uRos (Use of R in Official Statistics)

Book of Abstracts (2021)

Contents

A Reportıng And Comparıng Applıcatıon For Provınce Based Covıd-19 Cases In Turkey Wıth R Shıny

Authors

- Cenk İÇÖZ (Eskişehir Technical University Statistics Department)

Abstract

Covid-19 pandemics which emerged in Wuhan, China has been effective still and caused deaths of many people from all over the world. Governments stepped up into a race of taking precautions to prevent the spread of the disease and eventual deaths. Those precautions include partial and whole lockdowns, wearing protective masks in public and shutting down public schools. Turkish government and health ministry in the beginning of the February 2020 announced the normalization of the precautions that they took and started a new procudure of evaluating the risks province- wide. A R-Shiny application is designed to compare weekly announced provincial base number of cases per hundred thousand and to report the differences automatically with interactive graphs and reports. The aim of designing the app is to enable to make comparisons to whole public.

References

No References available

Advanced Industrial Turnover Index using Statistical Learning Algorithms

Authors

- S. Barragán (Dept. Methodology and Development of Statistical Production, Statistics Spain)
- L. Barreñada (Dept. Statistics and Operations Research, Complutense University of Madrid)
- J.F. Calatrava (Dept. Methodology and Development of Statistical Production, Statistics Spain)
- J.C. Gálvez Sáenz de Cueto (Dept. Methodology and Development of Statistical Production, Statistics Spain)
- J.M. Martín del Moral (S.G. for Industrial and Services Statistics, Statistics Spain)
- E. Rosa-Pérez (S.G. for Industrial and Services Statistics, Statistics Spain; Dept. Statistics and Operations Research, Complutense University of Madrid)
- D. Salgado (Dept. Methodology and Development of Statistical Production, Statistics Spain; Dept. Statistics and Operations Research, Complutense University of Madrid)

Abstract

The modernisation of the production of official statistics should make use not only of new data sources but also of novel statistical methods applied to traditional survey and administrative data. The incorporation of both new data and new methods must improve the traditional quality standards in the production of official statistics. Here we present an application of statistical algorithms to improve the timeliness under a controlled compromise of accuracy of the Spanish Industrial Turnover Index (ITI). We follow the philosophy of the GREG and Sanguiao-Zhang estimators to propose specifically for a cut-off sampling design a combined projective-predictive estimator where already collected data are plugged-in in the estimator whereas not yet collected data are predicted. We use collected values up to time t to train an extreme gradient boosting algorithm on regression trees. We also use both microdata and data collection paradata of the immediately past history of the survey. Preliminary results on data from February 2021 invite us to cherish the hope that the use of these techniques may improve the quality dimension of timeliness while accuracy is kept under controlled. Currently, the survey managers of the Spanish ITI receive three batches for each reference month at days t+20, t+27, t+38 to finally release the first version of the index at t+51. Promising preliminary results are obtained for ITIs for different geographical and economic breakdowns using the preceding estimators trained with the successive batches as well as an initial estimation of the ITI without data of the reference period. Early conclusions from this experimental exercise are: Statistical learning algorithms with high predictive capacity on early data of the reference time period allow us to improve timeliness under a controlled compromise of accuracy; Subject-matter knowledge is crucial for its incorporation into the statistical model, especially on the selection and construction of regressors; Both microdata and paradata are relevant for high-quality predicted values; There is still a wide range to further investigate improvements on the predictions both on exploring more algorithms and further complementary data (other surveys, administrative data, and new digital data); Outliers are extremely hard to model and predict and they clearly need management and processing by subject-matter experts.

References

No References available

Autocoding based on Multi-Class Support Vector Machine by Fuzzy c-means Method

Authors

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

Abstract

In recent years, data in official statistics are getting larger and more complex. For developing an autocoding method for the coding task of the Family Income and Expenditure Survey data, which is large and complex data, we have developed the multi-class Support Vector Machine (SVM) by k-means method. In this method, SVM is applied individually to each cluster obtained by k-means method to create the discriminant surface by considering the feature of an individual cluster and obtain a more accurate result for the autocoding. However, since the k-means method is one of the methods of hard clustering in which each data is classified to only one cluster, the k-means method tends to need many clusters to obtain a better solution. In this case, the result has less solution robustness, and a large amount of computation is necessary for obtaining a better solution. To overcome this problem, this study proposes a new classification method based on multi-class SVM that is a combined method of SVM and a fuzzy clustering method. In fuzzy clustering, we can utilize the degree of belongingness of data to clusters and reduce the number of clusters. In this study, we utilize the fuzzy c-means method as the fuzzy clustering method. Several numerical examples show a better performance of the proposed method with the Family Income and Expenditure Survey data.

References

No References available

Automating Open Data Publication in Public Health Scotland

Authors

- Csilla Scharle (Public Health Scotland)

Abstract

Public Health Scotland (and formerly ISD Scotland) have been publishing open data on the Scottish Health and Social Care Open Data platform since 2017. The platform operates on a CKAN data management system, which has been maintained manually through the web interface by our small open data team up until the end of 2020. With the start of the covid-19 pandemic in 2020, we saw a sudden increase in interest for timely, reliable and well-structured health and social care information. With this sudden growth in both demand and volume of open data to be updated and maintained on the site, our small team had to make some big changes to our upload process to be able to keep up. Updating and replacing 17 files daily and over 500 data files each month manually on the platform had become unmanageable – for both our human team and our technical setup. Initially we developed python code to automate batch uploads of daily covid-19 data – however, we soon shifted to using R instead to align with the skills and experience within our team and support available more widely in our organisation, as well to fit better with our technical infrastructure. Scripts developed in R are scheduled to run on the PHS RStudio Server environment with CRONR, not requiring staff to be logged in or online. Similarly, automated checks are in place to run testing the files match the expected format and in the folders ahead of upload, with automated email notifications flagging up errors or issues that need human response to resolve. Another automated email is sent to confirm successful upload and archival of the daily data files, completing the process without any human interaction required. While this process is functional, we have some future improvements in the pipeline, including more robust error-handling of the pre-upload checks and data delays.

References

No References available

Automation forecasting of regional macroeconomic series

Authors

- Priscila Espinosa (Universidad de Valencia)
- José Manuel Pavía (Universidad de Valencia)

Abstract

In the same way as many other of agents, economic agents must make decisions in uncertain environments. Since the last crisis and the current pandemic, COVID-19, the number and source of uncertainties have financial-economic increased in magnitude and intensity. Regional agents, closer to the local reality, are looking for both mechanisms capable of showing synthetically the economic situation the regional economy is going through. This paper presents the experience of the Valencian Community (Spain) in the generation of a model of dynamic ensembles of economic forecasts from different organizations implemented in a web application developed with Shiny. The application allows the automation of the annual economic forecasting process, thus facilitating its management by the main economic agents.

References

No References available

Building a content delivery pipeline for a website in R

Authors

- Gregor de Cillia (Statistics Austria)
- Bernhard Meindl (Statistics Austria)
- Alexander Kowarik (Statistics Austria)

Abstract

Statistics Austria (STAT) is currently working towards the release of a new website. An important part of it, namely tables and graphs, will be created using a three step R workflow. 1.The first step of the process is the data import. For the start, two data sources will be supported: STAT's statistical data base STATcube and the open government data (OGD) portal of STAT, which is also used by other open data portals and therefore this part might especially reusable. The R package STATcubeR handles the data import; It uses a REST API for STATcube and direct CSV downloads for OGD data. 2.The next step is creating the interactive contents: graphs and tables. For that purpose, R packages were developed which utilize the javascript libraries datatable.js and highcharts.js. Both rendering packages use the data format from STATcubeR as an input and create the contents as html-widgets. 3.The last step is to make the content available for the content management system (CMS) where the website articles are written. This is done using a plumber API hosted at rstudio connect. The contents are first created on the R server and saved into a database. The API then accesses the database based on a content-id and returns javascript code for embedding. A graphical user interface was developed to assist non-R users with this process. The GUI was built in shiny and makes it possible to import and edit data to be fit for tabulation or plotting. Afterwards, the user can define customized tables and or graphs using in the GUI. The contents can then be saved and the content id is displayed to the user. The content id can be copy/pasted to the CMS which performs the embedding under the hood. Some of the packages involved in this workflow are available as open source and could be quite useful to other organizations in a similar situation.

References

No References available

Computing in the Statistical Office

Authors

- Mark van der Loo (Statistics Netherlands (CBS))

Abstract

Computing with data is at the hart statistical production[1]. Yet, the area of technical computing often falls between the stools of data analysts, IT developers, and methodologists. Typically, data analysts have strong domain knowledge but lack skills in technical computing and software engineering, IT developers have strong software engineering skills but little domain knowledge or knowledge of technical computing. Methodologists frequently have technical computing skills but typically lack strong software development skills. This combination of organizational roles often leads to suboptimal choices when it comes to building (local) statistical production systems. Simply because the necessary skills are not present in a team, or the skills needed are not recognized. This situation is underlined by the lack of attention payed to computing with data in documents such as the Generic Activity Model for Statistical Organizations[2]. In this talk I will analyze the importance of computing with data for activities in the Generic Statistical Business Process Model (GSBPM). Next, I will give a broad overview of skills that are needed to produce statistical output. Finally, I will argue that statistical offices should consider introducing the role of Research Software Engineer (RSE) into their workforce. A RSE combines intimate understanding of data analyses purposes with a solid knowledge of computing with data and software engineering. The current situation is that probably many NSIs have people with an RSE profile in their staff, but they receive no explicit recognition or growth opportunities for their specific set of skills. I therefore also discuss a simple junior/medior/senior model representing growth of RSEs in their role.

References

- [1] MPJ van der Loo (2021) Computing in the Statistical Office.JIAOS 37(3) pp 1023-1036.;[2] Modernstats. Generic activity model for statistical organisations. United Nations Economic Committee for Europe; 2019. Version 1.2.

Dealing with the change of administrative divisions over time

Authors

- Kim Antunez (Ecole nationale d'économie et statistique (ENSAE), France)

Abstract

The administrative divisions of countries change over time. In some countries, some territories change their codes, or names, merge or divide every year. This is the case in France and its communes, counties, regions, departments… As a result, it can be tricky to compare territorial databases from different dates. COGugaison (https://github.com/antuki/COGugaison) is an R package for manipulating French spatial databases produced at different dates by providing the list of existing towns and their history since fifty years and useful functions for transforming databases into geographic codes of other years. The R package CARTElette (https://github.com/antuki/CARTElette) contains geographical layers corresponding to the annual division of the French territories and functions that allow you to load them directly into the {sf} format in R. At this time, those packages only concern France and are therefore only documented in French. Presenting this issue and these packages in this European conference could be a good opportunity to see if (and how) they should be extended to other countries.

References

No References available

Determinants of house price in Europe

Authors

- Mihaela Cornelia Sandu (Faculty of Business and Administration, University of Bucharest)
- Irina Virginia Drăgulănescu (Faculty of Business and Administration, University of Bucharest)

Abstract

This paper aims to analyze the determinants of house price in Europe. As independent variables and also as determinants for house prices in Europe we will consider: new constructed dwelling, stock of existing dwelling, households (an indicator of the size of the regional real estate markets), unemployment, purchasing power index (an indicator of purchasing power in different countries), hospitals (a proxy for the quality of available public infrastructure), interest rate, disposable income (a proxy of affordability in housing stock). A panel data for period 2005 – 2019 for European countries will be analyzed using R package 'plm'. We expect a positive relationship with households, purchasing power index, hospitals, interest rate, disposable income and a negative relationship with new constructed dwelling, stock of existing dwelling and unemployment.

References

No References available

Determining the Business Cycle of Turkey

Authors

- Muhammed Fatih Tüzen (Turkish Statistical Institute)
- Fatma Aydan Kocacan Nuray (Turkish Statistical Institute)
- İlayda Kuru (Turkish Statistical Institute)

Abstract

In this study, it is aimed to examine the basic characteristics of the cyclical fluctuations in the Turkish economy and to determine the business cycles (contraction and expansion). By using the Bry and Boschan (1971) algorithm, the turning points in the business cycles were obtained. In order to determine the business cycle, Turkey's monthly Industrial Production Index, monthly and quarterly Gross Domestic Product data were examined and analyzed. As a result of the analysis, the average business cycle for the Turkish economy was calculated as 5 years. It has been observed that this result is compatible with the related studies in the literature and the cycle characteristics of developing countries. Peak and trough points were obtained with the algorithm named "Harding-Pagan (Quarterly Bry-Boschan) Business Cycle Dating Procedure" in the BCDating R package released in 2019. Apart from this, various R packages are used for analysis, visualization and reporting. "dplyr" and "tidyr" were preferred for data manipulation, "ggplot2" and "plotly" for data visualization, and RMarkdown for reporting and presentation.

References

No References available

Efficient Computation of approximate Kantorovich-Wasserstein distance for large spatial maps

Authors

- Stefano Gualandi (University of Pavia, Mathematics Department)
- Federico Bassetti (Politecnico di Milano, Mathematics Department)
- Fabio Ricciato (European Commission, DG Eurostat, Unit A5, Methodology; Innovation in Official Statistics)

Abstract

In this talk, we present a new R package, called Spatial-KWD and available on the CRAN [1], for the computation of approximate Kantorovich-Wasserstein distance (KWD) between pairs of 2-dimensional distributions defined over a regular grid. KWD is used in various application fields, including image processing and computer science, where it is known with other names, e.g., Earth Mover Distance. In spatial statistics, KWD is a natural choice a measure of (dis-)similarity between spatial maps representing the density of some physical or social quantities, e.g. density of pollutants or density of people, but insofar its adoption has been impeded by the high computational complexity of previous implementations. The computation of the distance (similarity) between two spatial maps can be achieved using the theory of Optimal Transport. The core algorithm of the package implements the Network Simplex algorithm customized for 2-dimensional spatial histograms [2]. The package delivers an approximate solution within a tight deterministic bound. It exposes a tunable parameter to trade-off approximation bound with computation resources (memory, time). Additional features are made available in the package to provide flexibility and adapt to the requirements of different applications. For example, in case of spatial maps with irregular contour, it is possible to compute the KWD over the convex hull or, alternatively, to restrict to geodesic paths within the irregular region. When comparing distributions with different total mass, different options are implemented to deal with the mass. An additional function allows to restrict the computation of the distance within a given focus area embedded in a larger map, so as to measure the local similarity around smaller regions of interest (e.g., a specific urban area within a country). The efficiency of the package is evaluated based on large set of semi-synthetic instances covering the whole Belgium at different grid resolutions. On the larger maps, at resolution of 125m x 125m for a total of 2.5 million of tiles, the package is able to compute the KWD approximation within 1.29% of the exact solution (worst case) in around 10 minutes on a standard desktop computer, while at the resolution of 1km x 1km the resolution completes in around 3.9 seconds within 0.12% of the exact solution. The Spatial-KWD package is the results of a joint collaboration with Eurostat. The source code of the package is freely available online at https://github.com/eurostat/Spatial-KWD. The Spatial-KWD package has been used recently to assess the spatial accuracy of different methods for spatial density estimation of present population based on mobile network operators data [3].

References

- [1] Gualandi S. CRAN Package SpatialKWD, 2021/5/7. http://ftp.uni-sofia.bg/CRAN/web/packages/SpatialKWD/SpatialKWD.pdf; [2] Bassetti F., Gualandi S., Veneroni M. On the computation of Kantorovich-Wasserstein distances between 2D-histograms by uncapacitated minimum cost flows. SIAM J. Optim., 30(3), 2441–2469, 2020.; [3] Ricciato F., Coluccia A. On the estimation of spatial density from mobile network operator data. arXiv preprint, arxiv.org/abs/2009.05410

EMOS coding labs – presentation of preliminary results

Authors

- Eurostat (-)

Abstract

For Eurostat and EMOS community Coding Lab is a way to disseminate outputs that document and share statistical processes beyond just data, but also including code, algorithms, protocols and workflows. In the Statistics Coded users can interactively dialogue with data, explore, reuse and contribute to statistical products. They are distributed in the form of interactive and portable notebooks (in formats such as Jupyter, Rmarkdown and Observable) that combine explanatory text, executable code (in languages such as R, Python and JavaScript), output results and visualisations.

References

- https://ec.europa.eu/eurostat/cros/content/coding-labs_en

Extending data validation with standardised metadata from SDMX registries

Authors

- Olav ten Bosch (Statistics Netherlands)
- Mark van der Loo (Statistics Netherlands)

Abstract

Standardisation of metadata is crucial in official statistics. A harmonized statistical picture of society and economy can only be produced if metadata such as classifications, code lists and variables are agreed among organisations. Internationally these metadata elements are stored in Statistical Data and Metadata Exchange (SDMX) registries. Examples are the Global Registry and the Eurostat Registry. They provide the concepts, variable definitions, data flows, structures and code lists underlying international official statistics. Moreover these registries provide a standardized application programming interface (API) to automatically access specific versions of the metadata from any programming environment. This makes them a key element in international validation processes and hence important for data validation practices in general. The R-package validate is a popular tool in official statistics to validate data. Based on a set of predefined validation rules, varying from variable checks to multivariate or statistical checks, the software can execute analyses on possibly large datasets, providing the user with extensive feedback on the health of their data. The software supports the majority of checks that are actually needed in today's practical International data validation exercises. Validation results are presented graphically and in a machine readable standardized validation report, which can be used as input for consecutive processes. Common rules can be re-used among multiple validation processes. For more details we refer to the online validation cookbook offering recipes for the most common validation scenario's. In the ValidatFOSS21 project we extended the R validate package with data validation based on SDMX metadata. The aim was to make data validation as easy as possible by re-using metadata already provided in SDMX registries. The SDMX metadata such as the dimensions, attributes, code lists and data representations are automatically retrieved from the respective registry based on the identifiers specified and cached for consecutive validation runs. The architecture presented is generic and can be used on any SDMX 2.1 compliant SDMX registry from any international organisation. Alternatively the approach can be used on local DSD files or an organisation-internal registry. In this presentation we present the results of this endeavor. We address the main approach, our experiences working with the international SDMX registries, some examples of the use in practice and we show the setup of a new chapter of the validation cookbook on connecting to SDMX.

References

- Global registry: https://registry.sdmx.org; Eurostat registry: https://webgate.ec.europa.eu/sdmxregistry R-package Validate: https://cran.r-project.org/package=validate; Validation cookbook: https://cran.r-project.org/web/packages/validate/vignettes/cookbook.html; Awesome official statistics software: http://awesomeofficialstatistics.org

Flash reports with flexdashboard for the characterization and quality levels of available datasets

Authors

- Almiro Moreira (Statistics Portugal)
- António Portugal (Statistics Portugal)
- Bruno Lima (Statistics Portugal)
- Cecilia Santos (Statistics Portugal)
- João Poças (Statistics Portugal)
- Jorge Magalhães (Statistics Portugal)
- Paula Cruz (Statistics Portugal)
- Paulo Saraiva (Statistics Portugal)
- Salvador Gil (Statistics Portugal)
- Sofia Rodrigues (Statistics Portugal)

Abstract

In a very near future, Statistics Portugal will be facing two major challenges: the production of annual census statistics based on administrative data and the move into a more regular and efficient production of essential indicators, with a reduced burden on respondents (including enterprises). Under the cooperation established between Statistics Portugal and relevant external organizations (such as the Portuguese Tax and Customs Authority), a very significant volume of administrative data from various sources is currently available for internal use and sharing with researchers and other institutions. With a view to facilitate data analysis and exploration, and similarly to what was already being done within the business surveys domain, the production and use of flash reports (or similar reports such as dashboards) has been adopted, to visualize, for example, the characterization and quality levels of available datasets, for a variety of ends. Another great advantage of flash reports is that it is possible to identify potential major problems early on and to act on them, through periodic snapshots of key operational data. In the flash report created for the information on invoicing data received from the Portuguese Tax and Customs Authority, for instance, indicators on the evolution of monthly and year-on-year invoicing are included, alongside the distribution by activity and consumer groups and a brief comparative trend analysis with the data collected within the monthly short-term business statistics surveys. Due to its ease of use and interactivity, and taking into account the intended objectives and the amount of the information in consideration, the open-source package flexdashboard, based on R Markdown, has been chosen for that purpose. Within its main features, it is worth highlighting the following ones: a wide variety of components is supported; possibility to specify row and column-based layouts in a flexible and easy way; extensive support for text annotations; possibility to create storyboard layouts; the fact that default dashboards consist of standard HTML documents, enabling their deployment on any web server or even a simple attachment to an email message. Currently, reports implemented in flexdashboard are restricted to internal users and a relatively low, although important, number of data sets. However, it is intended to extend its use and scope, namely through the creation of a common template that can be easily adapted to different data sets, regardless of the source of the information.

References

No References available

Identifying outliers in multivariate databases - distance and density-based methods in housing statistics

Authors

- Antal Ertl (Hungarian Central Statistical Office and Corvinus University of Budapest)

Abstract

In the following paper, I would like to contribute to the literature of outlier-handling, concentrating on economic statistical data, namely observations in housing statistics. In order to create indices for changes in price, data cleaning, as well as model-optimizing is required – and for both, identifying outlying observations is crucial. By applying various techniques, such as distance based and density-based outlier-detection methods, I would like to highlight the importance of dealing with outliers, and discuss the difficulties one might encounter. Housing statistics is a special case, as there is a high correlation between price and the area of the dwelling in question, but it still serves as a fine example of handling outliers in economic and transactional data. While there is great literature on newly created alternatives for outlier detection, there is relatively little research on how these fare with big, economic datasets. For this reason, I use online data on housing rental prices to demonstrate different outlier detection methods implemented in R, such as Feature Bagging Outlier Detection (FBOD) and Local Outlier Factors (LOF). I show that identifying outliers is a rather nuanced thing, where statisticians could benefit from using advanced algorithms, while also highlighting some of the difficulties one might encounter when implementing these methods.

References

No References available

Measuring Inequality in Sampling Surveys

Authors

- Sebastian Wójcik (Statistical Office in Rzeszów (Statistics Poland))

Abstract

One of the core activities of official statistics is to conduct representative surveys. Weights are an integral part of sample surveys. Among the various R packages for data analysis, only some of them implement methods suitable for weighted data. In the field of inequality analysis, there is a package called "ineq" which contains several implementations of inequality measures (e.g, Atkinson, Gini, Theil), but no weighting. Specialists of the Statistical Office in Rzeszów added the Palma ratio, the 20:20 ratio, the Hoover index, the Jenkins index, the Cowell and Flachaire index for continuous variables, and the Leti index, the Allison and Foster index for discrete variables. All methods have been implemented with the possibility of weighing data. Moreover, the standard deviation and confidence interval for all indicators were estimated using the bootstrap method. Finally, this new tool was used to analyse data from a representative survey Participation of Polish residents in trips conducted by the Statistical Office in Rzeszów.

References

No References available

Minimising MSE in the rolling windows implementation

Authors

- Virgilio Pérez (Universidad de Valencia)
- José Manuel Pavía (Universidad de Valencia)
- Cristina Aybar (Universidad de Valencia)

Abstract

To increase the sample size, we sometimes use information from nearby objects/subjects, if these nearby objects/subjects behave similarly. To verify this assumption, we have used a large database, created based on about 400 public opinion surveys (barometers) over more than 30 years, obtaining more than 1,000,000 observations. We have found that, indeed, the behaviour of certain variables changes very smoothly over time, such as political ideology, among others, so it is not unreasonable to increase the sample size of a subset by implementing the rolling windows technique. But should we assign the same weight to all the data in each subset generated with this method? We have considered different scenes, using different combinations of weights, obtaining that one of the best proposals is the combination of weights that minimises the MSE of the estimator. To automate this process, we are developing a package in R that allows us to perform these calculations, for each subset and for any window size, a tool that may be of great use to the scientific community, in any knowledge area.

References

No References available

Missing values treatment in R

Authors

- Adrian Dușa (University of Bucharest)

Abstract

The traditional and well established missing value in R is called "NA". Unlike other statistical packages such as SAS, Stata or SPSS, where multiple missing values might be defined, R has the unique value NA to indicate a missing value. However, not all missing values are equal, especially for survey data in social statistics. There are research situations where there is a qualitative difference between the reasons why a value is missing: respondents who do not know the answer, are very different from the respondents who do not want to respond, and this has direct implications on missing values imputation procedures. Such values are recorded but they should be treated as genuine missing values in statistical operations. This presentation introduce the package declared which helps declaring multiple missing values in R, with a direct application on converting between various statistical formats using the package DDIwR.

References

No References available

Modern graphics for presenting the results of life-events surveys on satisfaction with public administration

Authors

- Sylvana Walprecht (Destatis)
- Daniel Kühnhenrich (Destatis)

Abstract

The Federal Government Service Centre for Better Regulation at the Federal Statistical Office of Germany (Destatis) supports the Chancellery and ministries its expertise in data collection and analysis. As part of this assistance and on behalf of the Federal Government, Destatis has been conducting life-events surveys every two years since 2015 to determine the satisfaction of citizens and companies with government services in Germany. A number of new graphic types such as Sankey diagrams, word clouds and network diagrams were introduced to illustrate results of the 2019 survey. These were implemented in R via plotly, wordcloud2 and igraph packages.

References

No References available

Modified Stahel-Donoho estimator for outlier detection in official statistics

Authors

- Kazumi Wada (Statistical Research and Training Institute, MIC)

Abstract

Outliers are often unavoidable in survey statistics, and it is necessary to avoid their influence on the final products. Nevertheless, multivariate outlier detection methods may not be commonly used yet, unlike univariate methods. The modified Stahel-Donoho estimators were adopted to detect multivariate outliers by Statistics Canada in 1997. NSTAC implemented the estimators on R in 2010 with a few modifications suggested at the Euroedit project for evaluation. The implemented R function is now practically used by the Statistics Bureau in Japan after comparisons with other methods such as Fast-MCD, BACON, and Epidemic Algorithm (EA). Suggestions by the Euredit project improve the performance; however, the improved method is suffered from the curse of dimensionality. The limit is 11 variables for datasets of 100 observations with a 32 bit PC. An R function for higher-dimensional datasets is also implemented in 2013 using foreach function of doParallel package. We are now reviewing those functions to enhance convenience for users. The presentation will explain why multivariate methods are necessary. Then the features of the estimators will be illustrated with a real data example.

References

No References available

Network Visualization of Multi-data Sources using R

Authors

- Rui Alves (Statistics Portugal – INE)
- Shirley Ortega-Azurduy (Statistics Netherlands – CBS)
- Christina Pierrakou (Hellenic Statistical Authority – ELSTAT)

Abstract

The use of multiple data sources in official statistics poses many challenges such as data discrepancies, incoherent concepts, indirect relations between sources and redundancies of information, just to name a few. These issues are far from being trivial and may be easily either overlooked or underestimated. Big data along with administrative data are major data sources that are not produced for the purpose of official statistics and require strong analytical tools. In fact and for a long time, it's been perceived that their potential has not being fully explored. In this paper, an innovative visualization implemented in R is introduced. The use of the package visNetwork[[1]] renders a dynamic and comprehensive platform-free network for statistical researchers in the field of official statistics of Travel and Tourism and assists them during the assessment of multi-purpose data sources. In the last decade data sources have expanded horizontally (more domains or areas) and vertically (more data for each domain or area) making it increasingly complex to have an up-to-date overview of these sources such that user can manage, search and extract implicit information and their connections. To overcome the aforementioned challenges, there is a need for a proper user-friendly solution to visualize, explore and help to deal with the complexity at hand. Network visualization is a valuable and flexible alternative to present a dense and complex set of heavily interconnected data sources, particularly when it comes to interactivity. This visualization was envisioned and developed under the ESSnet programme on Big Data 2018 -2020[[2]]. Herein; one of the goals was to provide an interactive graphical representation of connections between data sources (multi-purpose data sources, survey data and web scraped data), variables, domains, countries and experimental results. Our approach was inspired by network analysis which has become an increasingly popular tool to deal with the complexity of interrelationships. The two primary aspects of networks are a multitude of separate entities and the connection between them which are referred to as nodes of a graph (for example, statistical domains or data sources), while the connections are edges or links. The elements of this map of relationships; nodes - which can also be URL's - and edges, can be graphically displayed with different sizes, lengths, shapes and colours in order to provide a more intuitive and accurate view. Although there are several applications designed for network analysis, R has grown into a powerful tool for statistical analysis. The strength of R network analysis in comparison to other stand-alone network analysis software is that R enables reproducibility and provides robust tools for data analysis. Moreover, the use of R language is quite common at National Statistic Institutes (NSI) and, therefore, it's not only faster to produce and disseminate results, but also to share code and use it in current existing workflows. The visNetwork R package allows an interactive visualization of networks, and R's ability to read and write multiple formats makes it compatible with shiny, R Markdown documents, and RStudio viewer. The developed R script is "self-contained" in the sense that data are embedded in the code; therefore it recreates data and a visNetwork object. The visualization is displayed on RStudio viewer and exported as a stand-alone HTML file, where interactivity is a key feature. The level of relevancy of a source is imbedded since the nodes with more connections (edges) are automatically placed on the centre. You can also zoom in and out on the plot, move it around and re-centre it. Hoovering a selected node will present additional information and in specific nodes external webpages will be open upon double-click. Moreover, filtering options are available using two combo-boxes. One of the main advantages is the dynamic layout of the VisNetwork. It engages the user into the analysis of multi-purpose data as he/she can interact with the tool by dragging a node as the graph will literally pull all other direct and eventually indirect connected nodes along with it. In this way, the user has a visual "feel" of the importance of choosing a data source or domain (node) in the whole network. The aggregated value of this network is analytical engagement which is otherwise impossible in a static alternative. And it's also quite fun to play with.

References

- [1] Thieurmel, B., "visNetwork: a package for network visualization", release 3.0.8, Augustus 2019. https://github.com/datastorm-open/visNetwork; [2] [Essnet2018] ESSnet on Big Data 2018 - 2020 - Eurostat grant ESTAT-PA11-2018-8 Multipurpose statistics and efficiency gains in production. https://ec.europa.eu/eurostat/cros/essnet-big-data-2_en

Official estimates of community Covid positivity: estimating rare occurrences using multi-level regression with post-stratification in R

Authors

- Melissa Randall (Office for National Statistics, UK)

Abstract

Following the onset of the COVID-19 outbreak in the UK it was crucial to understand how COVID-19 was spreading across the population in order to control the pandemic and its effects. Decisions regarding the continued need for control measures to contain the spread of SARS-CoV-2 rely on accurate and up-to-date information about the number of people and risk factors for testing positive. The Office for National Statistics set-up the Covid Infection Survey to do this, repeatedly swab participants and generate an official estimate of covid positivity for the UK. Analysis needed to deal with relatively small numbers of positive tests and ensure estimates were unbiased. Multilevel regression and post-stratification (MRP) is a statistical technique used for correcting model estimates for known differences between a sample population and a target population. It has been shown to be an effective method of adjusting the sample to be more representative of the population for a set of key variables; previously it has been used in polling to predict the US election results, but it isn't widely used in other settings. MRP consists of two steps. First, an MRP is used to generate the outcome of interest as a function of (socio)demographic and geographic variables. Next, the resulting outcome estimates for each demographic-geographic respondent type are poststratified by the percentage of each type in the actual overall population. For the Covid Infection Survey, the result was to produce an estimate which takes account of differences by age, sex, time and region. All analysis for the Covid Infection Survey is run in R - this presentation will share further detail on the method and the resulting estimates.

References

No References available

Peer-Reviewing Code in Statistical R Packages: Standards, Processes, and Tools

Authors

- Mark Padgham (rOpenSci)
- Noam Ross (rOpenSci)

Abstract

Peer-review of statistical algorithms is challenging and labor intensive, and even software described in peer-reviewed literature rarely has its source code reviewed. Here we describe a peer-review system to support reviewing R code implementing statistical algorithms, developed by rOpenSci, an organization promoting open science through code and data sharing and peer code review. We have created a series of standards, guidance for authors and reviewers, and automated tools to facilitate review of statistical code. These include (1) a series of community-developed standards for documentation, testing, code architecture and programming APIs, of packages, with specialized sub-standards for specific areas such as regression, machine learning and Bayesian algorithms, (2) a series of packages for automated testing and literate programming to document standards compliance as part of code development, and (3) an open peer-review process supported by automated testing, reporting, and editorial handling. Together with our community of editors and reviewers, these tools form the backbone of our new peer-review system, and can also be adopted by individuals and organizations for internal review and validation processes.

References

No References available

Probabilistic record linkage using reclin

Authors

- Jan van der Laan (Statistics Netherlands (CBS))

Abstract

A lot of additional value can be gained by combining information from different datasets. For example, by linking data on income from the tax office to a survey on education, the value of different diplomas can be determined. Sometimes use can be made of unified identifiers present in both data sets. For example many government institutions use the same personal identifiers in all data sets concerning citizens allowing us to link those data sets using those identifiers. However, it is also common that, when linking two data sets, in at least one of the data sets such an identifier is missing. In that case linkage is often performed on a set of identifying variables such as last name, date of birth and address information. Unfortunately, these types of variables will often contain errors, for example people have moved, making address information no longer correct, and there can be spelling errors in names. Probabilistic record linkage is a method that can be used in these cases. In probabilistic record linkage, a model is estimated that estimates the probability for each combination of records from the two data sets that these records belong to the same target object (e.g. person or business). In the workshop, an overview will be given of the different steps in probabilistic record linkage process. Next to classic probabilistic record linkage using EM, the workshop will also discuss estimating the probabilities using machine learning methods. The methods will be applied using the reclin package which is an R-package for probabilistic record linkage.

References

No References available

Quality control with R on Simplified Business Information

Authors

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

Abstract

Simplified Business Information (IES) is a mandatory annual declaration, for enterprises and individuals with organized accounting, delivered electronically. This declaration collects, in a centralized way, information for accounting, tax and statistical purposes. Implemented in 2007, IES has been a facilitator to enterprises' compliance with their legal obligations. IES system has internally integrated around 2000 validation rules, which constitutes a guarantee that data itself is fully coherent. The next step in increasing its value is the integration of this data with other sources. One of the quality controls for IES uses data collected by INE - International Trade (CI) - regarding annual amounts of enterprises' imports and exports. Data from IES is joined to data from CI for a group of selected enterprises and their values are compared. This validation is performed in R with packages like {tidyverse} for data manipulation; {validate} to declare data validation rules and data quality indicators; and {dcmodify} to modify data using external rules. Data verification rules are pre-defined based on knowledge in the analyzed domain. These rules, as new variables to be created, are stored in a table, and confronted to our data. Each observation is evaluated according to the rules and new variables are created using 'if clauses' adapted from these rules.

References

No References available

R contribution to data quality: the E-invoice case

Authors

- Bruno Lima (Statistics Portugal)
- João Poças (Statistics Portugal)
- Sofia Rodrigues (Statistics Portugal)
- João Lopes (Statistics Portugal)
- Paula Cruz (Statistics Portugal)
- António Portugal (Statistics Portugal)

Abstract

Statistics Portugal (INE) receives monthly, from the Portuguese Tax Authority (AT), about 80 million records regarding taxable amounts aggregated by issuer and acquirer's VAT number. These amounts result from invoices issued by individuals or legal entities that have their head office or permanent establishment in the Portuguese territory. Records are obtained from an implemented mandatory E-Invoice system as part of the administrative simplification and anti-fraud measures from the Portuguese authorities. To become statistical data, these administrative data must be treated and validated, to ensure its quality, reliability, consistence, and completeness. In this data cleansing process, we also perform a more in-depth and specific analysis of anomalous data or lack of information. In this context, we implemented an iterative procedure for outlier's detection and the imputation of values to replace those outliers. The process is based on R packages: {tidyverse} (mainly {dplyr}, {ggplot2} and {purrr}) for data manipulation, visualization, and functional programing; {isotree} for outlier detection; {imputeTS} for univariate time series imputation; and {targets} that allows to implement a reproducible workflow.

References

No References available

R for Big Data processing: Online Job Advertisements

Authors

- Matyas Meszaros (Eurostat, European Commission)
- Andrea Ascheri (Eurostat, European Commission)
- Fernando Reis (Eurostat, European Commission)
- Gabriele Marconi (Sogeti)

Abstract

The Web Intelligence Hub (WIH) is an initiative by Eurostat aimed at building capabilities throughout the European Statistical System to collect various data from the web to enhance statistical information in multiple domains. Among web data, Online Job Advertisements (OJA) have great potential for labour market and skills analysis. This experimental study by Eurostat is based on the work done within the Work Package B of the ESSnet project on Big Data II [1], namely the R code developed by Destatis [2] to calculate concentration of labour markets using online job advertisements. The methodology developed by Destatis has been adapted and replicated on all the 27 EU Member States. For this study, more than 100 million distinct ads have been analysed coming from more than 300 sources. To ensure reduced running time of the code, we introduce parallel processing on all 27 countries. The R code establishes a connection with AWS database services where there data are stored and from where they are fetched in parallel for several countries at the same time. The code for calculation of the concentration index is also parametrised for parallel running to be able to adjust it to the available memory. The OJA data is automatically combined with other information retrieved from Eurostat database [3] and GISCO services [4] during the calculation. In addition, a methodology to deal with company names is proposed that (i) merges similar company names strings based on keywords and exclusions and (ii) identifies intermediary/staffing agencies based on keywords and predefined rules. Finally, for the first time the code of an experimental statistics published by Eurostat is shared on the Eurostat Github account [5].

References

- [1] https://ec.europa.eu/eurostat/cros/content/WPB_Online_job_vacancies_en; [2] https://github.com/OnlineJobVacanciesESSnetBigData/Labour-market-concentration-index-from-CEDEFOP-data; [3] https://ec.europa.eu/eurostat/data/database; [4] https://ec.europa.eu/eurostat/web/gisco/; [5] https://github.com/eurostat/oja_hhi

reclin: a package for probabilistic record linkage

Authors

- Jan van der Laan (Statistics Netherlands (CBS))

Abstract

Record linkage is the process of combining two data sets, linking records belonging to the same object (person/business) to each other. Sometimes the data sets contain a unique identifier that can be used to reliably link the two data sets to each other. However, often at least in of the data sets such an identifier is missing. In that case linkage is often performed on a set of identifying variables such as last name, date of birth and address information. Unfortunately, these types of variables will often contain errors making record linkage more difficult. Only linking records without errors can lead to only a small fraction of the records being linked which in term causes issues with selectivity and data size. Probabilistic record linkage tries to take into account errors in the linkage keys. A model is estimated that estimates the probability for each combination of records from the two data sets that these records belong to the same target object (e.g. person or business). These probabilities can then be used to link the two datasets taking into account of errors. In the presentation an overview of probabilistic record linkage will be given together with how reclin fits into this.

References

No References available

Selecting auxiliary variables in R

Authors

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

Abstract

When imputing the missing values in survey or administrative data it may be necessary to identify the "best" predictors of a target variable, in order to use them as explanatory variables in an imputation model or to calculate the distances when applying a donor-based imputation method. This one possible example of using a "working" model, implicit or explicit; e.g. a model that will not be used for inference purposes but just for solving specific problems (imputation, calibration, etc.) without introducing additional noise or bias into the final results (typically estimation of means, totals, proportions, marginal distributions, etc.). In these cases, we need to identify a subset of "auxiliary" variables that will play the role of predictors of the target variable in the working model; in this seek the preference is given to parsimonious models, relatively simple to be applied. The R environment provides many functions, mainly in additional packages, that can be used to select the "best" predictors of a target variable. Most of them are tailored to fitting regression models. The tutorial will give an overview of "simple" and "complex" methods that can be used in R when fitting a working model to survey or administrative data, taking into account the nature of the target variable (categorical or continuous) as well as that of the available variables (all continuous, all categorical, mix of categorical and continuous). The tutorial will also give some hints on selecting the predictors by means of well-known supervised statistical learning methods (e.g. ranfomForest).

References

No References available

Selective Editing Using Contamination Model in R

Authors

- Ieva Burakauskaitė (Statistics Lithuania)
- Vilma Nekrašaitė-Liegė (Vilnius Gediminas Technical University, Lithuania and Statistics Lithuania)

Abstract

The aim of the conducted study was to optimize the statistical survey data validation process using selective editing when predictions of the relevant statistical survey indicator are obtained using the contamination model. All the calculations were carried out with the statistical programming language R and its package SeleMix that is designed to execute the selective editing method. Selective editing identifies observations affected by errors that have a major impact on the quality of sample estimates. This way the data editing process can be focused on the corresponding observations while preserving human resources and time costs though maintaining the quality of sample estimates. Selective editing was applied to the data editing process of the quarterly statistical survey on service enterprises (turnover indicator) of Statistics Lithuania. Predictions of the target variable were obtained using the contamination model. An impact of a potential error on the sample estimate was evaluated using the score function with a standard structure – a difference between the observed value of the target variable and its prediction multiplied by a sample weight and a suspicion component. Discrete and continuous suspicion components were used and an impact of the suspicion component on the effectiveness of selective editing was investigated. Main results of the study as well as remarks for future studies are provided in the presentation.

References

No References available

Shiny demo: Mobility scan

Authors

- Josue Aduna (Livemobility, Netherlands)

Abstract

This is a Shiny application designed and developed to foster sustainable mobility behavior under a specific initiative that I currently work in: Livemobility (see https://www.livemobility.com/). Broadly speaking, Livemobility is a platform that rewards people for sustainable commuting behavior and helps companies to save money, avoid environmental pollution, improve public health and save travel time. This is achieved through a digital ecosystem that analyses mobility behavior and generates personalized insights to improve mobility efficiency. This Shiny app makes use of web interactive settings together with Google Maps APIs to provide relevant indicators of impact, generate geographic scans and create mobility profiles. Example: https://www.youtube.com/watch?v=0v4lfH4rsfs&t=4s

References

No References available

Simulation of synthetic microdata: an overview of packages

Authors

- Jiří Novák (Czech Statistical Office; Prague University of Economics and Business, Faculty of Informatics and Statistics)
- Lubomír Štěpánek (Czech Statistical Office; Prague University of Economics and Business, Faculty of Informatics and Statistics)

Abstract

Simulation of synthetic microdata is a wide range of methods that offer excellent protection against data disclosure. These methods create new synthetic microdata from the original data set, which usually does not contain the original values, but it should preserve the relationships between the variables and should maintain the hierarchical structure contained in the data. The great advantage of these methods is that they allow statistical offices and agencies the dissemination of datasets, which would otherwise have to remain hidden and confidential under normal circumstances. The aim of this paper is to offer an overview of methods and their packages (the most known are simPop and synthpop) for synthetic simulation of microdata and to compare their efficiencies, stability and drawbacks of the packages.

References

No References available

simviz: a package to visualize simulated telecommunication mobile network event data

Authors

- B. Oancea (Dept. Innovative Tools in Official Statistics, National Institute of Statistics (INS); Dept. Business Administration, University of Bucharest)
- D. Salgado (Dept. Methodology and Development of Statistical Production, Statistics Spain; Dept. Statistics and Operations Research, Complutense University of Madrid)
- S. Barragán (Dept. Methodology and Development of Statistical Production, Statistics Spain)
- M. Necula (Dept. Innovative Tools in Official Statistics, National Institute of Statistics (INS), Romania)

Abstract

The incorporation of mobile network data into the daily production of official statistics is proving to be a tough task with multiple issues to be tackled (legal, business, and technological conditions to access, statistical methodology and quality issues, etc.). Initiatives in the European Statistical System (ESS) can be found in different National Statistical Institutes (NSIs) and joint projects and working groups as the ESSnet on Big Data I and II and the ESS Task Force on MNO Data. As one of the approaches to develop the statistical methodology, synthetic network event data provides a way to research on different aspects of both this new data source in official statistics and the statistical methods needed to provide high-quality statistical outputs. One key aspect to work with these data and with the novel proposed methodology is to have a visualization tool dealing with appropriate data structures. In this contribution we present the first functionalities of the package simviz, focused on providing specific data formats and vizualization functions for these simulated data. The package includes the following characteristics: - Standardization of data structures through a combination of csv, xml, and xsd files. - Use of sf and stars functionalities to represent and manage data. - Use of grammar of graphics (mainly through ggplot2) to visualize multiple aspects of the simulations. This package is intended to be used in the development of the end-to-end statistical process going from raw telco data to final statistical outputs proposed in [1].

References

- [1] Salgado, D., Sanguiao, L., Oancea, B., Barragán, S. and Necula, M. An end-to-end statistical process with mobile network data for official statistics. EPJ Data Sci. 10, 20 (2021). https://doi.org/10.1140/epjds/s13688-021-00275-w

Spatial density estimation based on mobile network operator data within R: the MNO-simulator

Authors

- Marco Ramljak (Utrecht University, Netherlands)

Abstract

Data generated by the cellular network of a mobile network operator (MNO) represent a rich potential source for estimating the spatial distribution of mobile phones at some given time and, from there, gain insight into the temporal variations of the spatial distribution of humans – relevant for applications in, e.g., demography, tourism statistics and urban planning. Given the non-trivial estimation problem, current research highlights multiple research gaps and interests, concerning, e.g., estimation strategies, available data input, in-production execution, etc.. Several of these research projects are complex and involve the development of non-trivial software with the necessary rigor, prior to the actual simulations and analyses. This poses a high barrier for researchers. Through our individual thematic research within MNO-research we generalized and modularized our developed software workflow into an easy to use tool (MNO-simulator) within a software framework known to many researches in the field, namely R. R is an excellent statistical programming language to develop an end-to-end analysis workflow for this, given the availability of established spatial statistics resources such as the sf package, as well as the specialized package mobloc, focusing on radio propagation modelling. We developed a modular and computationally efficient workflow that enables the research on MNO-related research tasks (via simulation) as well as the real world application of spatial density estimation (with accessible MNO data). Furthermore, useful functions concerning the modelling and estimation process are consolidated in a custom package, such as tessellation based estimators (Voronoi and variants thereof) as well as estimators based on probabilistic models (SB, MLE, DF). The workflow contains three modules: (1) Toyworld Generation, which gives flexible options of generating multiple scenarios concerning mobile phone density and network density as well as assigns mobile phones to cells based on an individually specified generative model; (2) Estimation, which helps with specifying geo-location modules (creating estimation models) and executing all state of the art estimation strategies; and (3) Evaluation, which offers multiple quantitative evaluation metrics, such as the Kantorovich-Wasserstein Distance from the SpatialKWD package taking the spatial nature of the estimation problem into account. All modules work independently from each other, meaning that the user can individually decide which module is needed for the analysis project, as well as contain multiple predefined graphs that help to depict the descriptive details at any stage of the analysis project. Furthermore, our tool works with any kind of data availability – studies can be conducted with complete synthetic scenarios (no real-world data available), semi-synthetic scenarios (some real-world data available, e.g., census data or (partial) cell coverage data), and of course real scenarios if real world data are available. A future presentation (and paper) will serve as a guide on the usage of the MNO-simulator.

References

No References available

Structured Effects with Generalized Nonlinear Models

Authors

- Heather Turner (R Foundation; Statistics Department at the University of Warwick, UK)

Abstract

Generalized linear models are a standard tool in statistical analysis. In particular, the special cases of logistic regression and log-linear models are go-to methods for modelling binary and count (or rate) data respectively. However, using a linear predictor for the mean response can be quite limiting - models with satisfactory goodness of fit are often overly complex and difficult to interpret. Using a nonlinear predictor can be a way to solve both these issues. This talk will introduce the R package gnm, which provides facilities for specifying and estimating Generalized Nonlinear Models (GNMs). The application of GNMs will be demonstrated through two case studies in sociology and demography. The first will demonstrate the use of a structured main effect, by using Diagonal Reference Models to model the diluting effect of social mobility on health inequality. The second will demonstrate the more common use case of a structured interaction, by using the Lee-Carter model to model mortality trends.

References

No References available

Synthetic data with xgboost and advanced calibration

Authors

- Johannes Gussenbauer (Statistics Austria)
- Matthias Templ (Zurich University of Applied Sciences)
- Siro Fritzmann (Hotelplan, Switzerland)
- Alexander Kowarik (Statistics Austria)

Abstract

Synthetic data generation methods are used to transform the original data into privacy-compliant synthetic copies (twin data) that can be used for training data, open-access data, internal datasets to speed up analyses and much more. With our proposed approach, synthetic data can be simulated in the same size as the input data or in any size, and in the case of finite populations even the entire population. Synthetic populations can even be used as input for microsimulation models and for testing complex sampling designs. This work aims to show a new and powerful synthetic data generation method in combination with an improved calibration method to adjust the synthetic data to known population margins. The proposed XGBoost-based method is compared with known model-based approaches for generating synthetic data using a complex survey data set, namely the European Union Statistics on Income and Living Conditions (EU-SILC). The XGBoost method shows strong performance especially with synthetic categorical variables and outperforms other tested methods. Moreover, the structure and the relationship between variables are well preserved. The tuning of the parameters - an important step in the application of XGBoost - was done by a modified k-fold cross-validation. After the data generation it is recommended to adjust the synthetic data to known population margins. For this purpose, we implemented a simulated annealing algorithm which is capable of using multiple different population margins at once. The algorithm is thus able to calibrate simulated population data containing information about persons in households. In addition, the algorithm is efficiently implemented making it feasible the adjust populations containing 100 Million and more people.

References

No References available

The unexpected value of R in official statistics

Authors

- Edwin de Jonge (Statistics Netherlands (CBS))

Abstract

R is an incredible statistical environment, in which many statistical procedures are available and can be readily used.It is predominantly used in the statistical and data science community and has an increasing important role in the creation of Official Statistics. Although engineered for statistical analysis, R plays in many offices also an auxiliary role in the production, visualization and publishing part of official statistics. The presentation addresses several expected and unexpected usages of R in the production process of statistical offices and sketches possible directions for statistical offices for using R.

References

No References available

thestats: An R package for gathering and

analyzing Turkish higher education statistics

Authors

- Olgun Aydin (Gdansk University of Technology, Poland)
- Mustafa Cavus (Warsaw University of Technology, Poland Eskisehir Technical University, Turkey)

Abstract

Open data has become an essential contributor to scientific studies recently. It provides transparency on the results of the studies from the reproducible research point of view. It is possible to find open datasets regarding official statistics, finance, and many others. These datasets are published mainly by statistical institutions, central banks, governmental bodies, and third-party corporations. In Turkey, official statistics are mainly published by the Turkish Statistical Institution, Central Bank, and other governmental agencies. Thanks to this, researchers can reach out to various data points regarding the economy, finance, higher education, and various domains. For example, the Turkish Higher Education Council provides a web portal for higher education statistics in Turkey. On the portal, detailed statistics regarding universities, faculties and departments can be found. As students' mobility, preferences, and the participation of graduates in the workforce significantly affect both global and regional economies in Turkey, data provided through the portal would most likely help researchers investigate the effect of higher education on the economy. Even though this portal provides very useful information, it does not offer any possibilities to download or query the data in a comfortable way. This study introduces a user-friendly R data package, thestats, developed to make the higher education statistics accessible in an easy way. The package allows researchers to query the data which is scraped from the portal and . The data is scraped annually using the power of rvest package, as the portal's data is updated yearly. After scraping data from the portal, thanks to data manipulation scripts, datais cleaned, and R data files are created using usethis, which is an R package for handling creation of R data _les and wrapping them into an R package. Thanks to this, there is no additional work that needs to be done by users. It is enough for the users to install thestats and start exploring the Turkish Higher Education Statistics using functions provided by the package.

References

No References available

Using R in Jupyter for statistical production

Authors

- Susie Jentoft (Statistics Norway)

Abstract

During the past few years, Statistics Norway has continued to increase its use of open source software, including R. While RStudio is used as the development environment for most R programming situations, Jupyter is gaining speed towards being the default tool for running statistical production processes. In this presentation, I provide a case study; an example of how Statistics Norway uses Jupyter for drawing a sample for the Labour Force Survey. I present our experiences of using R in Jupyter in a production environment, with its strengths of ease and user-friendliness, and having documentation close to where the code is run. Difficulties we have experienced include the lack of oversight on data and objects in the environment, and lack of tools for developing packages.

References

No References available

Using R to determine the impact of an image on viewers

Authors

- Nicolae-Marius Jula (University of Bucharest, Faculty of Business and Administration)

Abstract

In this paper, we are addressing some aspects regarding the likeability of an image. Besides the obvious answers as the quality, resolution, size, we suggest that there are also other properties that make people like and remember an image. This study answers the questions like: what is the best time frame to post on social media for the highest impact, is the dominant color, main subject, or the number of subjects having an impact on clients/followers? We are answering these questions using AI packages like YOLO for classification and object detection, countcolors for finding pixels by color range, and imager for an array of functions for working with image data.

References

No References available

Using the R Programming System in Sample Surveys

Authors

- Mubariz Nuriyev (The State Statistical Committee of the Republic of Azerbaijan, Center of Scientific Research and Statistical Innovation)
- Saleh Movlamov (The State Statistical Committee of the Republic of Azerbaijan, Center of Scientific Research and Statistical Innovation)
- Zarifa Naghiyeva (The State Statistical Committee of the Republic of Azerbaijan, Center of Scientific Research and Statistical Innovation)

Abstract

The article discusses the advantages of the R program for statistical surveys and the evaluation of observational data, and also examines the issues of sample design technology. With R codes were developed the application to sample households. At the same time, are analyzed various aspects of typical and non-non-typical applications. Has been calculated descriptive statistics for estimating household consumption expenditures: means; variation; standard deviation; standard error of the mean; the coefficient of variation; an application for calculating limit of error of means. Are presented the algorithm for development the application and hierarchical scheme for data generalization. To reduce the standard error of the mean, the data were processed on a territorial basis.

References

No References available