The Progress of R in Romanian Official Statistics

Ana Maria DOBRE (e-mail: dobre.anamaria@hotmail.com)
National Institute of Statistics, Romania
Cecilia Roxana ADAM (e-mail: roxana_adam2004@yahoo.com)
National Institute of Economic Reasearch "Costin C. Kiritescu"
of the Romanian Academy, Romania

ABSTRACT

The present paper exposes an overview of the state-of-the-art of R statistical software in the official statistics in Romania, predominantly in the social statistics. Examples on data analysis and econometric models of Small Area Estimation successfully completed are given.

The scientific approach includes also a summary of the applications of R in other statistical offices around the world. Other countries like United Kingdom or Netherlands are truly experienced in the use of R.

We conclude with a series of proposals on the future research opportunities and other potential analysis procedures of R in the social statistics.

Keywords: *R Software, Official Statistics, Social Statistics, Econometric Models, Small Area Estimation*

JEL Classification: *C13, C18, C88*

INTRODUCTION

In 2011, opportunities to develop a long-way but strong implementation of the Small Area Estimation techniques have risen in the Romanian official statistics because of the lack of availability of figures on international migration. Nevertheless, the computational method chosen was R Software because it seemed to enclose all the advantages needed for developing the estimation model.

The small team from international migration using R grew and started to promote R both in official statistics and academic research, including universities. This team was the core of the Romanian R User Group, founded on 4th of April 2013.

The examples on the use of R in other countries underlie the need for spreading R in the Romanian official statistics. A transition and implementation strategy for this new environment could be needed.

LITERATURE REVIEW

Eurostat itself may require in the near future the use of R in the statistical offices across European Union. The arguments could be easy to deduce: low costs, easy customization and use of packages, technical support provided by a large community of users, continuous upgrade and harmonisation in data formats at European statistics offices level.

An argument for Eurostat is considering the use of R is given in the following. In 2012, Eurostat released the report "Analysis of the future research needs for Official Statistics". In the mentioned paper Eurostat provides an analysis of the research tools needed in the statistics offices. In the report is mentioned that "open source architectures will expand in the future, R software is an example." According to Eurostat approach, an integrated use of commercial software and open source software is a foreseen strong tendency, e.g. data and code sharing between products, use of R programs under SAS or SPSS. Within the European Statistical System, large commercial products like SAS and SPSS are often used in production processes and open source software such as R Software are used for methods and technology development, experimentation and statistical innovation. Also, Small Area Estimation is presented as a hot topic in the report, SAE methods connected with visualisation tools being an important future research area. In this report, Eurostat presents the results of the ROS (Research needs in Official Statistics) Survey, conducted between 2010 and 2011. The questionnaire was sent by e-mail and was available online on CROS Portal. The sample had 442 respondents from NSIs, research institutes, universities and others. The results of the survey show that at NSIs, SAS and R is the most applied software and universities have a strong preference for R. Research institutes use different software with the main focus on R, Stata, and SPSS.

A strong documented (Todorov, 2010) implementation of R software is the one of United Nations Industrial Development Organization. Todorov explains that the statistical techniques available in R e.g. linear and non-linear modeling, classical statistical tests, time-series analysis, clustering, as well as its data manipulation and reporting tools make this software an ideal integrated environment for both research and production in the official statistics. The state-of-the-art of R at UNIDO continued in 2012 (Todorov, Templ, 2012) along this line and presents three important areas of data processing and data analysis, typical for the activities of a national or international statistical offices. These areas consist in: missing data and imputation methods, editing and outlier detection and statistical disclosure control.

R-EVOLUTION IN ROMANIAN NATIONAL INSTITUTE OF STATISTICS

In this section we aim to present an overview of the state-of-the art of R in official statistics of Romania, year-by-year. This timeline is created predominantly for the social statistics.

In 2011 appeared the need for a computational tool to develop a Small Area Estimation model. In the last months of the year 2011 a small team of statisticians from International Migration Department started to use R for this purpose. R has been choosen since it is by far the most used open source statistical software among data scientists and academic communities.

In 2012 appeared the first steps and results in using R for Small Area Estimation with packages *JoSAE* (Breidenbach, 2011) and *nlme* (Pinheiro et. al., 2014). The package *JoSAE* is an implementation of the classical methodology of Rao (2003).

In October 2012 the Romanian R Team (www.r-project.ro) was founded.

2013 was a full year, with a plenty of activities. In April, was organized the first Workshop on R – State-of-the-art statistical software commonly used in applied economics, held as a section within EUB-2013 International Conference. There were presentations and free speaking about the advantages of implementing R in academia and official statistics in Romania.

In May 2013, the Small Area Estimation methodology has been successfully completed. The model was based on two data sources – Labour Force Survey and Population and Housing Census. Accurate estimates at NUTS 3 level (county level) have been obtained, outlining figures on international migration statistics.

From June to December, courses on R have been held under the aegis of National Centre for Training in Statistics. About 50 employees have been trained on the course "Introduction in Small Area Estimation Techniques with Applications in R". The course was conceived as an introduction to R, econometric modelling and Small Area Estimation techniques. The structure of the course was as follows:

- Introduction to R: installation, on-line community and resources, GUIs
- Overview of R data types: object-oriented programming, vectors, lists, dataframes, matrices
- Importing and exporting data from/to the following formats: txt, Excel, csv, SQL, SPSS, SAS, DBF
- Functions in R

- Graphics in R: histograms, scatterplots, box-and-whiskers plots, boxplots, scatterplot matrices, 3d plots
- Regression models: linear model, multiple linear regression, logit, probit
- Specification and choosing the independent variables in modelling
- Small Area estimation techniques

As part of the research activities of Romanian R Team, in August 2013 has been released the Romanian version of the well-known book R for Beginners (Paradis, 2005).

In 2013, R started to be used as main tool for data editing and imputation in business surveys in the Romanian official statistics.

As a follow-up of the timeline of progress of R in Romanian official statistics, in March 2014 was organized the second of a series of events dedicated to the use of R Project in Romania: International Workshop New Challenges for Statistical Software - The Use of R in Official Statistics. The workshop was an opportunity to develop new ideas and cooperation in the field of official statistics and academia.

In 2014, the National Centre for Training in Statistics has already planned new courses based on the use of R:

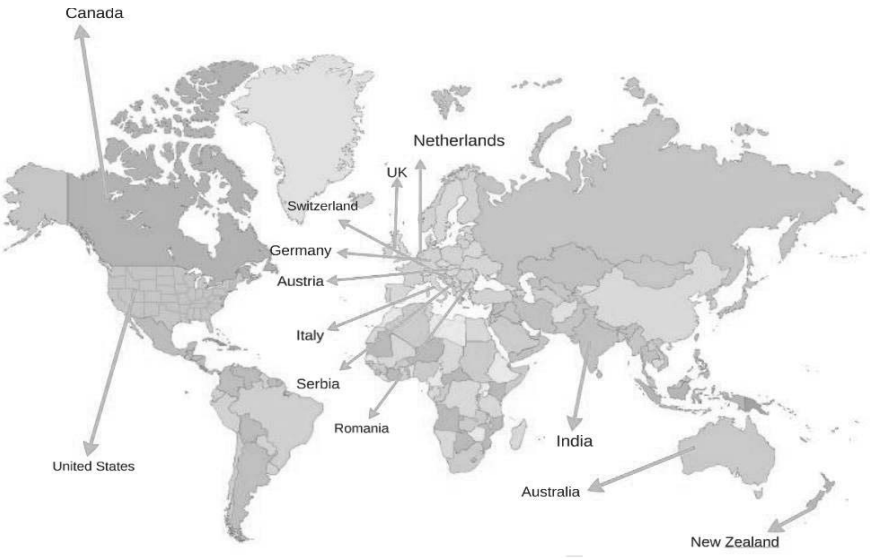
- "R Statistical Software – Presenting Advantages of its use for Data Analysis"
- "Introducing Statistics, the Need for Official Statistics"
- "Statistical Analysis – from Theory to Practice"
- "Concepts, Models and Techniques for Data Analysis"

R IN OTHER STATISTICAL OFFICES

In this section we will expose the spread of R in other statistical offices around the world, according to Figure 1.

Statistical offices using R

Figure 1



In the following we will detail the use of R in many of the countries on the map.

Austria is a pioneer in the use of R in academia, official statistics and business field. Vienna is the location of the headquarter of "R Foundation for Statistical Computing", developed by R Development Core Team in order to: provide support for the R project, provide a reference point for individuals, instititutions or commercial enterprises that want to support or interact with the R development community and hold and administer the copyright of R software and documentation (R Development Core Team, 2005). In Statistics Austria, R is used since 2004 and the experts from there even developed add-on packages for their needs and methodologies. The CRAN Task View for Official Statistics was developed by an expert from Statistics Austria (Templ, 2014).

Istat (National Institute for Statistics, Italy) is using R for sample design, for calibration and calculation of sampling variance, for selective editing, for record linkage, for statistical matching and for small area estimation. Istat has donated software libraries to R and has started to migrate from SAS even since 2009 (European Commission News, 2009).

Netherlands is another country with best practices in using R. Statistics Netherlands developed packages for their methodologies. R is used on three forms, depending on its use in statistics production, statistical research, or research in methods and computation (Van Der Loo, 2012). These three forms are the following: the production installation, the analyst installation and the research installation.

In *United Kingdom,* at the Office for National Statistics, R is used since 2004, mostly in producing statistics. The Economic and Social Data Service (ESDS) is using R software for analysing large scale government surveys (Walthery, 2012).

For the *United States* governement, there is an emerging awareness and recognition of the power of R in their Big Data Initiative. David Smith (2012), Chief Community Officer at revolution Analytics, has highlighted the US approach in using R: harmonize spill estimates from various sources, and to provide ranges of estimates to other agencies and the media; analyze data from clinical trials; research and development of models to predict river flooding; provide a tool to track pollution. In the USA, R is used in agencies like CIA, Food and Drug Administration, National Institute of Science and Technology, Consumer Financial Protection Bureau and San Francisco Estuary Institute.

Beside of being the birthplace of, *New Zealand* promotes in its official statistics the use of R. Institutes like Ministry of Business, Innovation and Employment and Department of Conservation use R language for statistical analyses (Statisphere, 2013).

FUTURE RESEARCH OPPORTUNITIES ON R IN ROMANIAN OFFICIAL STATISTICS

Other possible applications of R in the Romanian official statistics are presented below, according to the CRAN Task View Official Statistics and Survey Methodology (Templ, 2014). Almost all these procedures have dedicated packages; other procedures are enclosed in some packages.

- Complex survey design: algorithms for drawing survey samples and calibrating the design weights; computing point and variance estimates; performing simulation studies; comparing different point and variance estimators under different survey designs; comparing

different conditions regarding missing values, representative and non-representative outliers; create complex survey design (stratified sampling design, cluster sampling, multi-stage sampling and probability proportional to size sampling with or without replacement); selecting samples using probability proportional to size sampling and stratified simple random sampling; univariate stratification of survey populations with a generalisation of the Lavallee-Hidiroglou method (Lavallee, Hidroglou, 1988); estimate (Horvitz-Thompson) totals, means, ratios and quantiles for domains or the whole survey sample; estimation of variance for complex designs by delete-a-group jackknife replication for otals, means, absolute and relative frequency distributions, contingency tables, ratios, quantiles and regression coefficients even for domains; estimate certain Laeken indicators (at-risk-of-poverty rate, quintile share ratio, relative median risk-of-poverty gap, Gini coefficient) including their variance for domains and stratas based on bootstrap resampling; compare point and variance estimators in a simulation environment; incorporation of clustering, stratification, sampling weights, and finite population corrections into a structural equation modelling analysis; post-stratification, generalized raking/calibration, GREG estimation, trimming of weights; calibrate either on a total number of units in the population, on mariginal distributions or joint distributions of categorical variables, or on totals of quantitative variables; calibrate for nonresponse for stratified samples.

- Editing and visual inspection of microdata: convert readable linear (in)equalities into matrix form; applies deductive correction of simple rounding, typing and sign errors based on balanced edits; selective editing for continuous scaled data; robust location and scatter estimation and robust principal component analysis with high breakdown point for incomplete data; visualize missing values using suitable plot methods; profile or explore large statistical datasets.
- Statistical discosure control: generation of confidential (micro)data; simulation of synthetic, confidential, close-to-reality populations for surveys based on sample data; provide confidential tabular data.
- Seasonal adjustment: decomposition of time series; graphical user interface for the X12-Arima seasonal adjustment software
- Computing indices and indicators: estimate popular risk-of-poverty and inequality indicators (at-risk-of-poverty rate, quintile share ratio, relative median risk-of-poverty gap, Gini coefficient); tail modeling

of Pareto distributions for semi-parametric estimation of indicators from continuous univariate; computing various inequality measures (Gini, Theil, entropy, among others), concentration measures (Herfindahl, Rosenbluth), and poverty measures (Watts, Sen, SST, and Foster); computing empirical and theoretical Lorenz curves as well as Pen's parade.

- Statistical record matching between two or more datasources: perform statistical matching between two data sources sharing a number of common variables; linking and deduplicating data sets; nearest neighbor matching, exact matching, optimal matching and full matching amonst other matching methods.
- Small Area Estimation

More than the application on international migration statistics, the SAE method could be used for employment, poverty or education level estimates in social statistics.

The estimates obtained by Small Area Estimation modelling would contribute to policy efforts aimed at reducing poverty, inequality and social exclusion, helping to progress towards the goals of the Europe 2020 Strategy or to design better social and economic policies.

CONCLUSION

R Software represents a huge challenge for Romanian official statistics. The current status of using R is an example of best practice.

An ideal situation would be that where both statistical researchers and IT experts from official statistics would embrace the use of R.

A proposal for future would be a strategy for implementing R and for migrating from Visual Fox and SAS to R. This would be possible with a strong motivation for using R, strong training programme for employees, a wiki-like intranet for know-how sharing, matching migration path to work style, providing technical support and documentation.

ACKNOWLEDGEMENT

The authors are grateful to the R-omanian R Team (www.r-project.ro) and they give their special gratitude to everyone who made this project grow up.

Bibliography

1. Breidenbach, J. (2011) JoSAE: Functions for unit-level small area estimators and their variances. R package version 0.2., http://CRAN.R-project.org/package=JoSAE (Accessed on 9th of April 2014)
2. Breidenbach, J., Astrup, R. (2012) Small area estimation of forest attributes in the Norwegian National Forest Inventory. European Journal of Forest Research, 131, 1255-1267
3. Caragea, N., Alexandru, A.C., Dobre, A.M. (2012) Bringing New Opportunities to Develop Statistical Software and Data Analysis Tools in Romania, The Proceedings of the VIth International Conference on Globalization and Higher Education in Economics and Business Administration, ISBN: 978-973-703-766-4
4. Dobre, A.M., Caragea, N., Alexandru, C. (2013) R versus Other Statistical Software, Ovidius University Annals, 13, 484-488
5. EUB-2013 International Conference, http://www.eub-2013.ueb.ro/sections/ (Accessed on 10th of April 2014)
6. Europe 2020 Strategy, available at: http://ec.europa.eu/europe2020/index_en.htm (Accessed on 10th of April 2014)
7. European Commission News (2009) IT: Statistics institute: moving to open source increases cooperation, https://joinup.ec.europa.eu/news/it-statistics-institute-moving-open-source-increases-cooperation
8. Eurostat (2012) Analysis of the future research needs for Official Statistics, Methodologies and Working Papers, available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-12-026/EN/KS-RA-12-026-EN.PDF (Accessed on 10th of April 2014)
9. Ghosh, M., Rao, J.N.K. (1994) Small area estimation: an appraisal. Statistical Science, 9, 55-93
10. Lavallee, P., Hidiroglou, M.A. (1988) On the stratification of skewed populations. Survey Methodology, 14, 33-43.
11. Paradis, E. (2005) R for Beginners, available at: http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf (Accessed on 10th of April 2014)
12. Pinheiro J., Bates D., DebRoy S., Sarkar D., R Core Team (2014) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-115, http://CRAN.R-project.org/package=nlme (Accessed on 10th of April 2014)
13. R Development Core Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.
14. Rao, J.N.K. (2003) Small Area Estimation, John Wiley & Sons, Hoboken, New Jersey.
15. Rao, J.N.K., Sinha, S.K. (2008) Robust Small area Estimation under Unit Level Models, Proceedings of the Survey Research Section, American Statistical Association, 145-153
16. Smith, D. (2012) Applications of R in Government, available at: http://blog.revolutionanalytics.com/2012/06/applications-of-r-in-government.html (Accessed on 10th of April 2014)
17. Statisphere, Official Statistics System seminar 2013, http://statisphere.

govt.nz/seminars-training-forums/official-statistics-seminar-series/archived-presentations/r-language.aspx (Accessed on 10th of April 2014)
18. Templ, M. (2014) CRAN Task View: Official statistics and Survey methodology, available at: http://cran.r-project.org/web/views/OfficialStatistics.html (Accessed on 10th of April 2014)
19. Todorov, V. (2010) R in the statistical offce: The UNIDO experience. UNIDO Staff Working Paper, Vienna
20. Todorov, V., Templ, M. (2012) R in the statistical offce: Part II, UNIDO Staff Working Paper, Vienna
21. Van Der Loo, M. (2012) The Introduction and Use of R Software at Statistics Netherlands, available at: http://www.amstat.org/meetings/ices/2012/papers/302187.pdf
22. Vergil. V., Caragea, N., Pisica, S. (2013) Estimating International Migration on the Base of Small Area Techniques, Journal of Economic Computation and Economic Cybernetics Studies and Research, Bucharest, 3, http://www.ecocyb.ase.ro/nr.3.pdf/Voineagu%20Vergil.pdf, (Accessed on 10th of April 2014)
23. Walthery, P. (2012) updated by Rosalynd Southern (2013), The R Guide to UK Data Service key UK Service, UK Data Service, University of Essex and University of Manchester, available at: http://ukdataservice.ac.uk/media/398726/usingr.pdf (Accessed on 10th of April 2014)