Use of R in Official Statistics

7th International Conference 2019
Welcome

The global community of R users is growing, and the number of National and International Statistical Offices that are adopting R is growing as well. About six years ago, when this conference was organized as an international conference for the first time in Romania, we felt a bit like outlaws using Free and Open Source Software (FOSS) in an area where commercial packages rule the land. How times have changed: in the mean time FOSS, and in particular R is considered a driving force of innovation in academia, industry and government. The popularity of R is demonstrated by the hundreds of local R user groups, the thousands of R packages, and the RConsortium.

For the second time this year we have a full day of tutorials with topics including sample stratification, mapping data, building and maintaining R packages, data cleaning and processing. Make sure to take full advantage of the experts that came here to share their knowledge.

With about fifty contributed talks and around one hundred conference attendees, this uRos is the largest in its history, in Romania. We are grateful to the speakers, tutorial organizers and attendees for making this conference such a growing success.

We wish you an awesome conference. Welcome to uRos2019!
Organizing partners

- Statistics Romania
- Ecological University of Bucharest

Co-Organizing partners

- Statistics Netherlands
- Statistics Austria

Special Journal Issues

Romanian Statistical Review: http://www.revistadestatistica.ro/
Austrian Journal of Statistics: https://www.ajs.or.at/
Contents

Welcome iii

Program overview 1

Session overview 5

Tutorials 13

Integration of data sources in R through statistical matching (Marcello D'Orazio) 14
Mapping a Table of Data with Esri Shapefiles in R (Mervyn O’Luing) ........ 15
Modern package development in R (Bernhard Meindl) .................... 16
Statistical data cleaning for official statistics with R (Mark van der Loo and
Edwin de Jonge) ................................................................. 17
Use of R package SamplingStrata for the Optimal Stratification of Sampling
Frames for Multipurpose Sampling Surveys (Marco Ballin and Giulio
Barcaroli) ................................................................. 18

Keynotes 19

A missing value tour in R (Julie Josse) ............................................ 20
R at the Italian National Institute of Statistics (Istat): a twelve years story
(Giulio Barcaroli) ................................................................. 21

Conference presentations 23

uRos2019 v
A comparison between R, Julia and Python in building and prediction phases for high blood pressure in personalized medicine (Antonio Clim, Razvan Daniel Zota and Bogdan Oancea) .......................................................... 24

A generic shiny/js dashboard for data validation results (Olav ten Bosch and Mark van der Loo) .......................................................... 25

An R package for calculating repeat-sale price indices (Megan Kirby-McGregor and Steve Martin) .......................................................... 26

COMUNIKOS: COMMunicating UNcertainty In Key Official Sta-
tics (Dario Buono, Gian Luigi Mazzi, Massimiliano Marcellini, George Kapetanio-
s, James Mitchell, Tiziana Laureti, Edwin de Jonge, Florabela Car-
rausu and Ibtissam Sahir) .......................................................... 27

Data cleaning with R put into production (Aksel Thomsen) ............... 28

Detection of Business Cycle Turning Points in the Serbian Economic Activity (Isidora Jovandic, Petar Korovic and Katarina Stantic) .......... 29

From SPSS to R designing a scalable R infrastructure to support health anal-
lytics in Scotland (Terry McLaughlin) .......................................................... 30

Generalization and robustification of a ratio model (Kazumi Wada, Kei-
ichiro Sakashita and Hiroe Tsubaki) .......................................................... 31

Google Analytics, Shiny apps and EU/UK law (Joseph Adams) .......... 32

Headless Chrome Automation with the ‘crrri’ package (Romain Lesur) ... 33

Implementation of the cell-key method based random noise approach in R (Tobias Enderle and Bernhard Meindl) .......................................................... 34

Improvement of the reliability score for autocoding and its implementa-
tion in R (Yukako Toko, Shinya Iijima and Mika Sato-Ilic) ........................ 35

Insights on replicating research results in remote sensing data statistical analysis using R (Marian Necula and Mihaela E Paun) ................. 36

Integration of Survey Data in R Based on Machine Learning (Spaziani Mattia, Frattarola Dorian and D’Orazio Marcello) .......................................................... 37

Just BEAT it. Bethel Extended Allocation for Two-stage sampling package (Stefano Falorsi, Andrea Fasulo, Alessio Guandalini, Daniela Pagliuca and Marco D. Terribili) .......................................................... 38

Multivariate imputation by chained equation with packages ‘mice’ and ‘mi’ to impute for missing values in semi-continuous data: methods to properly address non-response bias (Sian Lloyd) .......................................................... 40

R tools for ILOSTAT: Rilostat and SMART (M. Villarreal-Fuentes and S. Ding) .......................................................... 41
Real time monitoring and reporting surveys using Shiny App on Survey Solutions API with R (Michael Wild, Antoniade-Ciprian Alexandru and Nicoleta Caragea) ................................................................. 42
Religiosity and Economic Growth (Serban Olah and Gabriel Roseanu) .... 43
Remittances' impact on receiving country. Multidimensional analysis at macro and microeconomic level using R. (Valentina Vasile, Elena Bunde- duchi, Stefan Daniel and Calin-Adrian Comes) ................................. 44
Reproducible Corporate Publications using pagedown (Romain Lesur) .... 45
Rethinking Human Well-Being: A holistic Flourishing perspective (Josue Aduna Castillo) ................................................................. 46
rtempo - TEMPO Online data interface to R (Marian Necula, Ana Maria Tiru and Bogdan Oancea) ................................................................. 47
Selective editing with categorical variables (M. Rosario Gonzalez-Garcia, Jara Poch, David Salgado and Teresa Vazquez-Gutierrez) .............. 48
Semiautomatic Reports with R Markdown (Max Grutter) ...................... 49
SMART: A web system for Small Area Estimation (S. Falorsi, A. Fasulo and F. Solari) ................................................................. 50
The ability of different imputation methods to capture complex depend- encies in high dimensions (Humera Razzak and Christian Heumann) ... 52
Trial of R training for technical supports to third countries (Kiyomi Shi- rakawa and Ken Nakamatsu) ............................................................... 53
Use of R in outliers detection (Audrius Taraila) .................................... 54
Use of R to increase the Automatic Coding Rate in Census2021. Statistics Portugal Challenge (Rui Alves and Almiro Moreira) ................. 55
Using R and RShiny to produce official statistics on Psychiatric Inpatient Care in Scotland - NHS National Services Scotland (Nikos Alexandrou and Ciara Gribben) ................................................................. 57
Utility evaluation of global recoding by accuracy of discrimination model (Natsuki Sano) ................................................................. 59
Utilization of big data for improving Consumption Trend Index - Estimation of the number of person per household based on the characteris- tic of purchase items- (Anri Mutoh, Masayo Yamashita, Yoshiyasu Tamura and Masahiro Matsumoto) ................................................................. 60
Web Scraping Notices of Insolvency Proceedings - Using publicly available data to enhance survey response quality (Joerg Feuerhake) .... 61
Program overview
## BRIEF AGENDA

**May 20, Monday**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 - 09:30</td>
<td>Registration of participants</td>
</tr>
<tr>
<td>09:30 - 10:00</td>
<td>Welcome message</td>
</tr>
<tr>
<td>10:00 - 11:00</td>
<td>Keynote speaker 1</td>
</tr>
<tr>
<td>11:00 - 11:30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>11:30 - 13:00</td>
<td>Parallel tutorial sessions I</td>
</tr>
<tr>
<td>13:00 - 14:00</td>
<td>Lunch break and Group photo</td>
</tr>
<tr>
<td>14:00 - 15:30</td>
<td>Parallel tutorial sessions II</td>
</tr>
<tr>
<td>15:30 - 16:00</td>
<td>Coffee break</td>
</tr>
<tr>
<td>16:00 - 17:30</td>
<td>Parallel Tutorial and Scientific Sessions</td>
</tr>
<tr>
<td>17:30 - 17:40</td>
<td>Go to Conference Room for a Project</td>
</tr>
<tr>
<td>17:40 - 18:00</td>
<td>Presentation</td>
</tr>
<tr>
<td>19:00</td>
<td>Official dinner</td>
</tr>
</tbody>
</table>

**May 21, Tuesday**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 - 09:00</td>
<td>Walk in</td>
</tr>
<tr>
<td>09:00 - 10:00</td>
<td>Keynote speaker 2</td>
</tr>
<tr>
<td>10:00 - 10:30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10:30 - 12:30</td>
<td>Scientific session I (parallel sessions)</td>
</tr>
<tr>
<td>12:30 - 13:30</td>
<td>Lunch break</td>
</tr>
<tr>
<td>13:30 - 15:00</td>
<td>Scientific session II (parallel sessions)</td>
</tr>
<tr>
<td>15:00 - 15:30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>15:30 - 17:10</td>
<td>Scientific session III (parallel sessions)</td>
</tr>
<tr>
<td>17:10 - 17:20</td>
<td>Go to Conference Room for Closing remarks</td>
</tr>
<tr>
<td>17:20 - 17:45</td>
<td>Closing remarks</td>
</tr>
</tbody>
</table>
SATELLITE EVENT –
unconfUROS hackathon
Conference Room of National Institute of Statistics, Blue Hall, Ground Floor

May 22, Wednesday
08:30 - 09:00 Welcoming participants
09:00 Opening, who is who?; Work starts.
12:00 - 13:00 Lunch break
19:30 End of day/Discussions

unConf implemented ideas will be promoted via conference website, GitHub and social media.

CONFERENCE VENUE

National Institute of Statistics, Bucharest, Romania
Address of National Institute of Statistics:
No.16 Libertatii Bvd., District 5, Bucharest, Romania
Session overview
SCIENTIFIC BOARD

Tudorel ANDREI (President, National Institute of Statistics - Romania)
Matthias TEMPL (ZHAW school of engineering, Switzerland)
Mark VAN DER LOO (Statistics Netherlands)
Alexander KOWARIK (Statistics Austria)
Bogdan OANCEA (National Institute of Statistics, University of Bucharest, Romania)
Adrian DUSA (University of Bucharest, Romania)
Elena DRUICA (University of Bucharest, Romania)
Kazumi WADA (National Statistics Center, Japan)
Marcello D’ORAZIO (FAO, Rome, Italy)
Valentin TODOROV (UNIDO, Vienna Austria)
Nicoleta CARAGEA (National Institute of Statistics, Ecological University of Bucharest, Romania)
Ciprian ALEXANDRU (National Institute of Statistics, Ecological University of Bucharest, Romania)
Alina MATEI (University of Neuchatel, Switzerland)
Bernhard MEINDL (Statistics Austria, Austria)
Claudiu HERTELIU (Bucharest University of Economic Studies, EMOS Board, Romania)
Ana-Maria CIUHU (National Institute of Statistics & Institute for National Economy, Romania)
Edwin DE JONGE (Statistics Netherlands, The Netherlands)
Gergeley DARÓCZI (System1, Hungary)
Kamarul Anjifin MANSOR (MARA University of Technology, Malaysia)
Marius Nicolae IULA (University of Bucharest, Romania)
Matyas MESZAROS (Eurostat, Luxemburg)
Mihaela PALIN (University of Bucharest)
Roxana ADAM (University of Bucharest & National Institute of Statistics, Romania)

ORGANIZING COMMITTEE

Nicoleta CARAGEA, National Institute of Statistics, Romania
Daniela Elena STEFANESCU, National Institute of Statistics, Romania
Viorela SALAGEAN, National Institute of Statistics, Romania
Bogdan Liviu PATARLAGEANU, National Institute of Statistics, Romania
Vitty Cristian CHIRAN, National Institute of Statistics, Romania
Mariana VASILE, National Institute of Statistics, Romania
Nicoleta JALBA, National Institute of Statistics, Romania

SECRETARIATE

Ana-Maria CIUHU, National Institute of Statistics
Roxana ADAM, National Institute of Statistics
Roxana GLĂVAN, National Institute of Statistics
Diana POPA, National Institute of Statistics
PROGRAMME
Conference Room and Blue Hall, National Institute of Statistics, Ground Floor

May 20, Monday

0830 – 0930 Registration of participants

0930 – 1000 Welcome message
Tudorel ANDREI, President of National Institute of Statistics, Romania
Nicoleta CARAGEA, President of uRos, National Institute of Statistics/Ecological University of Bucharest, Romania

1000 – 1100 Keynote Speaker 1: Julie JOSSE, Member of RFoundation and Rforwards,
Professor of Statistics at Ecole Polytechnique, France
Chairman – Mark VAN DER LOO, Statistics Netherlands (CBS)
Presentation - A missing value tour in R

1100 – 1130 Coffee break

1130 – 1300 Parallel Tutorial Sessions (I)

Conference Room, Ground Floor
Statistical data cleaning for official statistics with R
Mark VAN DER LOO and Edwin DE JONGE, Statistics Netherlands (CBS)

Blue Hall, Ground Floor
Use of R package SamplingStrata for the Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys
Marco BALLIN and Giulio BARCAROLI, Italian National Institute of Statistics (Istat)

1300 – 1400 Lunch Break and Group Photo

1400 – 1530 Parallel Tutorial Sessions (II)

Conference Room, Ground Floor
Modern package development in R
Alexander KOWARIK and Bernhard MEINDL, Statistics Austria

Blue Hall, Ground Floor
Mapping a Table of Data with Esri Shapefiles in R
Mervyn ÖLÜNING, Central Statistics Office, Ireland

1500 – 1600 Coffee break

1600 – 1730 Parallel Tutorial and Scientific Sessions

Conference Room, Ground Floor
Integration of data sources in R through statistical matching
Marcello D’ORAZIO - Food and Agriculture Organization of The United Nations
Blue Hall, Ground Floor

Scientific Session Methods for Official Statistics (I)
Chairman: Bogdan OANCEA, National Institute of Statistics, University of Bucharest, Romania

Implementation of the cell-key method based random noise approach in R
Tobias ENDERLE, DESTATIS, Germany
Bernhard MEINDL, Statistics Austria, Austria

Utility evaluation of global recoding by accuracy of discrimination model
Natsuki SANO, Statistical Research and Training Institute Ministry of Internal Affairs and Communication, Japan

Headless Chrome Automation with the criri package
Romain LESUR, Ministère de la Justice, France

From SPSS to R: designing a scalable R infrastructure to support health analytics in Scotland
Terry MCCLAUGHLIN, NHS National Services Scotland, Scotland

17:30 – 17:45 Go to Conference Room for a Project Presentation

17:45 – 19:00 Conference Room, Ground Floor
Project Presentation
Chairman: Bernhard Meindl, Statistics Austria

COMUNIKOS: Communicating Uncertainty In Key Official Statistics
Dario BUONO, Eurostat/GOPA
Gian Luigi MAZZI, EuroStat/GOPA
Massimiliano MARCELLINI, Eurostat/GOPA
George KAPETANIOS, Eurostat/GOPA
James MITCHELL, Eurostat/GOPA
Tiziana LAURETI, Eurostat/GOPA
Edwin DE JONGE, Eurostat/GOPA
Florabela CARAUSU, Eurostat/GOPA
Ibtissam SAHIR, Eurostat/GOPA

19:00 Conference Dinner
May 21, Tuesday

8:30 – 9:00 Walk In

9:00 – 10:00 Keynote Speaker 2: Giulio BARCAROLI, Italian National Institute of Statistics (Istat), Italy
   Chairman: Alexander KOWARIK, Statistics Austria
   Presentation - R at the Italian National Institute of Statistics (Istat): a twelve years story

10:00 – 10:30 Coffee break

10:30 – 12:30 Conference Room, Ground Floor
   Scientific Session Data cleaning, Sampling and estimation
   Chairman: Mark van der Loo, Statistics Netherlands (CBS)

   Just BEAT it. Bethel Extended Allocation for Two-stage sampling package
   Stefano FALCRISI, Istat, Italy
   Andrea FASULO, Istat, Italy
   Alessio GUANDALINI, Istat, Italy
   Daniela PAGLIUCA, Istat, Italy
   Marco D. TERRIBILI, Istat, Italy
   A generic shiny/js dashboard for data validation results
   Olov ten BOSCH, Statistics Netherlands, Netherlands
   Mark van der LOO, Statistics Netherlands, Netherlands

   Integration of Survey Data in R Based on Machine Learning
   Mattia SPAZIANI, Istat, Italy
   Doriana FRATTAROLA, Istat, Italy
   Marcello D’ORAZIO, Istat, Italy

   Selective editing with categorical variables
   M. Rosano GONZÁLEZ-GARCÍA, S.G. for Sectorial Social Statistics, Statistics Spain (INE), Spain
   Jara POCH, S.G. for Information and Communication Technologies, Statistics Spain (INE), Spain

   Trial of R training for technical supports to third countries
   Kiyomi SHIRAKAWA, Hitotsubashi University, Japan
   Ken NAKAMATSU, Takumi Information Technology, Japan

   Data cleaning with R put into production
   Aksel THOMSEN, Statistics Denmark, Denmark
10:30 – 12:30  Blue Hall, Ground Floor

Scientific Session R in Production: Data analysis
Chairman: Ana-Maria CIUHU, National Institute of Statistics, Romania

An R package for calculating repeat-sale price indices
Megan KIRBY-MCGREGOR, Producer Prices Division, Statistics Canada, Canada
Steve MARTIN, Producer Prices Division, Statistics Canada, Canada

The Ability Of Different Imputation Methods To Capture Complex Dependencies In High Dimensions
Humera RAZZAK, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany
Christian HEUMANN, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany

Detection of Business Cycle Turning Points in the Serbian Economic Activity
Isidora Jovandić, Statistical Office of the Republic of Serbia, Serbia
Petar Korović, Statistical Office of the Republic of Serbia, Serbia
Katarina Stanić, Statistical Office of the Republic of Serbia, Serbia

Religiosity and Economic Growth
Serban OLAH, University of Oradea, Romania
Gabriel ROSEANU, University of Oradea, Romania

A comparison between R, Julia and Python in building and prediction phases for high blood pressure in personalized medicine
Antonio CLIM, The Bucharest University of Economic Studies, Romania
Răzvan Daniel ZOTA, The Bucharest University of Economic Studies, Romania
Bogdan DANCEA, University of Bucharest, Romania

Remittances’ impact on receiving country. Multidimensional analysis at macro and microeconomic level using R.
Valentina VASILE, Institute of National Economy, Romanian Academy
Elena BUNUICHI, University of Medicine, Pharmacy, Sciences and Technology of Târgu Mures, Romania
Stefan DANIEL, University of Medicine, Pharmacy, Sciences and Technology of Târgu Mures, Romania
Calin-Adrian COMES, University of Medicine, Pharmacy, Sciences and Technology of Târgu Mures, Romania

12:30 – 13:30  Lunch Break

13:30 – 15:00  Conference Room, Ground Floor

Scientific Session Methods for Official Statistics (II)
Chairman: Alexander KOWARIK, Statistics Austria

SMART: A web system for Small Area Estimation
Stefano FALCORSI, Istat, Italy
Andrea FASULO, Istat, Italy
F SOLARI, Istat, Italy
Multivariate imputation by chained equation with packages ‘mice’ and ‘mi’ to impute for missing values in semi-continuous data: methods to properly address non-response bias
Sian LLOYD, Office for National Statistics, United Kingdom

Generalization and robustification of a ratio model
Kazumi WADA, National Statistics Center, Japan
Keichiro SAKASHITA, National Statistics Center, Japan
Hiroe TSUBAKI, National Statistics Center, Japan

Insights On Replicating Research Results in Remote Sensing Data Statistical Analysis Using R
Marian NECULA, The Bucharest University of Economic Studies, Romania
Mihaela PAUN, University of Bucharest, Faculty of Administration and Business, Romania

13:00 – 15:00  
**Blue Hall, Ground Floor**

*Scientific Session R in Production: Automation*
**Chairman:** Edwin DE JONGE, Statistics Netherlands (CBS)

Rethinking Human Well-Being: A holistic Flourishing perspective
Josue ADUNA, 3S Studio, Romania

**Improvement Of The Reliability Score For Autocoding And Its Implementation In R**
Yukako TOKO, National Statistics Center, Japan
Shinya IJIMA, National Statistics Center, Japan
Mike SATO-IIC, Faculty of Engineering, Information and Systems, University, Japan

**Web Scraping Notices Of Insolvency Proceedings - Using Publicly Available Data To Enhance Survey Response Quality**
Joerg FEUERHAKE, DESTATIS, Germany

**Use Of R To Increase The Automatic Coding Rate In Census2021. Statistics Portugal Challenge**
Rui ALVES, Statistics Portugal, Portugal
Almirio MOREIRA, Statistics Portugal, Portugal

15:00 – 15:30  
**Coffee Break**

15:30 – 17:00  
**Conference Room, Ground Floor**

*Scientific Session Dissemination and visualizations*
**Chairman:** Ciprian ALEXANDRU, National Institute of Statistics/Ecological University of Bucharest, Romania

**R Tools for ILOSTAT: Rilostat and SMART**
Mabelin VILLARRREAL-FUENTES, International Labour Organization (ILO), Colombia
Ding SHUTONG, International Labour Organization (ILO), Colombia

**Semi-automatic Reports with R Markdown**
Max GRÜTTER, Statistical Office of the Canton of Zurich, Switzerland
rtempo - TEMPO Online data interface to R  
Marian NECULA, The Bucharest University of Economic Studies, Romania  
Ana Maria ŢÎRU, National Institute of Statistics, Romania  
Bogdan OANCEA, University of Bucharest, Romania

Reproducible Corporate Publications using pagedown  
Romain LESUR, Ministère de la Justice, France

1530 – 1730 Blue Hall, Ground Floor  
Scientific Session Shiny apps and Big Data  
Chairman: Marcello D’ORAZIO - Food and Agriculture Organization of The United Nations  
Using R and RShiny to produce official statistics on Psychiatric inpatient Care in Scotland  
Nikos ALEXANDROU, NHS National Services Scotland, Scotland  
Clara GRIBBEN, NHS National Services Scotland, Scotland

Google Analytics, Shiny apps and EU/UK law  
Joseph Adams, National Records of Scotland, Scotland

Use of R in outliers detection  
Audrius TARAILE, Statistics Lithuania, Lithuania

Utilization of big data for improving Consumption Trend Index - Estimation of the number of person per household based on the characteristic of purchase items-  
Anri MUTOH, National Statistics Center, Japan  
Masayo YAMASHITA, National Statistics Center, Japan  
Yoshiyasu TAMURA, National Statistics Center, Japan  
Masahiro MATSUMOTO, National Statistics Center, Japan

Real time monitoring and reporting surveys using Shiny App on Survey Solutions API with R  
Michael WILD, World Bank, USA

Nicoleta CARAGEA, Ecological University of Bucharest/National Institute of Statistics, Romania  
Ciprian ALEXANDRUD, Ecological University of Bucharest/National Institute of Statistics, Romania

1719 - 1720 Go to Conference Room for Closing remarks

1720 - 1745 Closing Remarks

Contact  
Address of National Institute of Statistics:  
No.16 Libertatii Blvd., District 5, Bucharest, Romania  
Telephone center: +4021 3181824; +4021 3181842  
Fax: +4021 3124875; +4021 3181851; +4021 3181873  
e-mail: nicoleta.caragea [at] insse.ro
Tutorials
Integration of data sources in R through statistical matching

By Marcello D'Orazio

Food and Agriculture Organization (FAO) of the UN and Italian national Institute of Statistics (Istat)

Statistical matching (aka data fusion) consists in a set of techniques developed to integrate distinct data sources referred to the same target population (typically stemming out from sample surveys) to get insights on the relationship between variables not jointly observed in a single data source. The tutorial will give an idea of the standard setting of statistical matching and the corresponding underlying assumptions; it will show how to perform matching in R by means of facilities provided by the R package StatMatch, including analyses that should precede and follow the matching step.
Mapping a Table of Data with Esri Shapefiles in R

By Mervyn O’Luing

Central Statistics Office, Ireland

This is an introductory tutorial to mapping opensource data with two shapefiles, plot multiple time series maps and produce web based maps using the ‘ggplot2’ and ‘tmap’ R packages in R. Users will learn how to:

- create a CSV file with GDP data extracted from an official Tanzanian publication,
- create an R data-frame from this data,
- add a geospatial reference to this data-frame,
- download two Esri shapefiles, regions and water bodies, from the Tanzanian National Bureau of Statistics' (TNBS) website,
- create geospatial data-frames from the Esri shapefiles,
- join the region geospatial data-frame to the GDP data-frame,
- plot the data using the ‘ggplot2’ package, and overlay the region and water bodies geospatial data frames,
- plot multiple time series maps and produce interactive zoomable web based maps using the ‘tmap' package.
Modern package development in R

By Bernhard Meindl

Statistics Austria

In this tutorial, the aim is to learn about several tools that help you to improve building and maintaining R packages. Participants will learn about the following topics:

- setup continuous integration using travis
- perform automatic code and style checking
- improve unit tests in your package
- create better documentation
- automatically create nice looking documentation websites

This tutorial will be "hands-on" so it is advised to bring your notebooks with you.
Statistical data cleaning for official statistics with R

By Mark van der Loo and Edwin de Jonge

Statistics Netherlands, Statistics Netherlands

This tutorial will demonstrate how data quality can be systematically defined and improved using R. It will focus on data validation (data checking), locating errors, and imputing missing or erroneous values under restrictions. Examples will be drawn from the Structural Business Survey (SBS) where common restrictions include non-negativity rules and record-wise balance checks. I will present short introduction to the main principles, provide quizzes and discussions for the audience, and give short R-based exercises.

A number of R packages related to data validation will be demonstrated, including 'validate' (for data quality checks), 'errorlocate' (for error localization), 'simputation' (for imputation methods), 'rspa' (for value adjustment), and 'lumberjack' (for keeping track of changes in data). Special attention will be paid on how to combine the various data processing steps, and how to analyse and visualize the results. At the end of the course, participants will have insight into some of the methods common in data editing for business surveys as well as an overview of how to implement that with free and open source R and the mentioned packages.
Use of R package SamplingStrata for the Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys

By Marco Ballin and Giulio Barcaroli

Italian National Statistical Institute (ISTAT), Italy

The aim of this tutorial is to enable the participants to learn how to use the R package "SamplingStrata" in order to optimize the design of stratified samples. The package offers an approach for the determination of the best stratification of a sampling frame, the one that ensures the minimum sample cost under the condition to satisfy precision constraints in a multivariate and multi-domain case. This approach is based on the use of the genetic algorithm: each solution (i.e. a particular partition in strata of the sampling frame) is considered as an individual in a population; the fitness of all individuals is evaluated applying the Bethel algorithm to calculate the sampling size satisfying precision constraints on the target estimates. Functions in the package allow to: (a) prepare necessary inputs and check their validity; (b) perform the optimization step choosing the values of the most important parameters; (c) assign the optimized strata labels to the sampling frame; (d) select a sample from the new frame accordingly to the best allocation; (e) test the compliance of the design to precision constraints.
Keynotes
A missing value tour in R

By Julie Josse

Statistics at Ecole Polytechnique, France

In many application settings, the data have missing features which make data analysis challenging. An abundant literature addresses missing data as well as more than 150 R packages. Funded by the R consortium, we have created the R-miss-tastic platform along with a dedicated task view which aims at giving an overview of main references, contributors, tutorials to offer users keys to analyse their data. This platform highlights that this is an active field of work and that as usual different problems requires designing dedicated methods.

In this presentation, I will share my experience on the topic. I will start by the inferential framework, where the aim is to estimate at best the parameters and their variance in the presence of missing data. Last multiple imputation methods have focused on taking into account the heterogeneity of the data (multi-sources with variables of different natures, etc.). Then I will present recent results in a supervised-learning setting. A striking one is that the widely-used method of imputing with the mean prior to learning can be consistent. That such a simple approach can be relevant may have important consequences in practice.
R at the Italian National Institute of Statistics (Istat): a twelve years story

By Giulio Barcaroli

Italian National Institute of Statistics (Istat)

It was in 2006-07 that a small group of people inside Istat began to consider the adoption of R as a valid alternative to the use of proprietary software like SAS. There were many reasons for that: R was free and open, and was on the research edge, while SAS was expensive and without the same research background. But the odds were not negligible: SAS was deep inside the production processes, hundreds of people had been trained and used SAS in a systematic way, the assistance levels seemed not to be comparable. Notwithstanding this, also because of the general movement towards open source software, Istat decided to at least diminish its dependence on proprietary solutions (in particular SAS-based). Two initiatives were launched: migration of generalized software systems using open technologies (and R was on the top of these), and mass training on R. Some adhered enthusiastically, some other opposed strong resistance. Nowadays R and SAS coexist, with some important advancements: a relevant number of researchers make use of R in daily activities, the great majority of generalized solutions are R-based, and an important project aiming at building a generalized architecture for statistical production will adopt services many of which are functions of R libraries. The overall balance can be said to be positive, and to report this experience, including concrete examples of how it evolved over time, can be useful to other OS Institutes that undergo or could undergo a similar transition.
Conference presentations
A comparison between R, Julia and Python in building and prediction phases for high blood pressure in personalized medicine

By Antonio Clim, Razvan Daniel Zota and Bogdan Oancea

The Bucharest University of Economic Studies, The Bucharest University of Economic Studies, University of Bucharest

The starting point of our research was to propose a new approach in smart applications for patients’ remote monitoring model comprising blood pressure and heart rate measurement. We propose a different perspective in this area since a possible usage of R, Julia, or Python hasn’t been considered in remote monitoring of patients in personalized medicine domain. While Julia could sometimes outperform Python and R in terms of flexibility, ease of implementation, speed of execution, and machine-friendly approach, we showed that under certain conditions regarding the computing environment, R could attain the same level of performance as the other two languages under consideration. Using a set of standardized algorithms used in patient monitoring applications, we showed that R could be considered as a strong candidate for developing new applications in this area.
A generic shiny/js dashboard for data validation results

By Olav ten Bosch and Mark van der Loo

Statistics Netherlands, Statistics Netherlands

Monitoring data quality is essential for the whole statistical production chain. From raw input data to data delivery, at every step we need to know whether data is fit for purpose. In previous uRos conferences we presented the validation package R-validate [1-3], which is designed specifically for this purpose. In the scope of the ESSnet project Validat Integration [4] a generic report structure was defined [5] that can be used to describe any validation result of any validation tool in any validation language in a machine-readable way. Validation results generated by R-validate can now be generated in this JSON based generic message format using the new package validatereport [6].

Although these two packages already form a powerful and rather complete base layer to build validation processes, we believe that it is useful to also add a more interactive and visual way of interaction. In particular, we think it is useful to offer a way to view the (selected) validation results (both passes as well as fails) in a visual representative of the data being validated. For this we built an experimental validation dashboard.

Because the number of validation results may become huge this dashboard uses the efficient javascript library crossfilter for fast n-dimensional filtering and grouping of validation results. In addition we used d3.js [7] and dc.js [8] for graphical filtering of results. The HTML widgets library was used to integrate the result in a shiny dashboard, starting from the uRos2018 Unconf project validaty [9], where it works together with R-validate and R-validatereport. In this presentation we will show the first results, the challenges and the plans for extensions in future.
An R package for calculating repeat-sale price indices

By Megan Kirby-McGregor and Steve Martin

Statistics Canada, Statistics Canada

Residential property price indices are valuable indicators of economic activity and financial stability. Calculating pure, quality adjusted price changes using real estate data presents a number of obstacles, however, as properties each have a unique set of price determining characteristics. The repeat-sales method offers a means to construct a constant-quality price index by exploiting multiple sales for the same property over time, although calculating reproducible results on an on-going basis following this methodology is not trivial. This paper outlines an R package developed at Statistics Canada for calculating a variety of repeat-sale price indices found in the literature and used in practice, as well as supporting tools to get price data into a form that is suitable for calculating an index. The package is designed to be simple, easy to use, and suitable for operational use, without requiring much prior knowledge of the repeat-sales approach or R.
COMUNIKOS: COMmunica\textsubscript{NG} UNcertainty In Key Official Statistics

By Dario Buono, Gian Luigi Mazzi, Massimiliano Marcellini, George Kapetanios, James Mitchell, Tiziana Laureti, Edwin de Jonge, Florabela Carausu and Ibtissam Sahir

Eurostat/GOPA, Eurostat/GOPA

Traditionally, official statistics are published as point estimates of a given variable. Publishing a single number simplifies the communication to users and avoids possible confusion or misinterpretation. Incidentally statistical offices publish information about uncertainty often in textual form. More recently, statistical production methods that integrate new data sources (e.g. administrative registers, Big Data) create additional sources for possible errors and uncertainties, which are not measured and communicated to the users. The COMUNIKOS project (COMmunica\textsubscript{NG} UNcertainty In Key Official Statistics) is an Eurostat research initiative implemented with the support of the external contractor GOPA (GOPA Worldwide Consultants in joint venture with GOPA Luxembourg). The project aims to propose methods / metrics to present uncertainties in key official statistics, starting from a comprehensive review of the different sources, their impact on the disseminated data and their relevance for official statistics. A case study for measuring uncertainties in price statistics, following the standards production process and using scanned data, will conclude the research. For the implementation of the project GOPA combines expertise from researchers in the domain, statistical offices, evidence-based policy advisers, and consultants, under the close guidance and enhanced collaboration of Eurostat researchers. The project will be promoted and its results will be disseminated among the research community in official statistics along its implementation from January 2019 to June 2020.
Data cleaning with R put into production

By Aksel Thomsen

Denmark

In the last 4 years Statistics Denmark have put statistical data cleaning with R into production. We have integrated R into our self-developed Data Archive. The Data Archive is built on top of an Oracle database, and is our new home for micro data used in production - all the way from the first reporting by a statistical unit and until it is ready for publication.

This setup enables everyone to run R programs checking the validity of the reported data within the same GUI used for all manual processes in our production of data. Furthermore, we also achieve full logging of both changes in data and executions of the available R programs.

The execution of program results in both changes in data and a report. The validation reporting is done on three different levels:

1) An aggregated report on macro level for the statistic (based on Markdown!)
2) The overall quality of the individual statistical unit
3) Each potential validation error in the data

All this is done in R invisible to the user. Specialists in the Methodology office program all code regarding data cleaning. This ensures a common high quality and standardisation of the validation process across different statistics.
Detection of Business Cycle Turning Points in the Serbian Economic Activity

By Isidora Jovandic, Petar Korovic and Katarina Stancic


The business cycle (also called economic cycle or trade cycle) represents periodic but irregular movements in economic activity, measured by fluctuations in real GDP and other macroeconomic variables around their long-term trends. Having knowledge and understanding about the past, current and future position in the business cycle is a very valuable input for many different stakeholders, such as governments, other policy makers and businesses.

In our work, we have estimated a model for detecting past turning points using the Serbian GDP series and validated the findings by providing a first economic interpretation of the fluctuations. Based on that model, we have by tested several candidates for leading indicators for the business cycle movements and defined a simple initial model for the early detection of fluctuations. The whole investigation process was implemented in R and thoroughly documented, in a way that provides possibility to easily test, refresh or further enhance the model. This gives the Statistical Office of the Republic of Serbia an opportunity to develop the leading indicator further, by either adding time series, thus advancing from a simple to a composite indicator, or in other ways. Since being able to reliably anticipate future business cycle movements represents a huge aid in preparing and carrying out economic regulation actions successfully, the results of our investigation represent an aid in preparing the Serbian policy makers for the important decisions and challenges of the future.
From SPSS to R designing a scalable R infrastructure to support health analytics in Scotland

By Terry McLaughlin

NHS National Services Scotland

The Information Services Division (ISD) of NHS National Services Scotland provides health information, health intelligence, statistical services and advice that support the delivery of Healthcare and Social Care across Scotland. ISD employs over 300 analysts who, until very recently, have relied on IBM SPSS Statistics for the production of National Statistics publications and to deliver business-critical analytical functions. Recognising the need to streamline the production of our publications and improve how they are presented, and with a desire to work with modern data science tools, ISD's analysts have been the driving force in the adoption of R as the organisation's primary analytical tool.

This presentation will focus on some of the key challenges that were faced in designing and implementing a scalable R infrastructure for use by all analysts, and will describe the reasons why the organisation chose to build this infrastructure around RStudio Server Pro. We will also look at the tools and techniques that ISD are adopting for implementing Reproducible Analytical Pipelines, including package management, version control and code sharing, along with how guidance and training is being used to educate and encourage best practice across ISD's analytical community.
Generalization and robustification of a ratio model

By Kazumi Wada, Keiichiro Sakashita and Hiroe Tsubaki

Japan, Japan

It is known that data such as business sales and household income need data transformation prior to regression estimate so that the data has a homoscedastic error. However, data transformations make the estimation of mean and total unstable. Therefore, statisticians, who work for official statistics are obliged to provide mean and total values, may choose the ratio estimator to avoid transformation.

The ratio estimator is based on the ratio model, \( y_i = \beta x_i + e_i \), where \( i = 1, \ldots, n \) of \((x,y)\) is observed on each of the \( n \) units. It has a heteroscedastic error term \( e_i \sim N(0, [x^*\sigma]^2) \)

proportional to the variance of a single auxiliary variable \( x \). The ratio model is influenced by outliers as well as the regression model; however, the robustification is not straightforward because of the heteroscedastic error.

Our presentation explains how to modify the error term of the conventional ratio model to have a homoscedastic part, \( e_i \sim N(0, \sigma^2) \).

It enables to robustify the ratio model by means of M-estimation and the iterative reweighted least squares (IRLS) algorithm for computation. In addition, we also broadens the error term, \( e_i \sim N(0, [x^{<U + 0.3B3 >} \sigma]^2) \),

so that the model to obtain variation with regards to heteroscedasticity.

Monte Carlo simulations confirm the efficiency of the proposed generalized robust estimator, as well as their computational efficiencies. A comparison of scale parameters between the average absolute deviation (AAD) and median absolute deviation (MAD) is also made.

The proposed estimator will be illustrated using data of the 2016 Economic Census for Business Activity in Japan.
Google Analytics, Shiny apps and EU/UK law

By Joseph Adams

National Records of Scotland

Google Analytics (GA) is commonly used with Shiny apps to measure user-interactions. Since GA uses cookies to "remember" what a user has done, it is important to comply with relevant legislation when using this technology.

In this talk I will present our understanding of how the UK Privacy and Electronic Communications Regulations (PECR) apply to using cookies. I will also present practical information on complying with the PECR while using GA in a shiny app.
Headless Chrome Automation with the ‘crrri’ package

By Romain Lesur
Ministere de la Justice, France

Headless Chrome is a headless web browser that became extremely popular thanks to the node.js chrome-remote-interface and puppeteer libraries. A headless web browser can be used for different goals: web scrapping, testing web applications (e.g. for Shiny apps), screenshots or PDF generation of web pages.

Several R packages as RSelenium and webdriver offer high-level interfaces to headless web browsers.

In this communication, we will present a new package named crrri that provides a low-level interface to headless Chrome using the Chrome Developer Protocol (https://chromedevtools.github.io/devtools-protocol/).

It offers an access to the most advanced features of headless Chrome from R.

The crrri package has small system dependencies: the only dependency is Chromium/Chrome. Its API is close to node.js libraries: node.js scripts for headless Chrome can be easily transcripted in R.

Through several examples, we will show the different applications of headless Chrome.
Implementation of the cell-key method based random noise approach in R

By Tobias Enderle and Bernhard Meindl

Germany, Austria

The cellkey method (CKM) based random noise approach is a post-tabular Statistical Disclosure Control (SDC) technique. Random noise will be added to each table cell, according to some noise probability distribution and a mechanism to draw from that noise distribution in the so-called lookup step. The transition probabilities define the probabilities of transitioning from one state (e.g. a given original frequency count) to another (e.g. target frequency count). In other words, they define the probabilities of the random noise.

The EU project "Open Source tools for perturbative confidentiality methods" involves the implementation of the cell-key based post-tabular random noise approach in different SDC tools. It was decided to provide a package of the method itself (cellKey) as well as a package for the generation of perturbation tables (ptable) in R.

The goal of the ptable package is to produce transition probabilities that can be used for applying random noise to statistical tables by SDC tools like the R-package cellKey or the software µ-Argus. As suggested in Marley and Leaver (2011) and illustrated in Giessing (2016) the ptable package implements a maximum entropy approach to compute the transition probabilities given the users’ parameter setting such that certain characteristics for the random noise are guaranteed. There is also a shiny app for first time users and visual-style learners. In this way, users can visually learn how parameters affect the probability distribution of the noise.

The R package cellKey makes use of perturbation tables generated with the ptable package and allows to compute perturbed (weighted) frequency and magnitude tables. It features simple ways to define complex hierarchies and proves to be quite fast.
Improvement of the reliability score for autocoding and its implementation in R

By Yukako Toko, Shinya Iijima and Mika Sato-Ilic

National Statistics Center, Japan, National Statistics Center, Japan, National Statistics Center, Japan/Faculty of Engineering, Information and Systems, University of Tsukuba, Japan

We developed the supervised multiclass classifier for autocoding in our previous study. The classifier assigns corresponding classification codes based on reliability scores. The purpose of this study is the improvement of the reliability score for more accurate classification.

The previously defined reliability score considers both the uncertainty from data (probability measure) and the uncertainty from latent classification structure in data (fuzzy measure) in order to consider the unrealistic restriction of an ambiguous text description being classified to a single class, which gives our method a better accuracy of the result. However, this score still has problems to be addressed. The first problem is the reliability score does not consider the frequency of each feature over the codes in training dataset; therefore we cannot consider the confidence degree of existence of each feature. This problem could cause infrequent features to have significant influence leading to the classifier sometimes incorrectly classifying data. The second problem is that the reliability score has employed only partition entropy as a fuzzy measurement. When we consider generalizing our classifier, there is a necessity to apply various fuzzy measurement.

To overcome these problems, we propose improved reliability scores in this study. The new reliability score is defined considering the frequency of each feature over codes in data as the confidence of each feature. In addition, we employ a partition coefficient to the reliability score as another fuzzy measurement.

We will present the comparison of the results and the implementation of coding tasks with the newly developed R package of our classifier during our presentation.
Insights on replicating research results in remote sensing data statistical analysis using R

By Marian Necula and Mihaela E Paun

The Bucharest University of Economic Studies, University of Bucharest

Remote sensing data provides valuable statistical information and official statistics is engaged in active research on such data source and its statistical potential. This paper presents findings of a replication exercise on the research reported in several scientific papers in remote sensing data statistical analysis. While across scientific fields, research replication is presented, with all the bells and whistles associated, as being a major part in doing meaningful science, i.e. testing if the reported results are reproducible, in remote sensing data statistical analysis, the papers which engage in replication are very scarce. Based on several criteria, a number of high impact papers from the field were selected and, under a replication framework, the reproducibility potential of the findings reported in the original papers was scored. From 17 papers selected, for only 6 papers the results were fully reproducible in R. This study provides some insights in doing scientific replication work on remote sensing data statistical analysis using the R environment. The results, by no means, capture the scientific status-quo of remote sensing statistical data analysis, but can be regarded as insights for improvement of the reproducibility aspects of research in this field.
Integration of Survey Data in R Based on Machine Learning

By Spaziani Mattia, Frattarola Dorian and D'Orazio Marcello

Istat, Istat

Integration is a key step in a new statistical production processes within National Statistical Offices to exploit already existing data to produce and disseminate a richer set of statistics.

This work will focus on integration based on Statistical Matching (SM, also known as data fusion) that has the objective of integrating data from sample surveys, typically, with the goal of studying relationship between variables not jointly observed in the same survey (D'Orazio et al., 2006). In Istat these methods have been applied to integrate the survey on household income with the one on expenditures to get a more comprehensive picture of Italian households living conditions. The SM methods considered in past application ranged from simple hotdeck to more complex procedures designed to deal with complex sample survey data (Donatiello et al. 2016a, 2016b). All the experiments have been carried out in R using in particular facilities of StatMatch package.

This work will explore the possibility of using Machine Learning (ML) in integrating households survey data. In particular, some proposals for applying ML techniques in SM will be presented; they reflect findings of first experiments carried out by D'Orazio (2019). Since the household surveys collect mainly categorical variables our application will compare different ML classification methods ranging from simple Naïve Bayes to boosting based procedures. This is made feasible by the variety of R packages implementing modern ML techniques.

The paper will also compare results of new procedures with those provided by popular SM techniques implemented in the R package StatMatch. Compared to hotdeck techniques, the ML techniques in SM allow to skip time-consuming phases to select the matching variables, used in hotdeck to form donation classes or to calculate distance between units. Another advantage of ML techniques is their ability to detect complex interactions between target and predictors. The main disadvantages of ML is the effort required by their tuning.
Just BEAT it. Bethel Extended Allocation for Two-stage sampling package

By Stefano Falorsi, Andrea Fasulo, Alessio Guandalini, Daniela Pagliuca and Marco D. Terribili

Italian National Statistical Institute (ISTAT), Italy, Italian National Statistical Institute (ISTAT), Italy

NSIs use mainly two-stage stratified sampling design for select samples of household surveys from which estimates of very important parameters are derived (such as unemployed rate, average household income, poverty risk and so on). The accuracy of the estimates strongly depends on the sampling strategy. A crucial step, when defining the sampling strategy, is the sample allocation (see, e.g., Cochran, 2007).

Our open source software, called BEAT, determines the optimal sample allocation in two-stage stratified sampling design, allowing to control estimates accuracy of given parameters.

BEAT stands for Bethel Extended Allocation for Two-stage. It is the extension of another open source software called Mauss-R (Multivariate Allocation of Units in Sampling Surveys), implemented by Italian National Statistical Institute researchers. Mauss-R determines the sample allocation in multivariate and multi-domains case of estimates, for one-stage stratified samples. It extends the Neyman (1934) Tschuprow (1923) allocation method to the case of several variables, adopting a generalization of the Bethel's proposal (1989). BEAT develops this methodology but, moreover, it allows to determine the sample allocation in the multivariate and multi-domains case of estimates for two-stage stratified samples.

BEAT collects three function: i) beth that computes multivariate optimal allocation for different domains in one-stage stratified sample design; ii) beth2 that computes multivariate optimal allocation for different domains in two-stage stratified sample design, considering the design effect; iii) bethcv that given a multivariate optimal allocation, calculates coefficient of variation.

Furthermore, BEAT can manage the design effect due to the second stage of selection and the estimator effect due to the type of estimator will be used at the estimation stage (for instance, regression or calibrated estimator).
In conclusion, BEAT could be a very useful, simple and fast tool that can help researchers of NSIs, but not only, in defining the sample allocation in two-stage stratified sampling designs.
Multivariate imputation by chained equation with packages ‘mice’ and ‘mi’ to impute for missing values in semi-continuous data: methods to properly address non-response bias

By Sian Lloyd
Office for National Statistics, UK

The Quarterly Acquisitions and Disposals of Capital Assets Survey (QCAS) is a quarterly collection by the Office for National Statistics (UK) of data on the net expenditure on capital assets by businesses. QCAS estimates are produced in two periods of the quarter, month two (M2) and month three (M3). However, the revisions found between the estimates prompted a review of the current QCAS imputation method. The current method alongside, other single imputation methods, were tested for bias over a series of data from the second quarter of 2015 to the last quarter of 2018. None of the additional single imputation methods reduced the bias, and in some instances, bias was introduced. Subsequently, the author uses the predictive mean matching (pmm) function in both packages ‘mice’ and ‘mi’ for multivariate imputation of data that diverge considerably from distributional assumptions. All the response variables have some percentage of missingness and are semi-continuous. Semi-continuous data is defined here as heavily skewed data that has a continuous distribution except for a probability distribution mass at zero where the zeros are true responses. Further, all data are non-negative and can sometimes take relatively large values. The percentages of zero values and missingness range from 32.4
R tools for ILOSTAT: Rilostat and SMART

By M. Villarreal-Fuentes and S. Ding

International Labour Organization (ILO), International Labour Organization (ILO)

This article presents Rilostat and SMART, two statistical tools developed by the Department of Statistics of the International Labour Organization (ILO) to facilitate the user interaction with ILOSTAT, the largest repository of labour-related indicators. The package Rilostat allows data users around the world to access, extract and manipulate information from ILOSTAT. This document presents a thorough description of the package, including detailed explanations of all its functionalities, examples of reproducible data visualization and a Principal Component Analysis application carried out using information extracted with Rilostat from the Sustainable Development Goals (SDGs) collection available in the database. The Statistics Metadata-driven Analysis and Reporting Tool (SMART) allows the National Statistical Offices worldwide to easily generate and automate the production of analytical reports (such as the SDG national reporting) defined by means of an SDMX Data Structure Definition (DSD), either from processing micro-level data or from aggregated data by means of transcoding. It is a hybrid application that employs the .NET framework to build the user interface and R as the computational and reporting engine. These two R-based tools for ILOSTAT take advantage of all the benefits of the R software to give ILOSTAT data users simplified access to what they need.
Real time monitoring and reporting surveys using Shiny App on Survey Solutions API with R

By Michael Wild, Antoniade-Ciprian Alexandru and Nicoleta Caragea

World Bank, Ecological University of Bucharest and National Statistics Institute of Romania, National Statistics Institute of Romania and Ecological University of Bucharest

This paper presents a real time monitoring system as an essential component of data quality control program, designed to improve quality of survey data collected using computer-assisted personal interviews. In the authors’ approach, the Survey Solutions Computer Assisted Survey System (CASS) serves for designing questionnaires and collecting data through CAPI, CAWI or CATI method, including advanced techniques for minimizing missing data and data capture errors.

In nowadays, the reporting became a very dynamic management process in order to overcome the limitations in terms of time and resources; therefore, this Open Source solutions is a natural choice for any company involved in statistical research, and aiming for high quality survey data.

For real time reporting, we develop an R Shiny Application based on a REST API connection to Survey Solutions CASS, in order to control, manage and adjust the surveys running in the same company.

The paper presents how to use remote sensing data, to determine frame imperfection, as well as survey data and other survey paradata to monitor data quality and representativeness. The reporting data will be available for survey managers through different dashboards and facilitates monitoring of the so collected data as well as process it for data quality reports, almost in real time.
Religiosity and Economic Growth

By Serban Olah and Gabriel Roseanu

University of Oradea, University of Oradea

In this paper the authors would like to explore the relation between religiosity and economic growth using a sample of 44 countries spreaded all around the world, countries that have been selected for the World Value Survey, wave6. In the first part the authors analyze the scientific literature of the topic. The second part is methodological oriented and the authors build a multiple regression model using data from World Value Survey, World Data Bank and Barro-Lee data set. The last part is for conclusions.
Remittances’ impact on receiving country. Multidimensional analysis at macro and microeconomic level using R.

By Valentina Vasile, Elena Bunduchi, Stefan Daniel and Calin-Adrian Comes

Institute of National Economy, Romanian Academy, University of Medicine, Pharmacy, Sciences and Technology of Tirgu Mures, Romania

This research investigates the impact of remittances flows from the country of origin perspective on economic growth at both macro and micro level of the household in Romania and Moldova. We decided to carry out a comparative analysis due to the importance of these external financial flows to the economy. Although the share of remittances in the GDP of the two states differs due to the level of economic development, the constantly increasing labor migration is a common characteristic. In this research we applied time series regression model using \texttt{ts} \texttt{series} packages in R. The expected results of the research are to highlight the indicators influenced by the remittances in Romania compared to Moldova at macro and microeconomic level as well as the type and intensity of the generated impact. This research will help demonstrate that remittance-based economic growth is unsustainable and will highlight the long-term negative impact on the country of origin of these financial flows.
Reproducible Corporate Publications using pagedown

By Romain Lesur

Ministère de la Justice, France

The pagedown package is a new package for producing PDF documents from R Markdown. It uses a different technology stack: PDF documents are created using a web browser and HTML/CSS. It offers a new opportunity for creating reproducible publications and is well adapted to design corporate publications.

Using HTML and CSS for print documents is an opportunity for organizations that already possess skills in web design: they can re-use these skills to design print documents.

This communication will present, based on several examples, the different steps for developing a template compliant with a corporate design.
Rethinking Human Well-Being: A holistic Flourishing perspective

By Josue Aduna Castillo

Romania

Human well-being is a complex construct that has been addressed from different perspectives. It is a concept difficult to define and measure, yet the way in which it is approached drives society to a significant extent. Increasingly, contributions embrace the concept of flourishing to capture the breadth and depth of human life. This perspective usually relies on Aristotelian notions and include a combination of positive feelings and psychosocial functionings, yet morality is seldom considered despite the strong moral component in Aristotle’s account. This research explores the main constituents of well-being under a flourishing perspective that includes moral principles. This study analyzes cross-sectional data from the European Social Survey by performing Exploratory Factor Analysis with the statistical software R. A eudaimonic approach that includes moral attitudes is proposed and nine features were determined via literature review. The results reveal three irreducible dimensions of human well-being: hedonic, eudaimonic and moral. Thus, it is shown that current indicators of well-being fall short to provide a holistic picture of well-being and that the eudaimonic dimension provides a promising pathway to enhance the human condition, both individually and at a societal level.
rtempo - TEMPO Online data interface to R

By Marian Necula, Ana Maria Tiru and Bogdan Oancea

The Bucharest University of Economic Studies, National Institute of Statistics, University of Bucharest

rtempo provides an interface to the Romanian National Institute of Statistics TEMPO Online database, comprised of publicly accessible statistical datasets. The package was built under the Open Data standard, a set of specifications to which governmental agencies adhere when publishing different types of datasets, and on top of several R packages. Although, a previous initiative in this sense was carried two years ago, which resulted in a similar R package, the interface to the NSI's TEMPO database was completely changed and now it is based on Web services. This triggered a re-implementation from scratch of the rtempo package, encompassing a new development philosophy.

In particular, rtempo addresses the need for bulk download of datasets and overrides some inherent limitations present in the current web interface to the TEMPO Online database. In the current development stage, rtempo integrates different types of functionalities such as plotting specific datasets in ready-to-publish maps, mechanisms to combine and analyze the underlying statistical relationship between different datasets, such as correlations, regressions or classifications. In future development stages, rtempo will add functions to export the datasets in formats like xml, sdmx, json, descriptive statistics of datasets, correlation and regression analysis, classifications and presentation of the analysis results in different formats and/or standards, ready for publication.
Selective editing with categorical variables

By M. Rosario Gonzalez-Garcia, Jara Poch, David Salgado and Teresa Vazquez-Gutierrez

Statistics Spain, Statistics Spain

Selective editing is a key editing mode in the statistical data editing process to rationalise resources and to increase the quality of the process (timeliness and cost-effectiveness, especially). In all generic data editing and imputation strategies proposed by the Generic Statistical Data Editing Models (GSDEMs) it plays an important role to identify influential errors in the datasets. So far, these techniques only apply to quantitative variables so that business and household surveys need to use different methodologies and strategies. In the context of the optimization approach to selective editing by Statistics Spain (INE) we show how to formulate a new measurement error model for categorical variables for the same optimization problem driving us to novel local (item) score functions for this type of variables. We show the ongoing software development with a new package implementing this model. In a modular fashion, this new package extends the set of packages for selective editing already in production for some Short-Term Business Statistics in Statistics Spain (INE) thus impinging on the standardization of the statistical process. Besides, the underlying methodology paves the way for the use of machine learning techniques in the computation of different parameters. Our methodology is under test using data (especially the occupation variable) of the European and National Health Survey.
Semiautomatic Reports with R Markdown

By Max Grutter

Statistical Office of the Canton of Zurich

As the Statistical Office of the Canton of Zurich, the regional statistical office in the most populous Swiss canton, we regularly produce a large number of publications. For a year now, we have been adding semiautomatic short reports on recurring topics and data sets to our range of publications. At the push of a button, we can create an article for the homepage that automatically describes and graphically displays the most important developments.

We produce the reports in R Markdown according to cantonal corporate design guidelines. Our "Corporate Design Toolbox" with color- and ggplot-themes helps us with this. For the preparation of the report, we use pre-produced text fragments, which are used depending on the development of the relevant indicators.

The R Markdown templates we created fit perfectly into our infrastructure and follow the concept of data pipelines. At the push of a button, the current data is loaded from the Oracle database, the desired variables are calculated and text, graphics and tables are created for an attractively designed report without further user interaction. Ideally, there are only a few minutes between the release of the data and the publication of the two-page report.

At uRos2019 I would like to give an insight into the production of our so-called quarterly reports. In doing so, I will not only discuss the actual production process, but also point out the challenges arising from the interaction of R, R Markdown and especially Latex (we use Miktex distribution). In addition, I would like to present our Cluster Report. It is our first report inspired by the rules of the RAP (Reproducible Analytical Pipelines) and first reproduced in spring 2019 with updated data.
SMART: A web system for Small Area Estimation

By S. Falorsi, A. Fasulo and F. Solari

Italian National Statistical Institute (ISTAT), Italy, Italian National Statistical Institute (ISTAT), Italy

In the last ten years there was a growing interest from the Italian local governments of employment market data at sub-regional levels in order to optimize the local economic planning activities. To meet this need the Italian National Statistical Institute (Istat) started to disseminate yearly employment rate estimates at Local Labor Market Areas (LLMAs) level by means of Small Area Estimation (SAE) methods.

In addition Istat, in collaboration with local administrators, developed a web system called SMART (Small Area Estimation Tool). Users were allowed to obtain small area estimates of employment market indicators for every possible aggregation of the Italian municipalities. SMART produces model using Linear Mixed Models.

In 2010, a new project was launched aiming to update the web system SMART. Many innovations have been introduced in the system. The new SMART release is completely based on R codes instead of SAS as in the previous version. Furthermore methodological innovation have been implemented as a new SAE estimator based on the logistic linear mixed model and the flexibility in the model specification is increased. For each set of target areas defined by the user, the system allows to set, for the target variable of interest, the default model specification or an alternative one. In particular, it is possible to modify the fixed part of the model in terms of areas defining the model-groups (Estevao et al., 1995) and the poststrata defined in terms of sex and age.

The goodness of the small area estimates is tested through the application of the diagnostic procedures defined in Brown et al. (2001).

For each SAE estimator and for each small area the web system provides the estimate and the corresponding mean squared error. Estimates are rounded in order to avoid confidentiality disclosure problems.

Now the surveys data available on SMART are the Labour Force survey and the European Health Interview Survey. A new project should start soon with the aim of
integrating in SMART the information collected from the 2018 Italian population census that is based on a sample survey, called a Master Sample.

The production process underlying the web system SMART follows the guidelines released by the ESSnet project on SAE (SAE ESSnet, 2012).
The ability of different imputation methods to capture complex dependencies in high dimensions

By Humera Razzak and Christian Heumann

Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany

Multiple-imputation (MI) is a method for treating the problem of missing data. There are various competing computational algorithms available in the R environment to address missing data problems of categorical and continuous variables. In the case of a high amount of missing information, large sample sizes and complex dependency structures among categorical variables, the utility of the provided R packages is somewhat limited. A computationally expedient, fully Bayesian, joint modeling (JM) approach known as ”Dirichlet process mixtures of multinomial distributions” (DPMD), automatically models complex dependencies among variables. But this approach is limited to categorical variables only. We propose a simple and easy to implement combining algorithm which imputes continuous variables using various algorithms and uses the JM approach to detect complex dependency structures among categorical variables. We review, describe and evaluate software packages commonly available in R and compare the results with the proposed MI method by using as example an articial data set. The results suggest that the MI approach which combines the JM approach and various algorithms based on generalized linear models dominates various algorithms when applied solely.
Trial of R training for technical supports to third countries

By Kiyomi Shirakawa and Ken Nakamatsu

Hitotsubashi University, Japan, Takumi Information Technology, Japan

Hitotsubashi university has launched a new project for technical supports to National Institute of Statistics, Cambodia (NIS) in 2016.

The main purpose of this project is to store microdata of surveys and censuses that provided from NIS in Hitotsubashi university and to use those data at an onsite facility in the university.

However, the provided data sometimes is not well maintained. To improve the quality of those provided data, NIS is required to improve their data handling techniques. In addition, as well as other national statistics offices, the demand for applying open software is increasing in NIS.

Therefore, we started a training on data processing and data handling with R for NIS staff in 2018. Although the lecture we provided in 2018 was only basic skills of R for data handling, we will expand our lecture into more advanced contents.

The aim of this trial is to create support cases for countries that need to support tabulation, as well as NIS in Cambodia. In addition, we support the creation of documents such as codebook for microdata, classification tables and consistency test required for data cleaning, tabulation and secondary use.

Especially, we are considering to use some R packages such as ipumsr, dplyr and ggplot2 for data maintenance of accuracy, accessibility and visualization. These packages support reliability by prevent coding errors, improve productivity and versatility.

We will present the detail of our experience of training R in third country including difficulty and future works in our presentation.
Use of R in outliers detection

By Audrius Taraila

Statistics Lithuania

The R project is more often used for research purposes rather than for preparation of official statistics at Statistics Lithuania. Shiny is one of the numerous R packages. This package makes it easy to build interactive web apps straight from R. It allows R users to develop powerful web applications entirely in R without the need to understand HTML, CSS and JavaScript programming languages. It also allows to embed the statistical power of R directly into those web applications. Recently a tool has been developed at Statistics Lithuania using the R program's package shiny. The app allows the user to specify various input parameters using drop-down lists, slides, text fields, etc. The main objective of this tool is the outlier detection. There are three menu bar elements of this tool available: Data file upload, Main characteristics, Outliers detection. Data file upload enables uploading micro data from .xlsx or .csv format files. Main characteristics enables performing primary data analysis such as calculating basic numerical characteristics, correlations, drawing distribution, testing hypotheses or drawing different types of graphs. Several different statistical methods were implemented in Outliers detection, i. e. Selective Editing, Hidirogou-Berthelot, Cook’s distance, DFBETAS, Standardized Residuals and Leverage Points. Each of these methods is applied to detect outliers. After receiving the results, it is possible to export the results in various formats.
Use of R to increase the Automatic Coding Rate in Census2021.
Statistics Portugal Challenge

By Rui Alves and Almiro Moreira

Statistics Portugal, Statistics Portugal

R language and environment for statistical computing and graphics has been used in Statistics Portugal since more than 15 years and is currently widespread used throughout the organization. In this paper, we focus on the use of R within the ongoing project to increase the Automatic Coding Rate and Predictive Analysis of Expressions in Census2021.

In Statistics Portugal, Economic Activity and Occupation are two of the most commonly collected variables when it comes to social surveys. Their classification is performed according Portuguese classifications which are equivalent to the Statistical Classification of Economic Activities in the European Community (NACE) up to four digits and to the International Standard Classification of Occupations (ISCO), respectively.

Both classifications have a significant number of categories which itself poses some challenges to the coding process. In last 2011 Census, about 50

Nonetheless, the biggest challenge still remains the data to classify. In Census2021 we are planning collect all data with CAWI and these questions will be collected in as much detail as possible based on descriptions of the occupation/main tasks and main activity/what is done in the work place.

Nowadays, in current surveys, this that data is collected with an open-ended question and without any pre-coded aid or any kind of input restriction. This results in a high diversity of textual descriptions. The same word can be written in a multitude of variations due to spelling errors, (mis)use of abbreviation, caps, accentuation or hyphenation, just to name a few. This is quite understandable since interviewers (CAPI/CATI) input this data ”on-the-fly” and CAWI respondents are presented with an open field.

Until now about 50 percent of data coding is done manually by a team of coding experts. Knowing that manual coding is nowadays considered both time-consuming and error prone, Statistics Portugal started exploring automatic coding in order to
define the best solution for the implementation in social surveys, and particularly in Census2021.

This paper will address in detail how we are using R to: (1) Create and expand existing dictionaries, (2) Make automatic coding accessible, and (3) Monitor performance and provide useful data to validate results and improve performance.
Using R and RShiny to produce official statistics on Psychiatric Inpatient Care in Scotland - NHS National Services Scotland

By Nikos Alexandrou and Ciara Gribben

NHS National Services Scotland, NHS National Services Scotland

Background

The Information Services Division (ISD) of the National Health Service Scotland produces around 200 health and social care publications each year which are designated either official or national statistics. One of these is the Psychiatric Inpatient Care Publication, which has recently been transformed from a static PDF summary and report, with accompanying Excel tables produced using SPSS, to a set of interactive web pages and accompanying open data produced using R and RShiny. The previous publication was time-consuming to produce and involved considerable manual formatting and checking. Additionally, feedback from customers told us that the data we published could be more informative and user friendly. The new process is largely automated and the output is accessible to a far wider range of customers than the standard publication format.

Methods

The team held a publication incubator, which provided the opportunity to rapidly design and develop a new publication design in a collaborative, focused and fun environment. At this incubator, we split the team into Data Wranglers, Data Visualisers and a Project Manager. We used this time to select elements from the previous publication that would be simple to transform in order to create a prototype. During the week, we also met with stakeholders, including representatives from territorial health boards, the third sector, the government and professional bodies in order to understand what they needed from our publication. Over the 6 months following the incubator, the team transferred data production from SPSS to R using the Tidyverse packages and the outputs from Excel to D3 and RShiny. After we had produced a prototype RShiny app, we conducted user testing with colleagues internally and, after receiving their feedback, conducted similar tests with external customers. During testing, the issues we uncovered included potential disclosure of sensitive information and difficulty in meeting accessibility guidelines. We resolved the former by applying disclosure control to the R code and the latter by minimis-
ing unnecessary items in the layout of our Shiny application, enlarging the font size, and using good contrast between colours. We also chose to release the information presented in the publication through the NHS Scotland Open Data platform which allows customers who are confident in handling data to conduct their own analyses.

Results

The new Psychiatric Inpatient Care publication was released on 25th September 2018 providing customers with our data in a wide range of formats and levels of accessibility:

- The summary page (which contains only key points) is aimed at the general public,
- The trends page is for those who want to see more information but need some support,
- The explorer is for those who are already familiar with viewing data,
- The open data is for the very confident data users who want to conduct their own analyses.

The new method and format has saved time and has reduced the risk of manual errors as the team no longer has to manually format the data into Excel tables or perform manual disclosure control. We plan to expand the range of data available in the publication to include psychiatric care in non-psychiatric facilities and learning disability information in an upcoming release in order to ensure the new format publication covers the same range of topics as the previous publication. We have also developed RMarkdown code to automatically produce our publication summary which will save more time and further reduce the risk of errors. Our experience in transforming the way we publish has not only benefitted our own team but has allowed us to support other teams who are keen to modernise and streamline their own publications.
Utility evaluation of global recoding by accuracy of discrimination model

By Natsuki Sano

Statistical Research and Training Institute Ministry of Internal Affairs and Communication, Japan

As utility measure of statistical disclosure control, various information loss measures are proposed such as discrepancy between original data and protected data. We evaluate utility of protected categorical data based on accuracy of prediction by actual discriminant model. We employ global recording as a method of data protection, and conduct numerical experiments for Japanese census data with minimum frequency ratio p increased. Three hypotheses are conducted by numerical experiments. First hypothesis is when p increases, accuracy of discrimination model decreases. Second hypothesis is when the number of input variables in model increase, accuracy of discrimination model decreases. Third hypothesis is that high accuracy models in original data suffer from information loss by global recording. These hypotheses are supported except for some cases.
Utilization of big data for improving Consumption Trend Index -
Estimation of the number of person per household based on the
characteristic of purchase items-

By Anri Mutoh, Masayo Yamashita, Yoshiyasu Tamura and Masahiro Matsumoto

Official statistics agencies in Japan have been attempting to develop Consumption
Trend Index (CTI) by cooperating with academic researchers and commercial com-
panies as data holder since 2017. The CTI aims to grasp consumption trend quickly
and comprehensively. The CTI has two types: CTI macro and CTI micro. We are en-
gaged in the research of the CTI micro particularly. The CTI micro intends to indicate
the monthly trend of household average expenditure by the type of major items of
households.

The current CTI micro is estimated to compensate for the possible bias in the Fam-
ily Income and Expenditure Survey (FIES) by the Survey of Household Economy and
the Single Household Expenditure Monitor Survey, but there seems to be room for
further improvement. Therefore, we have decided to utilize "big data" that could
be obtained automatically in companies. However, the big data for the CTI micro
provided by companies needs to correspond to FIES, particularly the items of de-
mographic profile. One of the most important such item is the number of person
per household, which tends to be missing in the data from companies. This research
aims to estimate it based on the characteristic of purchase items.

We analyzed the above figure by logistic regression with L1 norm (LASSO) using the R
package "glmnet". The predictive variables of the model are the number of person in
households, and explanatory variables are purchase frequency or purchase price. As
a result, the characteristic of household appearing in purchased items has become
obvious. One of the most distinctive characteristic is in between one-person and
four-or-more-people household: the former purchases more services and high unit
price goods, and the latter purchases more foods and daily necessities which are
comparatively inexpensive and available in large quantities.

With the above analysis, we would be able to integrate the data provided by com-
ppanies with those from the Family Income and Expenditure Survey correctly.
Web Scraping Notices of Insolvency Proceedings - Using publicly available data to enhance survey response quality

By Joerg Feuerhake

Federal Statistical Office of Germany

German insolvency courts are obliged to publish a variety of notifications about opened, ongoing and closed insolvency proceedings. The notifications are meant to inform creditors about relevant topics concerning ongoing insolvency proceedings in order to give them the opportunity to claim insolvency assets. The information is published on a website that provides unstructured text on court decisions, schedules and further issues. This information can be used to augment existing data that is collected from insolvency courts and insolvency administrators by official statistics.

An augmentation of the existing surveyed data is very helpful in the insolvency statistics. Since insolvency courts and insolvency administrators report on the same cases while being institutionally separated record linking and completeness issues occur when integrating the collected data.

In the presented project R is used to search, scrape and process the aforementioned official notices of insolvency proceedings. The aim is to overcome the record linking and completeness issues mentioned above. However there are a few peculiarities that prevent a straight forward information gathering. Foremost the website on which notices are published offers only very limited search functionality to protect the privacy of insolvent debtors. For the same reason the website’s information is not indexed by search engines.

The proposed talk will outline technical and subject-related key features of the web scraping project. Also approaches to deal with restricted searchability and privacy concerns will be presented. Furthermore possible alternatives to the approach are discussed. Finally ideas to further use the processed information are proposed.
We hope you enjoyed your time at uRos2019. See you next time!