Use Of R in Statistics Lithuania

Tomas RUDYS (tomas.rudys@stat.gov.lt)
Statistics Lithuania

ABSTRACT

Recently R becoming more and more popular among official statistics offices. It can be used not even for research purposes, but also for a production of official statistics. Statistics Lithuania recently started an analysis of possibilities where R can be used and could it replace some other statistical programming languages or systems. For this reason a work group was arranged. In the paper we will present overview of the current situation on implementation of R in Statistics Lithuania, some problems we are chasing with and some future plans. At the current situation R is used mainly for research purposes. Looking for- ward a short courses on basic R was prepared and at the moment we are starting to use R for data analysis, data manipulation from Oracle data bases, some reports preparation, data editing, survey estimation. On the other hand we found some problems working with big data sets, also survey sampling as there are surveys with complex sampling designs. We are also analysing the running of R on our servers in order to have possibilities to use more random access memory (RAM). Despite the problems, we are trying to use R in more fields in production of official statistics.

Keywords: *R, official statistics, sampling, outlier detection*
JEL Classification: *C10, C88*

1. INTRODUCTION

Statistical software R[1] recently becoming more and more popular not even among sci- entists but also among official statistics offices. It can be used as a very powerful toll for production of official statistics and in some sense replace or work jointly with other statisti- cal systems like SAS[2], which is still very popular in a production of official statistics. Its worth to mention some other advantages of R: easy manipulation with large data sets; ability to import and export data to most common data formats; user-friendly graphical interface RStudio[3] (its also possible to run it from server), Tinn-R[4], Rcmdr[5], etc.; large commu- nity of users and forums where technical support, examples can be found; lots of literature, books and scientific papers; still growing amount of R packages and their updates; etc. One of the main advantages of R is that its free open source software and its highlighted that it comes with absolutely no warranty. Compared to other commercial software this argument leads the user to rely on R software, but not use it as a 'black-box'. To be more clear the user of R software has to be sure about what packages he is using, on other hand he can develop his own packages or functions.

As it was mentioned R software has lot of useful documentation. Regarding official statis- tics the main R packages are listed in CRAN (Comprehensive R

Archive Network) Task View *OfficialStatistics* [6]. There are short descriptions of packages for complex sampling designs, editing, imputation, statistical disclosure control, seasonal adjustment, statistical matching and record linkage, small area estimation, etc. Some other more specific software tools based on R which are created by official statistics specialist and used directly in a production of official statistics can be found in official websites of the organization. For example R pack- age EVER[7] for calibration, estimation and sampling error assessment in complex sample surveys, based on replication methods; ReGenesees[8] system for design-based and model- assisted analysis of complex sample surveys prepared by Istat; Calif[9] for calibration of weights of statistical social and business surveys prepared by Statistical Office of the Slovak Republic.

At the moment the IT structure for production of official statistics in Statistics Lithuania is mainly based on statistical analysis system SAS. In 2013 Statistics Lithuania started an analysis of possibilities to use R is some work processes.

2. USE OF R IN STATISTICS LITHUANIA

There are lot of good practices and examples of using R in official statistics. Taking into account these good practices Statistics Lithuania in 2013 arranged a work group which may task was to analyse the possibilities to use R in production of official statistics and give methodological and other support to the users of R in Statistics Lithuania. There where 5 members from methodology and quality division and 3 members from other subdivisions. In the work group 3 members where experienced R users and other where familiar with concepts and programming. Firstly the members of work group analysed the statistical process and jobs there just basic R knowledge can be applied. That is where just simple data manipu- lation (data import, export, data manipulation with data bases, aggregation, use of simple logical or mathematical function etc.) could be used. Secondly it was determined more com- plex statistical jobs which involved survey sampling and where more specific knowledge of R and its packages are needed (selecting samples, estimation, calibration, etc.).

Short courses on basic R were organized in Statistics Lithuania. The length of this basic R course was about 22 academic hours and it included:

- introduction to R (Console, Script, graphics windows);
- arithmetic operations and functions;
- vectors, matrices, arrays and operations with them;
- R data objects (data frame, list, matrix, vector, etc.)
- data import and export;
- graphical functions, data visualization;
- descriptive statistics in R;
- programming elements;

- writing R functions;
- probability distributions in R;
- data manipulation from data bases (package RODBC[10]).

About 58 employees had attended this basic R course in Statistics Lithuania.

At the moment in Statistics Lithuania quarterly Labor Cost Index is calculated directly with R. It is a relative indicator reflecting the change in labour costs per hour worked over a certain period of time. Earlier these calculations were done with SAS. The programs for cal- culation the Labor Cost Index were prepared by subject mater specialist in Labour Statistics division after the basic R course. No special packages were needed, just basic R programming elements were used.

In 2013 results for Farm Structure Survey were obtained using R. A Farm Structure Sur- vey is a statistical sample survey on farms producing agricultural products. Its target is to get information about the structure and types of agricultural farms, utilised area, fruit and berry plantations, crops, farm labour force between censuses. Farm Structure Survey is stratified sample survey, sample design was created using SAS. For estimation of totals of different indicators in different estimation domains (groups) Horvitz-Thompson[11] estimator was used. Survey data were uploaded from Oracle data base with RODBC package. Two main functions were used: `odbcConnect` - to create connection object and `sqlQuery` - to get data from data base simply using SQL[1] statement as an argument in this function. Raw data were prepared for estimation stage using simple logical, arithmetic and conditional sentences from basic R. For an estimation of parameters and accuracy measures package survey[12] was used. Function `svydesign` was used to describe survey object, `svyby` - to get estimates in domains (groups), `cv` - to get estimates of coefficient of variation. All programs in R, calculations and estimation were done by subject matter specialists in Agri- cultural and Environmental Statistics division.

From 2015 Statistics Lithuania started to use R for analysis of Official Statistical Portal data. Written programs were used to analyse information about how external users uses information from Statistics Lithuania Database of Indicators[14]. The output of this analysis is most popular indicators and how many times their were chosen by users, number of prepared statistical tables, average number of prepared statistical tables in a day, average time for preparation of statistical table, etc. The report is semi annual. Firstly this report was generated using R Markdown for generating dynamic documents, later just the main information and output tables were used for generating final report. At the moment this report is not public.

1. Structured Query Language (SQL) is a special-purpose programming language designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS)

More recently experts from Statistics Lithuania started analyze possibilities to use R in data editing. The analysis was started by analyzing the methods that could be used for outlier detection in Short Term Statistics (STS) surveys. Two methods that does not incorporate auxiliary information were analyzed: interquartile range rule [16] and the Hidiroglou- Berthelot[13] method. The first one is quite easy to program, because user do not need to use special R packages. Its possible to calculate interquartile range by using `IQR` function and calculate first and third quartiles by using function `quantile` with the argument `probs`. The R function to detect outliers with the Hidiroglou-Berthelot method was developed by Statistics Lithuania experts. Some basic R functions like `median`, `quantile`, `max` were used. Its well known that the use of auxiliary information could improve outlier detection. At Statistics Lithuania VAT data (Value added tax) were used to analyze if it is useful to incorporate this auxiliary information in outlier detection. As the main survey variable (from STS surveys) turnover has good correlation with auxiliary VAT data it is useful to analyze their linear relationship. Measures like hat diagonal index, standardized residuals[17], Cook distance[18] were used for outlier detection. These measures can be easily obtained using functions `hatvalues, rstandard, cooks.distance given` the object as an argument created with function lm. Recently selective editing method was also included in analysis. R package SeleMix[15] is used for outlier detection. Small example of outlier detection will be presented in simulation section.

Despite the fact that Statistics Lithuania recently started to use R in some projects, and will continue to do analysis were it possible to use it, there are some problems unsolved. First off all it is quite difficult to change the opinion of experts to start using R, because main IT structure for production of official statistics is based on SAS. All programs related to data analysis, manipulation and estimation is written in SAS. Secondly it is still an open question to use already developed R packages or develop own tools for more complex tasks. This leads to have a group of more experienced R user in Statistics Lithuania. One of the main concerns is survey samples, because complex designs often differs among counties. In this case a big analysis of available R packages has to be done and decision of available packages that could be used should be done by survey specialist. More easy to adopt already available packages for estimation if only the estimators are not complex. There are also some more technical problems, like working with big data sets. Sometimes it is not possible to import or export big data set to MS Excel, Access. The analysis of big data sets are also slow, sometimes merging big data sets are impossible. These problems occur that there is not enough RAM[1] in the computer and it looks like it is

1. Random-access memory (RAM) is a form of computer data storage. A random-access memory device allows data items to be accessed (read or written) in almost the same amount of time irrespective of the physical location of data inside the memory.

more hardware than software problems. In some cases this problem could be solved running R on server, there more RAM could be assigned for running R. In Statistics Lithuania RStudio server is running on virtual server to analyze these problems. It was noticed that in some cases RSudio server works quite well with big data sets.

3. SIMULATION STUDY. OUTLIER DETECTION

In this small simulation example an application of outlier detection methods mentioned earlier using statistical software R is presented. We are nor giving theoretical notations of the methods that are used in this simulation, because this short paper focuses more on overview on use of R in Statistics Lithuania. All theoretical notations can be found in references.

Real data from quarterly Short Term Statistical survey on service enterprises for first quarter of 2015 are used in this simulation. The parameter of interest is total turnover of enterprises in estimation domain (group). Estimation domains are taken as economical activ- ities according to Statistical Classification of Economic Activities (NACE rev. 2) at 4 digit level. As an auxiliary information VAT data were used.

Four methods were applied in this simulation. Interquartile range rule is referred as IQR. Measures: hat diagonal index, standardized residuals, Cook distance is referred as (LR). In this simulation study the element is taken as an outlier is one of these measures indicates that an element is outlier. The Hidiroglou-Berthelot method is referred as (B-H), and selective editing method is referred as (SE). Using these four methods an outlier detection is done for the main survey variable - turnover. The results are shown in the Table 1.

Number of outliers by different methods

Table 1

Domain	IQR	LR	B-H	SE
5229	16 (4.0 %)	6 (1.5 %)	19 (4.8 %)	5 (1.3 %)
6829	16 (6.8 %)	12 (5.1 %)	22 (9.4 %)	2 (0.9 %)
7911	13 (7.2 %)	8 (4.4 %)	7 (3.9 %)	29 (16 %)

In the table simulation results for three estimation domains according to NACE rev. 2 at 4 digit level are shown. The number of outliers are given for four different methods in columns. The share of outliers in particular domain is shown in percent. From the results we can see that methods find different number of outliers for particular domains. For domain 6829 maximum number of outliers (22) are found by B-H method and minimum (2) by SE method, the range is very big. About the same number of outliers were found by IQR method in all domains. In this example all constants that were needed for calculations were set the same for all domains. To conclude, more accurate analysis should be made on choosing outlier detection method. It should analyzed how and what different constants for different domains to use in these methods.

4. CONCLUSIONS AND FUTURE PLANS

R becoming more and more popular among official statistics offices. Recently Statistics Lithuania started to analyze possibilities to use R not even for research purposes. The projects started then Statistics Lithuania organized short courses on basic R for its experts. At the moment calculation of Labor cost index, some reports on using Statistics Lithuania Database of Indicators are done in R. The results for Farm structure survey was prepared with package survey. Also analysis of outlier detection is analysed in R. Upcoming plans are to pay more attention and analyze survey sampling methods in R and prepare short courses, develop outlier detection, analyze small area estimation methods, motivate Statistics Lithuania experts to use R, etc. Despite some challenges and difficulties Statistics Lithuania will focus on developing the statistical production process with R.

References

[1] **R Core Team.** *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ***http://www.R-project.org/***

[2] SAS statistical software. World Headquarters, SAS Institute Inc. 100 SAS Campus Drive Cary, NC 27513-2414, USA. ***http://www.sas.com/***

[3] **RStudio Team.** *RStudio: Integrated Development Environment for R*. RStudio Inc.,Boston, MA, 2015. ***http://www.rstudio.com/***

[4] **Faria, C., J.**, 2009, *Resources of Tinn-R GUI/Editor for R Environment.*. UESC, Ilheus, Bahia, Brasil, 2009. ***https://sourceforge.net/projects/tinn-r/***

[5] **Fox, J.**, 2005, *The R Commander: A Basic Statistics Graphical User Interface to R*. Journal of Statistical Software, 14(9):1-42.

[6] **Templ, M.** *CRAN Task View: Official Statistics & Survey Methodology*. ***https://cran.r-project.org/***

[7] **Zardetto, D.** *EVER: Estimation of Variance by Efficient Replication*. R package version 1.2. The Italian National Institute of Statistics.

[8] **Zardetto, D.** *ReGenesees: R Evolved Generalized Software for Sampling Estimates and Errors in Surveys*. R package version 1.7. The Italian National Institute of Statistics.

[9] **Frankovic, B., Vlacuha, R.** *Calif: Calibration of weights of statistical surveys*. Program version 3.2. Statistical Office of the Slovak Republic.

[10] **Ripley, B., Lapsley, M.** *RODBC: ODBC Database Access*. R package version 1.3-12. ***https://cran.r-project.org/web/packages/RODBC/RODBC.pdf***

[11] **Horvitz, D. G., Thompson, D. J.**, 1952, *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association, 47, 663685.

[12] **Lumley, T.** *Survey: analysis of complex survey samples*. R package version 3.30-3. ***http://r-survey.r-forge.r-project.org/survey/***

[13] **Hidiroglou, M.A.**, and **Berthelot, J.-M.** 1986, *Statistical Editing and Imputation for Periodic Business Surveys*. Survey Methodology, 12, 73-83.

[14] Official Statistics Portal. *Database of Indicators*. ***http://osp.stat.gov.lt/en/statistiniu-rodikliu-analize1.***

[15] **Guarnera, U., Buglielli, T., M.** *SeleMix: Selective Editing via Mixture models*.R package version 0.9.1. ***https://cran.r-project.org/web/packages/SeleMix/SeleMix.pdf***

[16] **Tukey, J., W.**, 1977, *Exploratory Data Analysis*. Addison-Wesley.

[17] **Pope, J., A,** *Influential Observations in Linear Regression*. U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, National Ocean Survey, Geodetic Research and Development Laboratory, 136 pages.

[18] **Cook, R., D.** ,1979, *The statistics of residuals and the detection of outliers*. Journal of the American Statistical Association, 74(365):169174.