commentary. It is a user-friendly way of communicating the main messages. The tool includes links to the data sources where the user has the possibility to further explore the data. A number of functionalities, such as downloading graphs, complement the user-friendliness of this R tool. With some minor additional adjustments, it was possible to make the layout adapt to mobile devices. In the production process, the tool first loads data, then rapidly generates a set of graphs based on those data and inserts texts from external files. The production process is much faster than the one for a traditional publication, and is therefore particularly suitable for products that are frequently updated, for example monthly. A set of chart functions created in Plotly makes it easy to add new graphs to the tool. Our project has proved that R with its packages can be adopted for statistical visualisation with limited resource investment. The production process, in full hands of statisticians, is much faster compared to traditional publishing. It is particularly suitable for products that are frequently updated, as in this case monthly.

References

No References available

GISSB – a package for network analysis on the Norwegian road network in R

Authors

- Sindre Mikael Haugen (Statistics Norway)

Abstract

Statistics Norway has created an R package called GISSB which contains the tools necessary to conduct network analysis on the Norwegian road network. This is an open source alternative to more traditional GIS software such as ArcGIS. The functions in the package enables the user to find the shortest path between any two addresses in Norway by supplying their street addresses and postal numbers. This can in turn be used to calculate the shortest driving time in minutes, or the shortest driving distance in meters, from inhabited addresses to public services such as hospitals. Furthermore, these results can be aggregated by different regional levels (such as counties or municipalities) to calculate the median or mean driving time/distance for the area in question. In addition to be very effective at computing driving routes for large numbers of addresses, the GISSB package makes network analysis more accessible to both employees at Statistics Norway and the general public alike, as no previous experience with GIS software is required. The required road network files are also publicly available from the Norwegian Mapping Authority. The GISSB package has been used to create official statistics which will be published soon. The analysis focuses on the driving routes in minutes and meters from every inhabited address in Norway to their nearest maternity ward and/or birth center. The population of addresses has been limited to addresses where fertile women reside. The results from this analysis shows that there are large regional differences in Norway regarding the population's accessibility to maternity wards and birth centers.

References

No References available

hEDA package

Authors

- Mervyn O'Luing (Central Statistics Office, Ireland)

Abstract

In this tutorial, we will download and install the hEDA package into the R environment on our computers. We will then read in and pre-process some data, and from these data create a data frame of the key variables. This will be our sampling frame. From this, we will identify our target and auxiliary variables. Then we will build our atomic strata and obtain summary statistical information for the target variables in each atomic stratum. We will inspect this output and then proceed to set our precision constraints for each target variable and domain. A solution represents a stratification of the atomic strata. The quality of each solution is measured by its sample cost. We are now ready to generate some initial solutions for the hEDA algorithm. Today we will generate solutions using the following methods: kmeans, pam, expectation maximisation and fuzzy clustering. We use the sample size of each solution to determine which solution to use as a starting point for the hEDA. However, in doing so we will also check the number of strata in each solution, which could be useful in determining what balance between exploration and exploitation we need to strike with our hEDA. Once we have our starting solution, we need to set the hyperparameters for the hEDA. Usually, it is advisable to fine tune these hyperparameters and today we will use sequential model-based optimisation to do so. Once we have fine-tuned the hyperparameters, we are now ready to use these with the hEDA and compare the resulting solution quality with that of the initial solution.

References

No References available

Histogram and Copula Synthetic Data Generation Methods in R

Authors

- Ehssan Ghashim (Canada Revenue Agency)

Abstract

We are presenting four differentially private (DP) methods developed from the broader "histogram" and "copula" families of synthetic data generation methods. We implemented the histogram method from Dwork et al. (2006) and the Enhanced Fourier Perturbation Algorithm (EFPA) from Gergely et al. (2012). Both aforementioned histogram methods transform continuous variables into categorical variables by constructing bins from the original data. Differential privacy is achieved by perturbing the frequencies of the constructed bins, either by adding Laplace noise (histogram method) or by applying the Fourier sanitization technique (EFPA method). Synthetic data points are then sampled from within the ranges of the constructed bins. To generate a synthetic dataset of the desired size, records are sampled from a weighted grid based on the perturbed bin counts. In search of more computationally efficient synthetic data generation methods, we implemented two differentially private (DP) copula methods in R. The DP copula framework samples synthetic data directly from the marginal distributions and the Gaussian copula function of the original data (Li et al. 2014). Because it allows for direct sampling, copula methods can handle high dimensional data more efficiently than histogram methods. In one of our approaches, we partitioned the dataset into groups by joint categorical levels and added Laplace noise to the group counts. We then applied the DP copula framework to the numeric variables within each category group. In the second implementation, rather than splitting the data by categorical levels, we handled both categorical and numerical features together by converting the original data into binary data (Asghar et al. 2019). This approach is more efficient and avoids some of the computational complexities of our first implementation, however, the generated synthetic data poorly preserves the relationships between variables.

References

- C. Dwork, F. McSherry, K. Nissim, and A. Smith. (2006) Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference '06, pages 265–284.; G. Acs, C. Castelluccia and R. Chen, "Differentially Private Histogram Publishing through Lossy Compression," 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 1-10, doi: 10.1109/ICDM.2012.80; Li H, Xiong L, Jiang X. (2014) Differentially Private Synthesization of Multi-Dimensional Data using Copula Functions. Adv Database Technol. doi:10.5441/002/edbt.2014.43 ; Asghar, H. J., Ding, M., Rakotoarivelo, T., Mrabet, S., & Kaafar, M. A. (2019). Differentially Private Release of High-Dimensional Datasets using the Gaussian Copula. arXiv preprint arXiv:1902.01499.

Impact of Main Skills Targeted by CVT Courses on Economic Activities of Enterprises

Authors

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

Abstract

In this paper we analyse the impact of Continuing Vocational Training (CVT) courses on turnover, value-added and gross operating surplus, labour productivity (apparent and wage adjusted) and other special aggregates (dimensional and qualitative) of enterprises' economic activity. The analysis is developed (for 28 European countries, during the 2010-2020 period) starting with the main type of skills targeted by CVT courses (general and professional IT skills, management and office administration skills, team working, customer handling, problem-solving and numeracy skills, foreign language and communication skills, technical, practical or job-specific and other skills and competences). As a methodology, we use R software to solve econometric models with panel data (time series and longitudinal data).

References

No References available

Improvement of Model Construction based on Reliability Scores of Objects for Autocoding

Authors

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

Abstract

This paper presents a new method for the improvement of a model construction for autocoding. In particular, we improve the training dataset based on reliability scores over objects for autocoding. We have developed a classification method based on the reliability scores, which have been defined considering both probability measure and fuzzy measure, for autocoding of data related to the family income and consumption survey. The developed classification method has introduced a method for the improvement of the training dataset based on the pattern of reliability scores over objects by using a clustering method. It has been practically implemented for the coding task of the Family Income and Expenditure Survey. However, a new method is a trial to simplify the previous procedure, including the clustering procedure based on the distribution of the reliability scores over the objects. That is, in this proposed method, the clustering procedure is excluded, and only the highest score of reliability scores for each object are considered. First, the proposed method performs the code assignment to evaluate data with the original training dataset. Then, we extract data whose reliability scores are relatively high evaluated data. After that, it adds the information of the extracted data to the original training dataset. Then, it performs the code re-assignment with the improved training dataset to the evaluation data. As it is assumed that data classified with high-reliability scores are clearer data, adding the information of those data to the training dataset is performed to obtain a better classification accuracy. In addition, avoiding the consideration of the distribution of reliability scores over the objects reduce the computational complexity related to the calculation cost. The numerical examples show a better performance of the proposed method.

References

No References available

Partially Differentially Private Synthetic Data Generation Using PCA – Mustafa Salamh

Authors

- Mustafa Salamh (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)

Abstract

In this study, we used principal components analysis (PCA) to generate -partially differential private synthetic data. The method is developed to preserve the correlation structure as observed in the original data set. The method can handle data containing both numerical and categorical variables. It utilizes the spectral decomposition theorem to map the original data set into a new orthogonal product space (OPS). Random samples are selected independently across all dimensions of the OPS and the synthetic data is finally obtained through the inverse map. A wide simulation study was conducted to investigate the performance of this method across several types of model-based artificial data sets. The goal of this simulation is to evaluate the performance of the PCA method with respect to model architecture, dataset composition, privacy-budget and data size. Based on our existing IT policies and security constraints, we developed tools in-house and created our own R-package to generate synthetic data. We also highlight the use of an R wrapper that allows us to leverage SQL code for generating synthetic data on Netezza servers.

References

No References available

Population forecasting

with Bayesian hierarchical models

Authors

- Violeta Calian (Statistics Iceland)
- Ómar Harðarson (Statistics Iceland)

Abstract

The goal of population projections is to predict the future values, and their uncertainty measures, of regional/total population by age, gender, time and other demographic or spatial characteristics, based on past observed values over a reasonably long time. The projection method should satisfy the following conditions: to be based on statistical modeling and be able to efficiently solve small area/population and rare events issues as well as to account for complex (auto-) correlation structures and to incorporate qualitative and quantitative prior information and/or expert assumptions. Currently there are three main classes of methods for producing (total) population projections: based on assumptions/scenarios, on functional models and based on Bayesian models. Statistics Iceland employed in the past a mixture of such methods, by forecasting fertility and mortality with functional models while modeling short term migration with econometric/ARDL models. In addition, modeling the time correlation between emigration and lagged immigration has further improved the predictions. In this paper we describe our new approach to population forecast which can successfully fulfill the requirements mentioned above. It is based on Bayesian hierarchical models and it is implemented in an open source R code. We provide two classes of solutions, depending on data: individual response models, if input microdata is available and aggregated/rates response models if only count data is available. Time and age (auto-)correlations are incorporated via Gaussian processes or flexible non-linear smooth terms while spatial and social-demographic characteristics are included in a natural multilevel model setting. For instance, fertility and mortality rates have a non-significant variation by municipality but depend on family related and education characteristics. Even when based on count data only, they provide very stable forecast models by age, time and space. Migration rates, by contrast, are highly fluctuating and sensitive to a richer set of factors which we address in some detail.

References

No References available

Preparing For A Journey: Introduction to Application Security in R

Authors

- Keith Douglas (Statistics Canada, Cyber Security Division)

Abstract

Production of official statistics using computational solutions involves secure design, implementation and maintenance of systems and data models. In this beginner's tutorial we introduce the 3 pillars of cybersecurity and an additional value many statistical agencies adopt (reputation). We discuss how they can play a role in application security for "small systems" and look at source code level examples designed for an audience familiar with R and typical libraries ("packages") from CRAN. The notion of the security boundary is introduced and related to the notion of the function in the functional programming sense. Participants will learn to recognize security sensitive parts of applications and practice some tips and techniques for improving the security stance involved. Completing this tutorial will hopefully motivate a journey towards more effective software development practices and towards better application security.

References

No References available

Query Score for Assessing Utility of Synthetic Datasets

Authors

- Alison Wardlaw (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)
- Philippe Bélanger (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)
- Ehssan Ghashim (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)
- Aditya Maheshwari (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)
- Rachel Ostic (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)
- Mustafa Salamh (Innovation and Analytics Division/Advanced Analytics and Artificial Intelligence Section/Canada Revenue Agency)

Abstract

Decision-makers and analysts are often interested in looking at summary statistics for certain subsets of the population. Replicating properties within these subsets is a consideration for the utility of synthetic data. In R, we implement a metric originally developed in SQL (Xiao et al. 2010) that assesses the robustness of our synthetic datasets to filtering on multiple features. To create a filter, we randomly choose upper and lower cut limits for each numeric variable and a level for each categorical feature. We apply the filter to both the original and synthetic datasets and compare results. The simulation code repeats this process for 1000 randomly generated filters and averages the difference between statistics calculated on the original and synthetic subset data. Large average differences are indicative of a synthetic dataset that has the potential to lead analysts to erroneous conclusions on filtered samples. Currently, we measure the difference between the number of records remaining after filtering, although the method is flexible and could be used to verify that a feature mean or other moment is preserved upon subsetting.

References

- X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. In ICDE, pages 225–236, 2010.

R and JDemetra+ 3.0: a new toolbox around seasonal adjustment and time series analysis

Authors

- Alain Quartier-la-Tente (INSEE France)

Abstract

In 2019, the first R interface to the seasonal adjustment software JDemetra+, RJDemetra, was published. It implements the two leading seasonal adjustment methods TRAMO/SEATS+ and X-12ARIMA and allows to manipulate workspaces used in production. Three years later, with the development of JDemetra+ 3.0, more than 12 new R packages are being developed and not only on seasonal adjustment: several tests on time-series data and presence of seasonal and trading days effects; creation of outlier regressors and trading-days variables (with user-defined calendars); manipulation of moving averages and trend-cycle extraction methods; seasonal adjustment in a faster way than in RJDemetra and with more functionalities; benchmarking and temporal disaggregation; seasonal adjustment of high frequency data; structural time series and state space models. This presentation aims at presenting the different packages, their performance and their new functionalities with simple examples.

References

No References available

R in Labour Statistics

Authors

- Tatyana Savchenko (Employment statistics division of the Labour Statistics Department, National Statistical Committee of the Republic of Belarus)

Abstract

The National Statistical Committee of Belarus uses R as the convenient and reliable tool for statistical production. R provides several possibilities for processing, analysis and visualization of results of state statistical observations. The purpose of the presentation is to share experiences in the use of R in labour statistics. Integrated statistics system of the Republic of Belarus (ISSS) is widely used in Belarus' statistics, providing automation of all processes of organization and maintenance of state statistics. However, additional software tools are necessary for some indicators that require non-standard and more complex calculations. The presentation will give an overview of the R applications used in the production of labour statistics.

References

No References available

R software for Statistical Disclosure Control – utility and problems

Authors

- Andrzej Młodak (Statistical Office in Poznań, Poland)
- Tomasz Józefowski (Statistical Office in Poznań, Poland)
- Tomasz Klimanek (Statistical Office in Poznań, Poland)

Abstract

Statistical Disclosure Control (SDC) is essential for modern data dissemination policy. The main aim of SDC is to find a trade-off between minimizing the risk of disclosure risk – i.e. the risk that an individual unit or its sensitive attribute will be identified on the basis of the avail-able data – and minimizing information loss resulting from applying SDC (i.e. maximizing the utility of disclosed data for the user), such as suppression or perturbation of original data. Among the most important software solutions that implement SDC techniques are packages of the R environment (such as e.g. sdcMicro, sdcTable, recordSwapping or cellKey), which are constantly developed. In the paper we will present our previous experience of using these tools for the protection of Polish statistics. The main advantages and drawbacks of R packag-es dedicated to SDC will be discussed. Their utility will be assessed taking into account their performance (especially their ability to handle larger databases, such as a census), availability of various solutions, consistency and continuity of operation (e.g. whether all operations are performed on the same type of object throughout the whole SDC process, which is necessary to be able to observe changes in disclosure risk due to consecutive suppression or perturbation steps), in-built quality measures and convenience of output results, for statisticians and data users. We present the newest functions implemented in these packages and the most frequent problems and challenges encountered when using them, making suggestions as to possible improvements.

References

No References available

R you working in the open? The importance of open source culture in adopting R in Official Statistics

Authors

- Kate Burnett-Isaacs (Innovation Program Manager, Statistics Canada)

Abstract

Many National Statistical Organizations (NSOs) are adopting R to improve data processing, advance data analytics and integrate data science into the production of official statistics. R is a powerful programming language that can help NSOs reap efficiencies and enhance products and production in new ways beyond traditionally used software. However, statistical agencies will leave efficiencies on the table if R adoption is not paired with the adoption of the open way of working. The open way of working involves instilling principles of transparency, collaboration, Agile development, inclusivity and community into the statistical and analytical work conducted by employees within a NSO. Transparency in official statistics means that employees can view and reuse others' work to learn from each other and reduce duplication. Collaboration means everyone can contribute to building the best solutions possible. The Agile approach of prototyping and releasing minimum viable products early and often allows for new discoveries and faster solutions. Inclusivity means that diversity is actively encouraged; everyone is heard and the best ideas are celebrated and support by the group. Community packages these concepts together and results in a common, shared, goal and culture within and across a statistical agency. These cultural pillars must go hand in hand with the adoption of R for a NSO to fully realize the benefits of these modern tools. This presentation will provide an overview on how R, alongside the open way of working, can provide efficiencies and benefits to NSOs and share concrete examples and use cases of how Statistics Canada has approached this cultural transformation.

References

No References available

R-shiny applications for traffic and economic indicators

Authors

- Ala'a Al-Habashna (Data Exploration and Integration Lab (DEIL), Centre for Special Business Projects, Statistics Canada)
- Nick Newstead (Data Exploration and Integration Lab (DEIL), Centre for Special Business Projects, Statistics Canada)
- Dennis Huynh (Data Exploration and Integration Lab (DEIL), Centre for Special Business Projects, Statistics Canada)

Abstract

Over the last few years, the Data Exploration and Integration Lab (DEIL) at Statistics Canada has conducted an increasing amount of work with open data using various open-source tools in an open ecosystem. This exploration and experience have given rise to the idea of using an "open project" approach. Furthermore, several products and visualization tools have been developed and released by DEIL in the past few years, such as the Linkable Open Data Environment (LODE) viewer (Statistics Canada , 2021), the Traffic Flow Dashboard (Statistics Canada , 2022), and the Real-time Local Business Conditions Index (RT-LBCI) dashboard (Statistics Canada, 2021). We propose to make a presentation about two R-shiny applications developed at the DEIL, Statistics Canada. The first dashboard is the Traffic Flow dashboard. This dashboard shows traffic count data that is obtained from traffic camera imagery using a computer vision-based system developed at the Data Exploration and Integration Lab (DEIL), CSBP, Statistics Canada. The system periodically pulls traffic imagery from the Application Programmable Interfaces (APIs) of municipal and provincial traffic camera programs. Vehicle detection was implemented using the pretrained and open-source You Only Look Once version 3 (YOLOv3) object detection model. The output of the model is used to generate real-time counts of the detected vehicles (cars, trucks, buses, and motorcycles), and the dashboard shows the total aggregate counts of the different vehicle types. The second dashboard is the RT-LBCI dashboard. The RT-LBCI is intended to provide a real-time signal on business activities at the local level following the disruptions brought about by the pandemic and through the recovery phase. The RT-LBCI utilizes near real-time and highly granular sources of data such as Google Places to track business closures and TomTom Traffic Stats to track traffic near dense retail areas to quantify the impact of these disruptions. It is acknowledged that this signal does not capture all dimensions of business conditions; however, the methodological developments of this index are ongoing. New data sources may be incorporated into the index computation over time and, consequently, the computation methods may too be revised over time.

References

- Canada, Statistics. (2021). Linkable Open Data Environment Viewer. Retrieved from https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2020014-eng.htm; Statistics Canada . (2022). Traffic Flow Dashboard. Retrieved from https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2022018-eng.htm; Statistics Canada. (2021). Real-time Local Business Conditions Index. Retrieved from https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2021017-eng.htm

R, machine learning, and innovation in the production of official statistics: some use-case thoughts

Authors

- Sandra Barragán (Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Spain)
- David Salgado (Dept. Statistics and Operations Research, Complutense University of Madrid, Spain)

Abstract

We argue that R is highly suitable and fitted for innovation in the production of official statistics. This is especially evident in the incorporation of machine learning techniques into the suite of official statistical methods. We provide three illustrative examples, namely (i) the improvement of cost efficiency in data editing, (ii) the enhancement of timeliness by producing early estimates with a mass imputation exercise, and (iii) the time disaggregation of design-based estimates by computing sampling weights for shorter reference time periods (from quarterly to monthly). All these use cases have been developed entirely in R. We present results with real data with an emphasis on the appropriateness of R for the innovation of statistical methodology in a statistical office. R allows subject matter experts, methodologists, and computer scientists to collaboratively produce trustworthy statistical prototyping software tools in a very fast fashion. This brings minimum viable products into life in a very dynamic way, which can be further developed and deployed to production. We claim and illustrate with examples that R excels at shortening and accelerating the path from methodological ideas to prototypes with real data paving the way for a rigorous combination of finite-population methodologies and machine learning. These ideas connect directly with initiatives in the European Statistical System with the creation of an internal working group for Open Source Software (OSS) and the impressive list of OSS tools classified according to the GSBPM (https://github.com/SNStatComp/awesome-official-statistics-software).

References

No References available

Recent and traditional approaches to outlier detection in panel survey data

Authors

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

Abstract

Outlier detection is part of data editing phase of surveys that observe numeric variables (typically enterprise and agriculture surveys). In the UNECE's glossary1 an outlier "is a data value that lies in the tail of the statistical distribution of a set of data values" and the underlying assumption is that "outliers in the distribution of uncorrected (raw) data are more likely to be incorrect". An outlier is often an influential value that has a great impact on the final estimate and therefore may deserve ad-hoc treatment. This work investigates outlier detection in longitudinal data when a set of quantitative non-negative variables are observed twice (or more) on the same sample of units. In the univariate case a popular technique in official statistics was suggested by Hidiroglou and Berthelot (1986); it identifies as outliers the units falling in the tails of the empirical distribution of the scores obtained by transforming the rations of the values observed in subsequent time occasions. D'Orazio (2022a) used the Hidiroglou and Berthelot scores as input of recent outlier detection methods developed in the field of statistical learning, in particular k-NN distance learning methods and isolation forest. This is rather straightforward in the R environment where the mentioned methods are already implemented in various additional packages: Hidiroglou and Berthelot outlier detection is implemented in the univOutl package (D'Orazio, 2022b); some distance-based outlier detection methods are implemented in the package DDoutlier (Madsen, 2018) (k-NN distance is calculated in many other R packages); the standard isolation forest is implemented in the package solitude (Srikanth, 2021) while the package isotree (Cortes, 2022) implements the "base" isolation forest algorithm as well as some of its variants. The empirical comparison in D'Orazio (2022a) shows that the "new" methods are rather promising; in particular, the isolation forest is very efficient and can also handle multi-modal distributions. This work goes a step further and investigates the behavior of these recent methods in the multivariate case; for this purpose, the comparison includes also well-known model-based outlier detection methods that assume a multivariate Gaussian distribution and the fact that it may be "contaminated", i.e. the data are the outcome of a mixture of Gaussian distributions. The first results of an empirical comparison using data of different nature (enterprise, agriculture holdings, households) show that model-based outlier detection procedures perform quite well but the isolation forest represents a valid alternative that has the additional advantage of being not dependent on a specific underlying statistical model. In addition, the isolation forest seems less sensitive to the specification of the input tuning parameters if compared to the distance-based outlier detection methods.

References

- Cortes, D. (2022) "isotree: Isolation-Based Outlier Detection". R package version 0.5.15, https://CRAN.R-project.org/package=isotree; D'Orazio, M. (2022a) "A note on outliers detection in longitudinal data". Submitted for publication; D'Orazio, M. (2022b) "univOutl: Detection of Univariate Outliers". R package version 0.3. https://CRAN.R-project.org/package=univOutl; Hidiroglou, M.A. and J.-M. Berthelot (1986) "Statistical editing and Imputation for Periodic Business Surveys". Survey Methodology, 12, pp. 73-83

Remote sensing for official statistics: ranking Romanian major cities in terms of an urban green index

Authors

- Marian Necula (National Statistics Institute of Romania)
- Bogdan Oancea (University of Bucharest and National Statistics Institute of Romania Iulia Toma, National Statistics Institute of Romania)
- Iulia Toma (National Statistics Institute of Romania)
- Ana-Maria Țiru (National Statistics Institute of Romania)

Abstract

Modernization process in official statistics promotes using new data sources in order to deliver new meaningful, timely and cost efficient statistics. Remote sensing data is a fairly new open data source, which is considered by some a low-hanging fruit for official statistics, and apparently satisfies the aforementioned constraints. This paper documents the results of deriving an urban green (vegetation) index for 41 major Romanian cities. As an intermediate result of SDG 11.7 objective, an urban green index could enhance the timely survey of urban infrastructure. Using Terra MODIS and Copernicus Sentinel 2, we estimated area and proportion of urban green vegetation proxied through NDVI (normalized difference vegetation index). The output was aggregated into a national ranking of urban greenness. Data access and processing were done using different R packages.

References

No References available

Seasonal Adjustment of Infra-monthly Time Series with JDemetra+

Authors

- Anna Smyk (INSEE France)

Abstract

Infra-monthly economic time series have become increasingly popular in official statistics in recent years. This evolution has been largely fostered by official statistics' digital transformation during the last decade, and the COVID-19 pandemic outbreak in 2020 has added fuel to the fire as many data users immediately asked for timely weekly and even daily indicators of economic developments. Many of those indicators display seasonal behavior and, thus, are in need for seasonal adjustment. JDemetra+, the official software for seasonal adjustment of monthly and quarterly data in the European Statistical System and the European System of Central Banks, has been augmented recently with a regARIMA-esque pretreatment model and extended versions of the ARIMA model-based, STL and X-11 seasonal adjustment approaches that are tailored to infra-monthly data and accessible through the {rjd3highfreq} R package. We give a comprehensive overview of the package's current developmental stage and illustrate selected capabilities,including code snippets, using daily births in France.

References

No References available

Small Area Estimation for Sampling Rare Populations

Authors

- Hisham Galal (UNHCR Regional Bureau for the Americas in Panama)

Abstract

Surveying rare populations is often hampered by the absence of representative sampling frames. Yet if the promise to "Leave No One Behind" is to be fulfilled in monitoring progress towards the achievement of the Sustainable Development Goals, then rare populations need to be brought out of the statistical shadows. Here we explore the use of small area estimation methods to produce area frames for rare populations. An application using a zero-inflated hierarchical Bayesian model fit with brms is presented for the construction of a sampling frame for the Venezuelan refugee-like population in Chile. We conclude with a discussion on the promises and perils of the proposed technique.

References

No References available

Spatial smoothing with the btb R package

Authors

- Kim Antunez (INSEE France)
- Julien Pramil (INSEE France)

Abstract

Spatial smoothing is a key method for analyzing spatial organization of data available at a small geographic level. Its aim is to provide simplified, clear mapping, relieved of the arbitrariness of territorial boundary lines ("Modifiable Area Units Problem" effect). From a theoretical point of view, spatial smoothing is a non-parametric estimation method for the intensity function of a point process with observed values in R^2. The theoretical intensity function in one point x is found by calculating the average points observed per unit surface on neighbourhoods containing x [Insee, 2018]. From a practical point of view, spatial smoothing relies on the choice of parameters:- The kernel describes how the neighborhood is approached. - The bandwidth quantifies the "size" of this neighborhood (to be chosen according to a bias/variance trade-off) - The geographical level from which the smoothed values are estimated (points, squares, etc.) - The treatment of edge effects makes explicit how geographical boundaries and the limits of observation territory are taken into account in the analysis. Several R packages make it possible to perform smoothing. For example, the spatstat package dedicated to the analysis of spatial point processes is very complete. It includes a smoothing function (density.ppp) and various functions for choosing optimal bandwidths (bw.diggle, bw.frac…). The btb ("beyond the border") R package, developed in 2018 by the French National Institute of Statistics and Economic Studies (Insee), implements a quadratic kernel estimation method with a variable bandwidth. The geographical level mobilized is a square whose size can be chosen. The main advantage of this package is that it takes into account border effects (maritime coasts for example). The method implemented is conservative (thanks to a normalization): before and after smoothing, the number of points observed is identical. The btb package also makes it possible to use quantile smoothing, which has the advantage of being less sensitive to extreme values and thus enriches the analysis of some variables, in particular income variables.In this presentation, we propose to present the main functionalities of btb with a concrete example of application: the housing prices in Paris in 2021.

References

- Insee (2018). Handbook of spatial analysis. Institut national de la statistique et des études économiques (chapter 8, spatial smoothing); - R package btb: https://github.com/InseeFr/btb

Statistical analysis of network data – a gentle introduction

Authors

- Eric Kolaczyk (Department of Mathematics & Statistics at McGill University, Canada)

Abstract

Among the many and varied data types available in the current digital age is a rich spectrum of network data, ranging from survey-based friendship networks to networks of human mobility. Our understanding of the potential impact of such data on official statistics, and how best they can be integrated with more traditional data, is arguably still in the early stages. In this talk, I will give a gentle introduction to some fundamentals of the statistical analysis of network data, largely through the lens of the R package igraph and several other complementary packages

References

No References available

The Cause of Inaccuracy for Wilson Confidence Intervals in R

Authors

- Mei Dang (University of Waterloo)

Abstract

For percentage estimates in custom tabulations, truncated and unadjusted Wilson confidence intervals, or CIs, can be constructed in R to varying degrees of accuracy. With CIs currently calculated by SAS macros, the migration between programming languages for flexibility and efficiency emphasizes the importance of finding accurate estimates in R. Comparing CIs and relevant variables used in its calculation, estimates are tested for both one-way and multi-way tables with all values rounded to 2 decimal places. In SAS, they are found using options in the tables statement of proc surveyfreq. Meanwhile, the process for R is broken into steps to calculate the same variables as SAS, performing calculations using formulas in SAS documentation. To evaluate base variables in R, the svymean() function from the survey package finds the percentage and standard error, and the fre() or cross_cases() functions find the total population size. Notably, the results from both programming languages are only truncated when the effective population size is greater than the total population size, so the standard error is only relevant when the results are truncated. After testing over 100 estimates, CIs and all relevant variables from R calculations match those from SAS for cases with no truncation. On the other hand, most CIs and the corresponding standard errors vary up to 1% for cases with truncation. Analysis reveal that the overall inaccuracy of CIs in R is due to that of the standard error, revealing a problem to further research.

References

No References available

The improvements offered by web scraping in R georeferencing packages

Authors

- Virgilio Pérez (University of Valencia)

Abstract

Georeferenced information is essential to get the most out of the data. Most of the information managed by companies and public administrations can be georeferenced, but the mechanisms currently in place do not allow the automation of geocoding processes, guaranteeing a minimum of quality. In this project, the functionalities of several R packages are discussed, which allow, with greater or lesser success, to convert a postal address into coordinates and vice versa. Since the error rate obtained in the different packages is high, an alternative based on web scraping is proposed.

References

No References available

The management of many users on different views of a single R Shiny application

Authors

- Yannick Wolters (Statistical Office of Lower Saxony, Germany)

Abstract

Using R Shiny applications is a great way of delivering data and information to recipients. However, when dealing with confidential data, it is necessary to control who can access these applications. And it is not just general access that has to be controlled, but also which part of the data or which evaluations are visible to whom. Furthermore, security aspects must always be considered when handling sensitive data, whereas handling many users requires consideration of scalability. All these aspects are not taken care of in a plain R Shiny project or with the setup of a vanilla Shiny Server. So we used ShinyProxy to obtain all of that. In the Statistical Office of Lower Saxony our group developed an application for all judicial authorities in Germany to deliver confidential data to verified and authorized users. Every user has to authenticate themselves using an open source Identity and Access Management system in which the underlying level of confidentiality is handled. These users can all access their data over the same Shiny application, but with different views depending on their level of authorization and confidentiality. Each session is separated from every other session using containerization via Docker, resulting in increased data security and higher scalability. By combining our application with ShinyProxy new possibilities opened up for us as a Statistical Office. We have now improved the data delivery to our customers without frequent IT effort while also minimizing maintenance and maximizing user-friendliness. It has officially been decided to handle statistics from all jurisdictions via our framework and we are already examining the embedding of further statistics.

References

No References available

The use of R and Shiny in the Belize 2022 Housing and Population Census: from ETL to advanced visualizations

Authors

- Gian Aguilar (The Statistical Institute of Belize, Belmopan, Belize)

Abstract

The Belize 2022 Population and Housing Census, conducted by the Statistical Institute of Belize, was the first digitalized Census in the history of the country. However, due to lacking national administrative data and incomplete telecommunications coverage, the census was primarily conducted via personal, house-to-house interviewing. Setting up the proper infrastructure for data collection, monitoring, supervision, quality assurance and early indicator analysis involved implementing, managing, and harmonizing a host of different software, platforms, and processes. In the middle of it all and helping to hold it all together, greasing the axles that turned the wheels of the Census, was the R programming language. R was used in almost all parts of the data collection process, from the creation and automation of the electronic questionnaire "assignments" to the monitoring, management and visualization of data collection using a complex Shiny dashboard. More specifically, R was used in: 1. The preparation of the original "sample" (based on the national building register) of Census units. 2. The Extract-Transform-Load (ETL) process of downloading collected census data and processing it and preparing it for use in databases, applications, and dashboards. 3. The automated, near real-time creation of dwelling-level census questionnaires based on the identified dwelling units during the first-phase "listing" of buildings and dwellings, via HTTP communication with the data-collection application. 4. The creation of a wide-scope, feature-rich Shiny dashboard for monitoring and managing many aspects of the Census data collection. Features encompassed: a. Census metrics, counts and key performance indicators. b. Interactive, feature-rich maps for monitoring coverage and progress, as well as for assisting in key aspects of data processing. c. Automated flagging and summaries of data inconsistencies. d. Automated payment processing for census enumerators, who were paid on a piece-wise basis. e. One-click creation of new census dwelling-level questionnaires based on uploaded building-level information. 5. The harmonization of census data and spatial data. 6. Census data post-processing, editing and preparation. The way that R was implemented in the Belize Census was through automated (scheduled) workhorse scripts and through the interactive Shiny dashboard. Its usage was one of the primary driving forces for Census success and there are many lessons available for (especially developing) countries that are preparing and planning for a Census.

References

No References available

Two-stage Sampling Design and Sample Selection with the R

package R2BEAT

Authors

- Giulio Barcaroli (Independent consultant)
- Andrea Fasulo (Italian National Institute of Statistics (ISTAT))
- Alessio Guandalini (Italian National Institute of Statistics (ISTAT))
- Marco D. Terribili (Italian National Institute of Statistics (ISTAT))

Abstract

R2BEAT ("R 'to' Bethel Extended Allocation for Two-stage sampling") is an R package for the allocation of a sample. Besides other software and packages dealing with the allocation problems, its peculiarity lies in facing properly allocation problems for complex sampling designs with multi-domain and multi-purpose aims. This is common in many official and non-official statistical surveys, therefore R2BEAT could become an essential tool for planning a sample survey. The package implements the Tschprow (1923) - Neyman (1934) method for the optimal allocation of units in stratified sampling, extending it to the multivariate (accordingly to Bethel's proposal (1989)), multi-domain and to the complex sampling designs case (Falorsi et al., 1998). The functions implemented in R2BEAT allow the use of different workflows, depending on the available information on one or more interest variables. The package covers all the phases, from the optimization of the sample to the selection of the Primary and Secondary Stage Units. Furthermore, it provides several outputs for evaluating the allocation results.

References

- Bethel, J. W. (1989). Sample allocation in multivariate surveys. Survey methodology 15 (1), 47–57.
- Falorsi, P. D., M. Ballin, C. De Vitiis, and G. Scepi (1998). Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall'istat. Statistica Applicata 10 (2), 235–257.; Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97 (4), 558–625.; Tschprow, A. (1923). On the two different aspects of the representative method: the method of stratified son the mathematical expectation of the moments of frequency distributions in the case of correlated observationsampling and the method of purposive selection. Metron 2, 646–683.

Using R and CANCEIS to edit and impute labor income on National Household Sample

Survey in Brazil

Authors

- Fernanda Karine Ruiz Colenghi Baptista (Instituto Brasileiro de Geografia e Estatística)
- Gabriel Henrique Oliveira Assunção (Instituto Brasileiro de Geografia e Estatística)

Abstract

The PNAD Contínua (National Household Sample Survey) produced by the Brazilian statistics bureau, IBGE (Brazilian Institute of Geography and Statistics), has as its main objective to capture information about the labor force and its percentage changes in rates in the short, medium and long terms; as well as on other topics, such as education, housing, other forms of work, among others. After the monthly database of PNAD Contínua is collected, non-response and measurement errors are processed in sequential steps in a system that uses CANCEIS (Canadian Census Edit and Imputation System) and SAS Enterprise Guide1 software to correct those errors and impute missing data through deterministic and probabilistic imputation tools. Deterministic imputations happen through pre-established rules in the SAS environment. The probabilistic imputations take place in CANCEIS, a package developed by Statistics Canada for missing data with the NIM - Nearest-Neighbor Imputation Methodology. The present work aims to show the probabilistic imputation of missing data with the scenario closest to that used by the IBGE to impute labor income. As the IBGE does not publish monthly microdata, the microdata from the 4th quarter of 2019 were read in the R environment with the PNADcIBGE package. Then, 10% of the records of work income were put missing for individuals who had more years of schooling, with the assumption that the non-response in labor income is correlated to a higher level of education. The data were generated in TXT format for reading the CANCEIS and the imputation was performed with the same configuration used in the PNAD Contínua, that is, the same explanatory variables, weight associated with the variables and the distance function with same parameters. After imputation, the database in TXT format was read using the R software and the Survey package was used to estimate labor income. 1

References

No References available

Using R and Github Actions to automate data reporting for the Sustainable Development Goals

Authors

- Maia Pelletier (Statistics Canada)

Abstract

Reporting the most up-to-date data on the Sustainable Development Goals (SDGs) is crucial to tracking Canada's progress towards meeting them and Statistics Canada is the central focal point for reporting Canada's data. Statistics Canada has two data hubs that report data for both the SDG indicators and domestic indicators through the Canadian Indicator Framework for the SDGs. Between the two frameworks, Statistics Canada reports data for over 200 indicators total. One of the many roles of the Statistic Canada's SDG team is to keep the reported data as up-to-date as possible. However, the data hubs were originally built for the data to be manually updated by the SDG team as new data became available. The team encountered challenges because manual updates can create an enormous workload burden, tracking the releases of dozens of sources can be difficult to keep up with, and manually editing data files is prone to human error. As the majority of the data reported in the hubs comes from Statistics Canada sources, the natural solution was to automate the updates by leveraging the Statistics Canada API. The data hubs require a specific data schema to be followed and data to be uploaded as CSV files. Thus, we chose to connect to the API through R using the StatCan API wrapper package {cansim} so that after retrieving data, it can be easily transformed using {dplyr} and {tidyr} and written to CSV files. The data for the hubs is hosted in Github repositories, so we use Github Actions and Workflows to perform a data refresh every time a change is pushed to the repository. The result of this work is over half of the reported indicators are updated through automated data refreshes, lightening the manual work required, taking away the risk of human error, and getting rid of the task of tracking new data releases. Freeing up this time will allow the team to take on more projects to modernize and innovate Statistics Canada's reporting on the SDGs.

References

No References available

Using R code to model small area estimates of household income in England and Wales

Authors

- Peshali Diyasena (Office for National Statistics)
- Matthew Plummer (Office for National Statistics)

Abstract

The Office for National Statistics (ONS) produces small area estimates of mean household income at 'Middle Super Output Areas' in England and Wales1. These estimates have previously been produced using models written in licensed platforms, and users were required to manually edit large parts of the code to run the models. In recent years, the ONS and UK government has focused on the use of open-source code for developing official statistics2, as well as implementing 'reproducible analytical pipeline' principles to automate code as much as possible3. For these reasons, we took the opportunity to translate the code used for modelling mean MSOA income to the R platform. Within this work, we have translated the code using various R packages, as well as 'hand-written' sections of code using base R functions, in order to apply various analytical models to the data including stepwise and mixed-model regressions and small area modelling. So far, we have been able to replicate the outputs produced from the original model code for estimates produced in 2018, giving confidence that the R implementation will produce quality estimates for years going forward. However, we believe that the R implementation offers several strengths to the original model code as we have been able to automate almost every aspect of the code, reducing the risk of human error being introduced in running the models. We have also taken the opportunity to refresh the guidance and supporting documents with the code to help users run models. Part of this has produced RMarkdown documents that produce key summaries of model outputs for users to interpret and inspect. In this talk, we will discuss the benefits of transforming the code to R, the challenges overcome during this work, and identifying lessons that can be applied for future similar projects in the ONS or other National Statistical Institutes.

References

- 1 MSOA estimates of mean household income for the year ending 2018 https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/incomeestimatesforsmallareasinenglandandwalestechnicalreportfinancialyearending2018; 2 Central Digital and Data Office (2021). The Technology Code of Practice. https://www.gov.uk/guidance/the-technology-code-of-practice; 3 Government Analysis Function. Reproducible Analytical Pipelines. https://analysisfunction.civilservice.gov.uk/support/reproducible-analytical-pipelines/

Using R to implement indirect estimation approaches: an application based on SDG Indicator 5.a.1

Authors

- Stefano Di Candia (FAO UN; OCS Department, Italy)
- Raymond Shama (FAO UN; OCS Department, Italy)

Abstract

The present paper aims to present the use of R in statistical analysis to develop and implement indirect methods to produce more accurate proxy estimates through the use of the so-called "projector estimator" proposed by Kim and Rao (2012)1. Specifically, the study enables the use of R facilities and additional packages to integrate data from independent complex surveys and estimate alternative and better proxy estimates for the SDG indicator 5.a.1. At the beginning of 2021, the global SDG database contained values of the indicator 5.a.12 for only 10 out of the 193 UN Member States. To address this gap, the FAO and UN Women developed methods for generating proxy measures from internationally-supported national surveys, such as the Demographic and Health Surveys (DHSs). DHSs gather data on agricultural land ownership, but these are not detailed enough to guarantee reliable direct estimates of indicator 5.a.1 according to the internationally agreed methodology3 endorsed by the UN Statistical Commission. The first section of the presentation analyses the methodological limitations of the 2018 Nigeria Demographic and Health Surveys (DHS)4 in generating proxy 5.a.1 estimates, by comparing them with the direct estimates computed from the 2019 Nigeria General Household Survey (GHS)-Panel5. The second section discusses the implementation of the projection estimator, which requires the availability of a common set of auxiliary variables in the two surveys to be integrated. To achieve an efficient projection, these variables also need to share a common structure and definitions. Basic R commands for recoding and harmonizing have been used extensively to satisfy this requirement. The third section illustrates the use of the Boruta package6 to implement the Boruta feature selection method for the identification of auxiliary variables to be included in the model from the GHS-Panel. Besides running the Boruta algorithm, Domir package7 was used to perform dominance analysis for predictive modelling functions. The fourth section discusses the use of the Survey package8 to generate a weighted multinomial logistic regression to estimate the projection parameters needed to compute synthetic values of the variable of interest. Furthermore, to assess the performance of the model at distinguishing between the two classes of the dependent binary variable, the ROC (Receiver Operator Characteristic) curve was built using the pROC9 package. In conclusion, by applying the estimated projection parameter, synthetic values of the variable of interest are obtained in the DHS dataset. This, in turn, allows producing estimates of SDG Indicator 5.a.1 using the package ReGenesees10. Results provide a nuanced understanding of the importance of using R in statistical analysis to improve sampling and estimation and, hence, enhance the accuracy of data and official statistics on gender-land inequalities.

References

No References available

Viewing Multiple Interactive plots with plotly and trelliscopejs

Authors

- Jeremy Selva (Singapore Lipidomics Incubator :: National University of Singapore :: Centre for Life Sciences)

Abstract

In a targeted lipidomics analysis workflow, quality checks are done to ensure a transition/compound measured in a sample is in good quality. This results in several quality control (QC) plots created for each transition and output as a pdf file. However, as technologies improved, many transitions can be measured in one sample quickly. Today, over 500 transitions can be easily be measured on large studies of a few thousand samples. Looking at several 500-page pdf files of static plots has its limitations. Thus, a different approach is required. For each transition, R package Plotly was used to create these quality control plots like Injection Sequence vs Peak Area scatter plot and Raincloud plots (or violin plot). R package Trelliscopejs is then used to view them as a trellis plot. QC statistics and other metadata can be converted to cognostics which can be used to filter the plots and provide additional information. Trelliscopejs objects can be output as a non-self contained html file which can be compressed in a folder and distribute to others. A Quarto report https://jauntyjjs.github.io/Trelliscopejs_In_Quarto_Example/ is created to show how it can be done using a published data set https://doi.org/10.1038/s41467-021-27765-9 used in clinical diagnostics for pancreatic cancer.

References

No References available