# R – a Global Sensation in Data Science

**Nicoleta CARAGEA** (nicoletacaragea@gmail.com)
**Antoniade-Ciprian ALEXANDRU** (alexcipro@yahoo.com)
Ecological University of Bucharest - Faculty of Economics
**Ana Maria DOBRE** (dobre.anamaria@hotmail.com)
National Institute of Statistics, Romania

## ABSTRACT

*The main objective of this paper is to expose the evolution of R, as the most used data analysis tool among statisticians and the academic researchers. Its flexibility and complexity simply gained the statisticians and data scientists.*

*The paper examines some of the reasons behind the popularity of R, using tools like SWOT analysis.*

*R software environment offers integrated tools for a very large area of data analysis, from computations and data mining to high-effects visualization. As an example, we performed in this paper an illustration of 3D plotting.*

***Keywords:*** *R Software, R Packages, Statistics, Data Visualization, 3D Plotting*
***JEL Classification:*** *C13, C18, C88*

## Introduction

*Motto:* "It is easy to lie with statistics.
It is hard to tell the truth
without statistics" (Andrejs Dunkels)

Nowadays, R is the most used and appreciated tool in data science. The R system implements a dialect of the influential S language but has its own GUIs and IDEs.

The fact that especially in Romania, but common also in other countries, most of private or public institutions use commercial statistical tools having a predictable cost, did not embarrassed the spread of R use and its growth.

In this context, R itself concerned on a huge industry of data science, data mining, open data – both in the private and public sector and in many fields, such as statistics, medicine, biology, geographic information system, social media, marketing, finance, engineering and so on.

This paper represents a further research of the authors (Caragea et. al, 2012).

## LITERATURE REVIEW

R first appeared in 1996, when the statistics professors Ross Ihaka and Robert Gentleman of the University of Auckland in New Zealand released the code as a free software package, under GNU General Public License..

R is considered the *lingua franca* for statisticians and recently for data scientists.

David Smith (2011), the Chief Community Officer of Revolution Analytics, considers that data science is a valuable rebranding of computer science and applied statististics skills. In fact, the terminology of data science is related to statistics, data mining, Exploratory Data Analysis, big data, artificial neural network, forecasting, decision tree. Nowadays, many companies hire "data scientists" and many conferences are held under aegis of "data science".

## SWOT ANALYSIS OF R PROJECT

A comprehensive and well-documented SWOT analysis of R software is necessary to understand its advantages and disadvantages and to define a possible causality of R's rapidly growth among data analysis tools, according to the illustration below.

**SWOT Analysis of R**

*Figure 1*

| | | |
|---|---|---|
| **Strengths:**<br><br>Open-source program<br>A fantastic user community that keeps growing |  | **Weaknesses:**<br><br>R keeps all the data in the RAM memory so it can consume very quickly the available memory |
| **Opportunities:**<br><br>R is the product of international joint of top computational statisticians and computer language designers | | **Threats:**<br><br>It is considered by many to be harder to learn than other similar software due to the fact that it has more types of data structures than the data set |

A detailed SWOT analysis is presented in the next section of the paper.

**STRENGTHS**

- Open-source program
- R and its GUIs and IDEs are completely freeware; the cost of using R are related only with training of users
- R is cross-platform: it runs on Windows, Linux, Mac OSX
- A fantastic user community that keeps growing
- User support through a very active mailing list, blogs, dedicated forums
- Being a challenge for every user to involve himself and to exchange knowledge
- Continuous develop and release at academic level, growing list of print books and e-books
- Linked with the way statisticians think and work (e.g.: keeping the track of missing values)
- Meets the changing needs of shifting global economy because of its flexibility
- Competitive tools for Geographic Information Systems
- Operations Research; SPSS has not this issue available
- R supports connection with the main commercial software, such as:

JMP, MATLAB, Spotfire, SPSS, STATISTICA, Platform Symphony and SAS.

- The freedom to teach with real-world examples from outside organizations, which is forbidden to academics by SAS and SPSS licenses
- The flexibility to mix-and-match models, scripts and packages for the best results
- R functions can nest inside one another, creating nearly infinite combinations
- Easy to create scripts with all the steps for an analysis, and run the script from the command line or menus
- R is an object-oriented language and has the advantage of operating on an object according to methods that make sense and also the methods can adapt to the type of object
- Intermediate results can be reviewed, and scripts can be edited and run as batch processes
- R stimulates critical thinking about problem-solving rather than a "push the button" mentality
- Every computational step is recorded in the background, and this history can be saved for later use or documentation
- The possibility to transform R code into HTML code so that it could be published on web (via Rpubs)
- Turn analyses into interactive web applications that anyone can use (via Rshiny) without necessary HTML or JavaScript knowledge
- R allows importing data from Microsoft Excel, Microsoft Access, txt, SAS, SPSS, Visual Fox Pro, Oracle, MySQL, and many more formats
- R can handle a few millions of records on a regular PC, and there are some great packages that support handling larger data than the actual RAM

**WEAKNESSES**

- Data collection should be available from other tools; MySQL or PostgreSQL are popular among useRs for this purpose
- R keeps all the data in the RAM memory so it can consume very quickly the available memory
- Direct Marketing not available
- Guided Analytics not available
- The help files and the vignettes for packages are written for relatively advanced users; documentation is sometimes impenetrable to the non-statisticians
- R is not very user friendly and it needs basic knowledge of

programming language; that will limit R's long-term growth because GUI users far outnumber programmers
- The default GUI of R is limited to simple interaction and does not include statistical procedures; the user must type commands for importing data, computing and plot graphs

**OPPORTUNITIES**
- R is the product of international joint of top computational statisticians and computer language designers
- Users' contribution to program's ongoing development; anyone is welcome to provide bug fixes, code enhancements, and new packages
- Share new techniques with other R users around the world via online community
- Re-use and reproduce new discovered techniques on analytic operations that the user is going to perform
- Very large area of use - statistics, business analytics, finance, journalism, mapping, forecasting, social networking, spatial analysis, engineering, science, drug development, computational biology, and many more
- Easy to export results to usual formats and get data visualization like maps, 3D surfaces, image plots, scatter plots, histograms, bar plots, pie charts, multi-panel charts and many more
- IT skills in R are very appreciated on the labour market
- R for mobile devices: R has version for OS X ("R Programming Language" on Itunes), as well as a server-based implementation of RStudio for Android ("R Instructor" on Google Play)
- R supports big data and performs big data analysis
- R supports multicore task distribution and parallel computing
- R offers many facilities to learn basic statistics
- Currently available CRAN Task Views on various topics and the possibility to extend with some others

**THREATS**
- It is considered by many to be harder to learn than other similar software due to the fact that it has more types of data structures than the data set
- It is necessary for the user to carry out the macro language of R and to control the management of the output; SPSS and SAS allow user to skip those issues until he needs them

## INTEREST OF USING R

The interest of using R can be quantified by various tools such as Google Trends. Bob Muenchen is the author of r4stats.com, a blog which analyses the popularity of data analysis software.

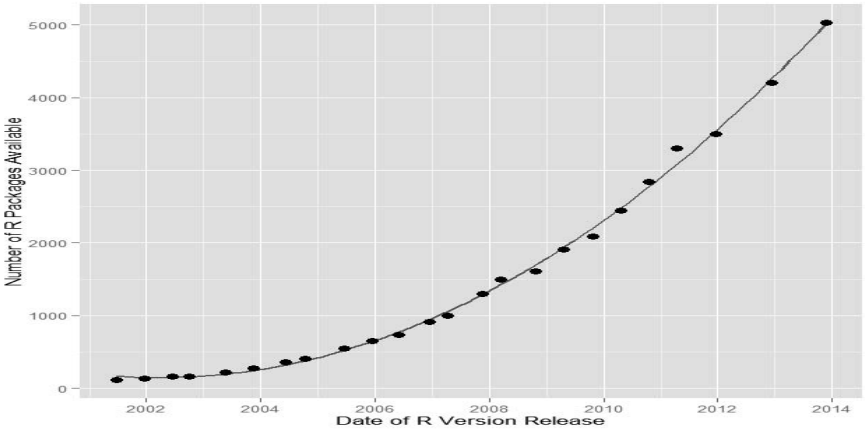
The evolution of R as the most used data analysis tool in the last decade is highlighted below by an analysis of R packages, given the fact that every package is a user contribution to the R system.

The packages can be found mainly on the R-project website: http://cran.r-project.org/web/packages/available_packages_by_name.html.

Figure 1 shows that the growth in R packages is following a rapid parabolic arc, specifically a quadratic fit with R-squared=.998 (Muenchen). The trend is more spectacular given the fact that it represents only CRAN packages, but not the packages from other seven repositories of R, such as Bioconductor.

**Available Packages on CRAN**

*Figure 2*

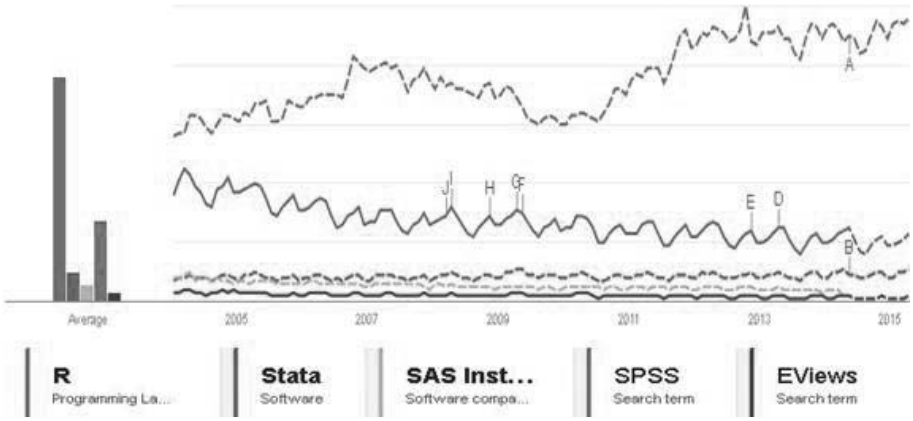



*Source: http://r4stats.com/articles/popularity/*

As prove of the spread of R, we have analyzed on Google Trends the popularity of Google searches for the most used statistical software: R, Stata, SAS, SPSS and Eviews.

**Interest over time of R, Stata, SAS, SPSS and EViews searches (2005-2015)**

*Figure 3*




*Source: Google Trends, http://www.google.com/trends/explore?hl=en-US#q=%2Fm%2F0212jm%2C%20%2Fm%2F05ymvd%2C%20%2Fm%2F01bp2d%2C%20SPSS%2C%20EViews&cmpt=q*

As seen in the Figure 3, starting from 2005, R had a tremendous increase comparing to the other analyzed Google searches. Point A on the graph is the starting point for forecasting. The forecasts show that R will keep in line until April 2015 (the limit date for forecast). Google Trends show also that the countries more searching for R in the period January-April 2014 are the following: Iceland, India, Senegal and South Korea.

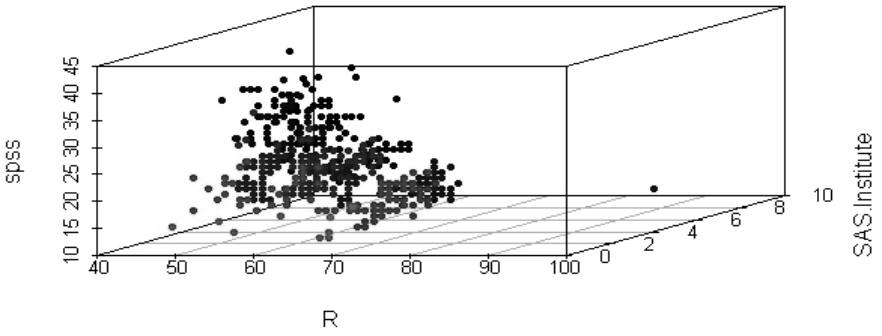
In the following statement we present three examples on R's 3d graphical capabilities, concerning the same data from Google Trends for the statistical software presented in the section above.

The first example is a 3d scatterplot created with *scatterplot3d* package (Ligges, 2003).

> s3d < -scatterplot3d(R, SAS.Institute, spss, pch=20, highlight.3d=TRUE, type="p")

**Interest over time of R, SAS and SPSS (2005-2014)**
**according to Google Trends**

*Figure 4*

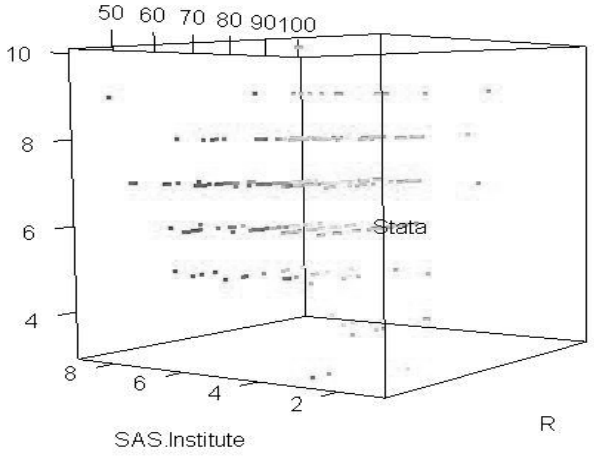



The second example is a 3d plot created with *rgl* package (Adler et. al, 2014).

> plot3d(R, SAS.Institute, Stata, col=rainbow(1000))

**Interest over time of R, SAS and Stata (2005-2014)**
**according to Google Trends**

*Figure 5*

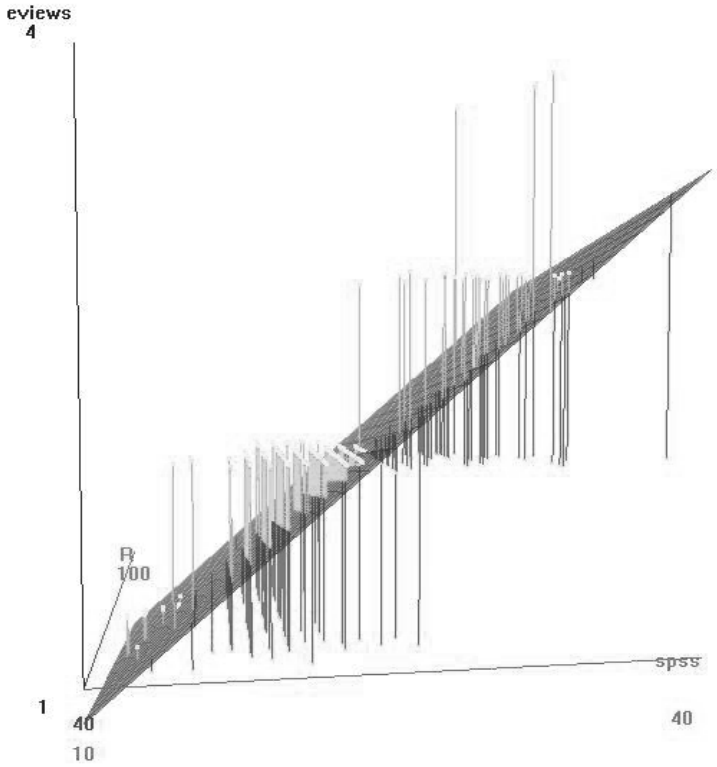



The third example is an interactive 3d plot created with *car* package ( Fox, 2011) that hasthe option to identify points by mouse clicking. This type of plot supports also regression models.

> scatter3d(x=R, y= eviews, z= spss, size = 10)

**Interest over time of R, EViews and SPSS (2005-2014) according to Google Trends**

*Figure 6*




Overall, the figures presented above show an ascendent trend of the interest of R versus other statistical software.

## CONCLUSIONS

R allows users and experts in specific fields of statistical computing and academic researchers to add new capabilities to the software. Is it not about writing new programs in R, but it is also convenient to combine related sets of programs, data, and documentation in R *packages*. More over, R is a full-fledged programming language, with a rich complement of mathematical functions, matrix operations and control structures.

In Romania is a small group in the official statistics involved in small area estimation based on R technique. This year has seen a definite increase

in R-omanian team activity by extending R use in the frame of the research institutes of Romanian Academy, also in universities and business field.

**Acknowledgement**

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

**References**

1. Adler, D., Murdoch, D. and others (2014) rgl: 3D visualization device system (OpenGL). R package version 0.93.996. http://CRAN.R-project.org/package=rgl
2. 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, pp.450-456
3. Fox, J., Weisberg S. (2011) An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion
4. Itunes Store, https://itunes.apple.com/gb/app/r-programming-language/id540809637?mt=8, accessed on 8th of April 2014
5. Ligges, U. and Mächler, M. (2003) Scatterplot3d - an R Package for Visualizing Multivariate Data. Journal of Statistical Software 8(11), 1-20.
6. Google Play, https://play.google.com/store/apps/details?id=appinventor.ai_RInstructor.R2, accessed on 8th of April 2014
7. Google Trends Data, http://www.google.com/trends/explore?hl=en-US#q=%2Fm%2F0212jm%2C%20%2Fm%2F05ymvd%2C%20%2Fm%2F01bp2d%2C%20SPSS%2C%20EViews&cmpt=q, accessed on 7th of April 2014
8. Muenchen B., The Popularity of Data Analysis Software, http://r4stats.com/articles/popularity/, accessed on 7th of April 2014
9. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
10. URL http://www.R-project.org/
11. Smith, D. (2011), Data Science: a literature review, available at: http://blog.revolutionanalytics.com/2011/09/data-science-a-literature-review.html, accessed on 8th of April 2014

**Trademarks**