RR by R in Official Statistics

Ciprian ALEXANDRU
Nicoleta CARAGEA
Ecological University of Bucharest and National, Institute of Statistics, Romania

Motto: "Golden Rule of Reproducibility:
Script everything"
Roger D. Peng

ABSTRACT

Nowadays there is a growing alarm about scientific research results that cannot be reproduced. The reasons comprise increased levels of control, complexity of experiments and statistics, and also time pressures on researchers. Journals, scientists and research institutions all have a part in promoting reproducibility. On the other side, the progress in computational science has barriers because the researchers do not have the ability to independently reproduce or verify published results.

The reproducible research using R in official statistics is one of the objectives that the Romanian Institute of Statistics takes into account to achieve in the near future. The first step is to implement the idea according to "publications produced in official statistics should be accompanied by the relevant information and data to reproduce the results and findings".

Key words: R, Packages, reproducible research, official statistics, data science
***JEL Classification**:* C10, C18

1. INTRODUCTION

Reproducibility is quite a new concept, developed in strong correlation with new trends in data sciences, regarding the increase of the availability of the data and development of IT technology. The basic idea of reproducibility is making the data and the computational methods, from the original study, available for the public, so that anybody having possibility to run the same code/analysis to be able to come to the same results (Peng, 2015).

Reproducible Research by R (RR by *R*) became almost a standard for the reproducible reporting taking into consideration the "power" of the *R* environment (Ihaka and Gentleman, 1996), in recent years. The computational tools are considerably a key factor of the research process, from the exploratory steps until the dissemination of the findings. The usability of the *R* environment, the interoperability with a lot of the others tools, like editors, databases engine and other programming languages, the open source license,

were key factors in the decision taken in many statistical offices to introduce *R* as a tool in the daily activities and research.

Research is a daily activity in the statistical offices, very important for developing new methodology, identifying the data sources for new challenges in statistics, in elaboration of statistical analysis, regarding to data production. Official statistics offer information for a wide range of users, from local to central government, also for researchers and journalists, for business and citizens, in general. The reproducible research by R in Official Statistics has two issues, one from the perspective of the researcher and one from the perspective of production of statistics.

In this paper we try to find the answer to some questions like: Research, where is it going? Replication or Reproducible? Why Reproducibility? What is necessary for Reproducibility? Which are the development tools, the presentation tools and the research pipeline in R?

2. LITERATURE REVIEW AND BACKGROUND

The European Commission supports numerous research projects in different fields of statistics, which are being carried out by public and private research institutes on behalf of the Statistical Office of the European Communities (EUROSTAT). The digital transformation is a real process which conducts our society to a data-society, with some possible threat emerging from the increase of the entropy of data availability. In the same time, the increasing of availability of data, as a trend, is already seen by the European Statistical System (ESS) as a strategic relevance for official statistics (ESS Vision 2020, 2014).

The challenges of official statistics come from these new data sources, huge amounts of data (Big Data), high dynamic data, maintain a comparability of the data, in time but also between EU members state, the timeliness and cost efficient. The role of private companies was changed and a partnership from private data source owners and official statistics should be a strategic issue.

One of the relevant responses to these challenges could come from the source of the challenges, the IT development, the data computational tools, the IT infrastructure and the human resource development. In the table below we synthesize the solutions offered by the *R* environment to those issues.

The main objectives in developing the ESS activities	Response from Reproducible Research by R
Quality in all activities	*R* environment is developed by statisticians, also through the reproducibility process the results could by validate by the academic community
Deliver coherent, **relevant and reliable statistics** based on internationally harmonized concepts and methodologies	Almost any functions for any statistical model are available in the R system and the methodology used by the researcher is available through a system by vignette.
Engages users proactively and meets their demands in a cost efficient and responsive manner	For an open-source software, R environment is open for inspection and modification to anyone who wants to see. For 15 years code improvement is a continuous activity (Inside-r.org, 2016).
Promotes efficiency and realizes productivity gains through **collaboration** in sharing methods, tools and technological infrastructure where it is allowed by appropriate data and human resources	The project leadership of the R include more than 20 leading statisticians and computer scientists from around the world. R packages are developed by other thousands people and are used by 2 million users.
Embraces opportunities provided by the **digital transformation** and harnesses new data sources to produce meaningful statistics	The newest software technologies are implemented in R core or in the packages.
Delivers information in an **interactive** and easily understandable way	Very advanced functions for interactive visualizations are available in R (Parkkinen, 2014).

3. KEY POINTS IN REPRODUCIBLE RESEARCH

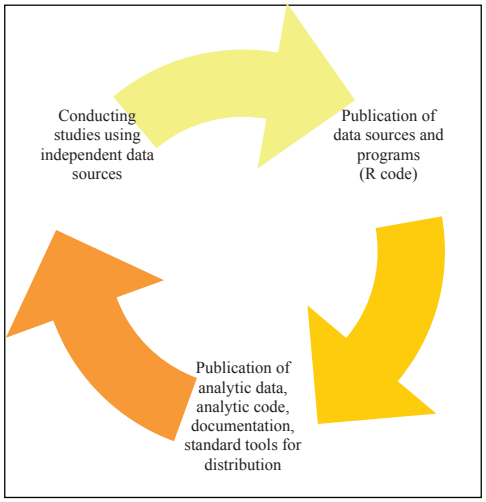
Replication is the state of the art in verification and validation of the results, characterized by independent researchers which came with same conclusion as the originals. Sometimes this could be a process impossible to reproduce taking into consideration technical aspects or uniqueness of the experiment. Other very significant barriers to replication of an experiment consist in lack of resources, human, material or financial. Also, when time is a characteristic of the experiment, reproducibility would require other conditions, because it would reproduce the phenomenon in another period of time. In this latter case fall many social phenomena, such as studies on events, based on data collected from social networks or time series analysis related to financial or economic phenomena.

Compared to replication of the experiment, in a reproducible research process, the original data and programming code are studied by independent analysts obtaining the same results of the original study, so the data analysis can

be repeated with same findings. The advantages of the reproducibility are: *enhancing the scientific evidence, ensures transparency, creates confidence* and *validates the data analysis*. Nevertheless, reproducibility isn't synonymous with correctness, merely a wrong assumption can be identified by an independent replication process, that is easier than reproducibility.

Reproducibility involves "the calculation of quantitative scientific results by independent scientists using the original datasets and methods" (Wilson, 2014).

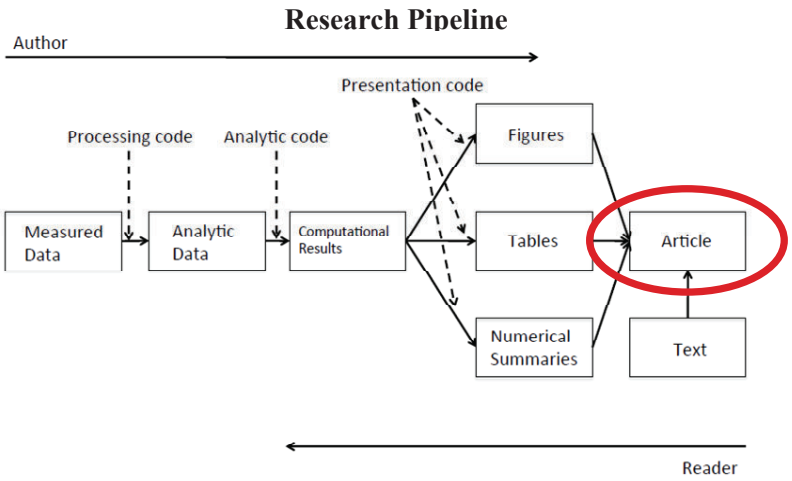
What is necessary for Reproducibility?



How easy can we make our work reproducible? It could be very useful in preparing data into a friendly format or to find a web server for uploading code and data with "almost" limitless availability. The work could be reproducible using some tools to help readers (well known or well documented, eventually) and making guidelines available for future use.

Reproducible research is considered a threshold between more time consumed in present and saving time for the future.

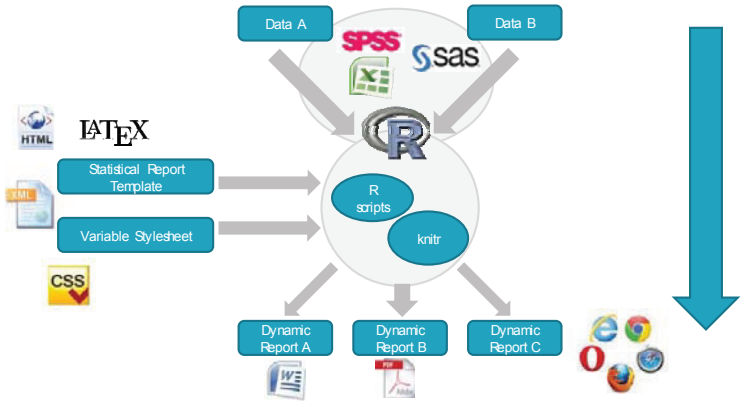
The schemes of reproducible research procedures are given by Peng, for general research and by Alexandru, for R research, as below:



Source: Roger D. Peng, *Reproducible Research: Concepts and Ideas*

The research pipeline in *R* environment starts with importing the data, which can be raw data or cleaned data. This process is very important for official statistics regarding the issues that come from surveys, data collections, online data collection and correlations between administrative databases.

Research Pipeline in R



The next step is statistical analysis by executing the code (*R* scripts) according to the theoretical model stipulated in statistical methodologies. The last step is presenting the results of the analysis, by graphs, tables or any kind of summaries, into a dynamic report, through different techniques, like HTML, XML, CSS, $L^{A}T_{E}X$.

4. DEVELOPMENT TOOLS FOR REPRODUCIBLE RESEARCH IN OFFICIAL STATISTICS

The basic principle of reproducibility is to combine computer code and software documentation in the same document (Xie, 2014). The results from a compiled code could be mixed with documentation; also the final report/ document may or may not include the source code. In any case, reproducibility implies an automatic generation of the reports, starting from scratch, from raw data, to data analysis.

Development tools consist in *documentation language*, that is human readable, and *programming language*, meaning a machine readable language. The most used documentation languages are the following:

- *Sweave*, that allows to embed the *R* code for complete data analyses in $\LaTeX$ documents (Leisch, 2002), and
- *knitr*, a *R* package for dynamic report generation, that allows various programming languages, like *R*, *Python* and *awk*, and, as output markup languages $\LaTeX$, HTML, and Markdown (Xie, 2015).

As programming languages, *R* (R Core Team, 2015) and Python are the most productive and very popular among the data analyst.

This mix of documentation and source code is well known as Literate Programming.

The last step into the pipeline of reproducible research process are data repositories and presentation tools, for dissemination of the final report and the final results. The advantages of these repositories are given by open access, a well known structure of the information and long time availability. In the last years, the most used repositories and publishing tools are:

- GitHub - powerful collaboration, code review, and code management for open source and private projects [https://github.com/]
- RPubs - web publishing from R [https://rpubs.com/]

Through these tools, the authors and the researchers have the opportunity to put together data and code for public distribution of their work. Sometimes, the official statistics have some legal restrictions regarding the access to the microdata, but this issue must by overcome by anonymisation of data, before the research pipeline itself.

R activities in Romanian official statistics

In Romania, in the National Institute of Statistics, since 2013 a team from demographic statistics is working in data analysis activities using R software.

The main results are related to the international migration methodology where small area techniques were applied in order to produce reliable statistical data at territorial level.

Other R activities involve managing data-bases from statistical and administrative source using Data-matching procedures.

Small area estimation is also in progress to be applied to improve the quality of poverty indicators at regional level on data provided by EU-SILC.

A new challenge for the National Institute of Statistics is to manage Big Data using R.

During 2013-2016, in the frame of annual professional continuing education programmes, over one hundred of statisticians were trained on 3-5 days courses for data analysis using R software.

The National Institute of Statistics, together with researchers from universities and research institutes have organized R-related scientific conferences and workshops, the most focused debates being the use of R in official Statistics. Statisticians from NIS participated at relevant R conferences sponsored by others and promoted the use and development of R and R-related software in Romania.

5. CONCLUSIONS

For sure, reproducible techniques are taken into consideration very seriously by the official statistics, in research departments, statistical analysis and in operative and production process, respectively. The challenges are quite high, considering the necessity of methodology changing according to the specificity of research pipeline. The tools used for the reproducibility process are sometimes significantly new, compared with old software and techniques. Nevertheless, enhanced graphical visualization, online interactions, real-time analysis, web integration of the reports, prompt and effective response, given the rapid changes in data sources are available from administrative sources, but also open data sources.

Researchers and statisticians have to manage and preserve statistical data with very complex metadata support to allow others to easily find, understand, and use archived files. At the moment, data, metadata and files are produced and achieved using other tools, but implementing reproducible research using R in official statistics in Romania is one of the most important challenges in next years.

6. References

1. **Ihaka, R., Gentleman, R.**, 1996, *R: A language for data analysis and graphics*, Journal of Computational and Graphical Statistics, 5(3):299–314.
2. **Leisch, F.**, 2002, *Sweave: Dynamic generation of statistical reports using literate data analysis*. In COMPSTAT 2002 Proceedings in Computational Statistics, number 69, pages 575-580, Physica Verlag, Heidelberg, ISBN 3-7908-1517-9. [bib | PDF]
3. **Parkkinen, J.**, 2014, *Interactive visualizations with R - a minireview*, [***http://ouzor.github.io/blog/2014/11/21/interactive-visualizations.html***]
4. **Peng, R.**, 2015 "Report Writing for Data Science in R", Leanpub, 3-7.
5. **R Core Team**, 2015, *R: A Language and Environment for Statistical Computing*., R Foundation for Statistical Computing, Vienna, Austria
6. **Stodden** and **Miguez** 2014, *Best practices for computational science: software infrastructure and environments for reproducible and extensible research*, Journal of Open Research Software.
7. **Wilson, B.**, 2014, *Implementing Reproducible Research*, Journal of Statistical Software, October 2014, Volume 61, Book Review 2, [***http://www.jstatsoft.org/***]
8. **Xie, Y.**, 2014, *knitr: A Comprehensive Tool for Reproducible Research in R*, Implementing Reproducible Research, Publisher: Chapman and Hall/CRC
9. **Xie, Y.**, 2015, *Dynamic Documents with R and knitr - Second Edition*, Publisher: Chapman and Hall/CRC, [http://yihui.name/knitr/].
10. *** Yale Law School Roundtable on Data and Code Sharing. *Reproducible research: addressing the need for data and code sharing in computational science.* Columbia University Academic Commons, 2010.
11. *** European Statistical System Committee (2014), *ESS Vision 2020*, Luxembourg, 4-5.
12. *** Inside-r.org, 2016,*What is R?*, ***http://www.inside-r.org/what-is-r***.
13. *** DOI, JIMMY; POTTER, GAIL; WONG, JIMMY; ALCARAZ, IRVIN; & CHI, PETER. (2016). *Web Application Teaching Tools for Statistics Using R and Shiny*. Technology Innovations in Statistics Education, 9(1). uclastat_cts_tise_27492. Retrieved from: ***http://escholarship.org/uc/item/00d4q8cp***