Bringing New Opportunities to Develop Statistical Software and Data Analysis Tools in Romania

Nicoleta Caragea and Ciprian Antoniade Alexandru and Ana Maria Dobre

2012

BRINGING NEW OPPORTUNITIES TO DEVELOP STATISTICAL SOFTWARE AND DATA ANALYSIS TOOLS IN ROMANIA

Nicoleta CARAGEA
Senior Expert, National Institute of Statistics, Romania/Lecturer at Ecological University of Bucharest
Bucharest, Romania
nicoletacaragea@gmail.com
Ciprian Antoniade ALEXANDRU
Dean, Faculty of Economics, Senior Lecturer at Ecological University of Bucharest
Bucharest, Romania
alexcipro@yahoo.com
Ana-Maria DOBRE
Expert, National Institute of Statistics
Bucharest, Romania
dobre.anamaria@hotmail.com

Abstract

In the last decade, open source programming technology is widely used among statisticians for developing a new statistical software and data analysis. This is R software environment and the main objective of this paper is to underline the importance of R for statistical computations, data analysis, visualization and applications in various fields. Regarding to this, the paper is primarily intended for people already familiar with common statistical concepts. Thus the statistical methods used to illustrate the R performance are not explained in detail. The main intention is to offer an overview to get started, to motivate beginners by illustrating the flexibility of R, and to show how simply it enables the user to carry out statistical computations.

Keywords
R packages, programming language, statistics, data analysis, regression models, data visualization.

1. INTRODUCTION

The use of data analysis tools has usually a high inertia to change, often because the level of knowledge that young people receive in the frame of tertiary general programs, but also because of reasons related with the quality of the programming technology used in universities and tradition of teachers in keeping poor updated topics. Another cause, especially in Romania, but common also in other countries, is that most of private or public institutions use commercial statistical tools (generally with a predictable cost), the aversion to change them being very high.

Obviously, the existence of several data analysis techniques assumes implicitly different criteria of these products, based on several features, such as integrated functions in that software, respectively the ability to work with many kind of data input, in terms of data type or size of data series. Other criteria to select tools for data analysis is the product price, but also the diversity of required output (in terms of how the output or reports are showing the results, but also the details included, graphical aspect, or ease of learning and use it).

In these circumstances, it is necessary to re-shape the thinking at university level, changing the approach of the academic programs by presenting to students the opportunity to use the state-of-the-art programming technology for data analysis. In this way, graduates will be able to decide which tools are more appropriate for their purposes, depending on the need of institution they are working for.

In this paper we want to present the R software, one of the most popular data analysis tools developed by statisticians and that is currently continuous upgraded by a large community of researchers in academia and top business institutions.

To some people, R is just a simple letter in the Latin alphabet. Now is the time to find out what is R actually about. R is a statistical programming language based on S language created by two academicians in 1993 in New Zealand and released in 1996, Ihaka and Gentleman (1996) [4]. It is called R from the simple fact that both first names of the two creators start with R - Ross Ihaka and Robert Gentleman.

R is a free and open-source software, which can be downloaded from http://www.r-project.org. Currently it has more than two million of users worldwide and thousands of contributors, so it is a real global phenomenon. It could be called a global brainstorming.

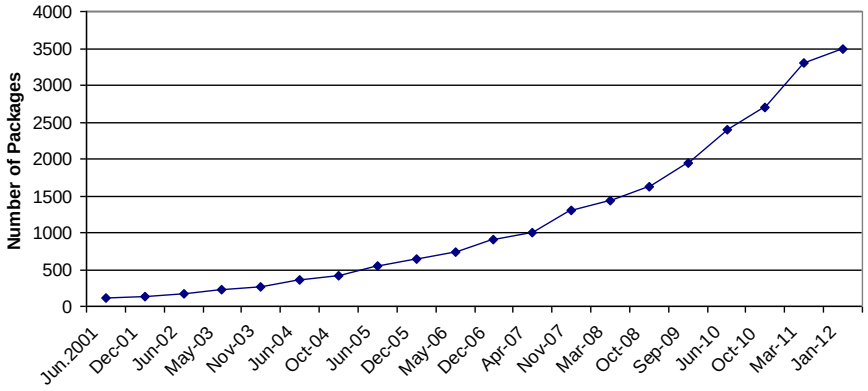
R was created by statisticians for statisticians, but in short time it spread in many other domains because of its unique ability to transform and to adapt.

Currently fighting with the three giants of the statistical computing – SAS, SPSS and Stata, R is in a continuous release and upgrade of component-based software. Some experts [12] have already forecasted that year 2015 will be the beginning of the end for SAS and SPSS.

This continuous process of evolving and improving raised better alternatives to the R base installation: RStudio, Deducer, Revolution Analytics, Red-R, JGR (Java GUI for R), SciViews-R. These GUIs (Graphical User Interfaces) are really user-friendly.

The capabilities of R are extended through packages, user-created add-on programs, which allow specialized statistical techniques, reporting tools, data-mining techniques etc. R also includes many pre-written procedures, called functions, which can be easily called in the packages. Every package is a research project that is reviewed at academic level [13]. The number of R packages available grew very quickly in the last years, as the following graphic shows on [14].

Fig.1 Available Packages in R



Sources: *http://r4stats.com/articles/popularity/ and http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Fox.pdf*

The packages can be found mainly on the R-project website: http://cran.r-project.org/web/packages/available_packages_by_name.html. To install a package it is necessary to start R, select a CRAN [15] mirror, from a list of software repositories that are scattered around the world. There is another easier way to initialize R avoiding this step: the user can only type *setInternet2(TRUE)* and after *install.packages ("packageYouWantToInstall")*. To call the function it is necessary to run *library ("packageYouWantToInstall")*. The users must pay attention on typing names of packages because the R is case-sensitive.
For the ones who want to migrate for other statistical softwares to R and need further assistance, it exists the following website: http://rconvert.com/.

2. LITERATURE REVIEW

Importance of using only one software which is able to perform all the stages of data analysis was firstly shown by Hodgess (2004) for some models in SAS and FORTRAN programming or a combination of Excel, FORTRAN and SAS. Currently, R software packages can make almost all type of data analysis, like preliminary plots, transformation, decomposition, Box-Jenkins models, sampling analysis, mapping, statistical regression, and forecasts.
A recent study [16] show that R has a very fast learning curve, after the first two courses students can perform analyzes on their data sets. This is due to the fact that programming language is similar to the C language, one of the most used in the world. Programming language usually helps to eliminate or reduce repetitive actions.
Learning facilities of software R are numerous, from instructions found on the package itself to online manuals, introductory textbooks, and textbooks Becker,

Chambers and Wilks (1988) [1], Spector (1994) [7], and Krause and Olson (2000) [5] or textbooks devoted to a particular issue, Modern Applied Statistics with S, Venables and Ripley (2002) [9], Venables and Ripley (2000) [8], Chambers and Hastie (1992) [2], Pinheiro and Bates (2000) [7], and Zivot and Wang (2002) [11]. Because the software is free, noncommercial, it can be installed on multiple computers, and students and professionals can use it on personal computers.
A software package shows the advantages of an open source system: independence from a seller of the product, lower costs, given the lack of a cost to purchase, easy customization, and usage of technical support provided by the existence of a large community of users and of specific blogs. Another advantage consists in the reuse of the source code developed by other specialists licensing under the GNU General Public License type, thus reducing analysis time and enables the specialist to allocate time bringing a whole new contribution given the specific analysis own.
R's popularity has grown in recent years and the trend is favorable, the estimations showing that in about three years will exceed the number of users of SAS and SPSS. Regarding the number of users of applications for analysis, data mining and software for large databases, for the period May 2010-May 2012, R was ranked first by 30% of respondents (Muenchen, 2012) [6].

3. SWOT ANALYSIS FOR R PROJECT

Strengths
Open-source program
Freeware, the cost of using R are related only with training of users
Free and open-source GUIs and IDEs [17]
Working of various operating systems: Windows, Linux, Mac OSX
Easy to install and configure
A fantastic user community that keeps growing
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
User support through a very active mailing list, blogs, dedicated forums
Used for statistical computations, data analysis, visualization and exciting applications in various fields
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
Excel integration via RExcel; SPSS has not this issue available
Competitive tools for Geographic Information Systems
Interactive Graphs; SPSS has not this issue available
Internet Control; SPSS has not this issue available
Operations Research; SPSS has not this issue available
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 (it benefits those organizations, so the vendors say they should have their own software license)
The flexibility to mix-and-match models, scripts and packages for the best results
The possibility to transform R code into HTML code so that it could be used on user's website [18].

Weaknesses
Data collection should be available from other tools; MySQL or PostgreSQL are popular among R users for this purpose
Direct Marketing not available
Guided Analytics not available
The help files and the vignettes for packages are written for relatively advanced users
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

Opportunities
Users' contribute to program's ongoing development
Share new techniques with other R users around the world via online community [19]
Re-use and reproduce new discovered techniques on analytic operations that the user is going to perform – this is difficult in SAS or SPSS
Very large area of use - statistics, journalism, mapping, finance, forecasting, social networking, drug development, computational biology, life sciences and many more
Easy to export data to usual formats and get data visualization like maps, 3D surfaces, image plots, scatter plots, histograms, bar plots, pie charts, multi-panel charts

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

4. STATISTICAL ANALYSIS WITH R

This part of the paper is primarily intended for people already familiar with common statistical concepts. Thus the statistical methods used to illustrate the R performance are not explained in detail. The main intention is to offer an overview to get started, to motivate beginners by illustrating the flexibility of R, and to show how simply it enables the user to carry out statistical computations.

4.1 The simple linear regression model

A basic model for this is a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The *y* variable is called the response variable and the *x* variable the predictor variable, covariate, or regressors. As a statistical model, this says that the value of y_i depends on three things: variables xi, the function $\beta_0 + \beta_1 x_i$, and the value of the random variable ε_i. The model explain that for a given value o f *x*, the corresponding value of *y* is found by first using the function on *x* and then adding the random error term ε_i.

To be able to make statistical inference, we assume that the error terms, ε_i have a normal distribution. This assumption can be rephrased as an assumption on the randomness of the response variable. If the x_i values are fixed, then the distribution of y_i is normal with mean μ and variance σ^2. If the x_i values are random, the model assumes that, conditionally on knowing these random values, the same is true about the distribution of the y_i.

4.2 Estimating the parameters in simple linear regression

Before using R to find estimates, we need to explain how R represents statistical models. Linear models are fit using R's model formulas.

The basic format for a formula is the ~ (tilde) is read "is modeled by" and is used to separate the response from the predictor(s). The response variable can have regular mathematical expressions applied to it, but for the predictor variables the regular notations +, −, *, /, and ^ have different meanings.

One goal when modeling is to "fit" the model by estimating the parameters based on the sample. For the regression model the method of least squares is used.

The method of least squares finds values for the β that minimize the squared difference between the actual values, y_i, and those predicted by the regression function.

The simple linear regression model for y_i has three parameters, β_0, β_1 and σ^2. The least-squares estimators for these are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \bar{x}$$

$$\sigma^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

To find the estimates, it is used the lm() function. The basic usage of lm function is of the form:
lm(formula, data=…, subset=…).
As is usual with functions using model formulas, the data= argument allows the variable names to reference those in the specified data frame, and the subset= argument can be used to restrict the indices of the variables used by the modeling function.
By default, the lm() function will print out the estimates for the coefficients. Much more is returned, but needs to be explicitly asked for, as an example below:

```
> summary(fit.lm <- lm(y ~ x1+x2+x3 + x4, data=dataset ))

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-14.5325  -0.1515  -0.1515  -0.1515  29.4675

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.515e-01  3.804e-02   3.984 6.92e-05 ***
X1          1.187e-03  9.893e-05  12.001  < 2e-16 ***
X2          1.781e-03  2.496e-04   7.136 1.18e-12 ***
X3          4.252e-03  2.672e-04  15.914  < 2e-16 ***
X4          3.292e-03  4.866e-04   6.765 1.58e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.044 on 3176 degrees of freedom
Multiple R-squared: 0.5516,      Adjusted R-squared: 0.551
F-statistic: 976.8 on 4 and 3176 DF,  p-value: < 2.2e-16
```

The decomposition of the total sum of squares (SST) into the residual sum of squares (SSE) and the regression sum of squares (SSR) allows us to interpret how well the regression line fits the data. If the regression line fits the data well, then the residual sum of squares will be small. If there is a lot of scatter about the regression line, then RSS will be big. To quantify this, we can divide by the total sum of squares, leading to the definition of the **coefficient of determination:**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

This is close to 1 when the linear regression fit is good and close to a when it is not. When the simple linear regression model is appropriate this value is interpreted as the proportion of the total response variation explained by the regression. That is, $R^2 \cdot 100\%$ of the variation is explained by the regression line. When R^2 is close to 1, most of the variation is explained by the regression line, and when R^2 is close to 0, not much is. **The adjusted** R^2 divides the sums of squares by their degrees of freedom. For the simple regression model, these are $n-2$ for RSS and $n-1$ for SST. This is done to penalize models that get better values of R^2 by using more predictors. This is of interest when multiple predictors are used.

4.3 Multiple linear regression

Multiple linear regression allows for more than one regressor to predict the value of *y*. These regressors may be separate variables, products of separate variables, powers of the same variable, or functions of the same variable. In the next example, we will consider regressors that are not numeric and categorical. They all fit together in the same model, but there are additional details. Much of the background for the simple linear regression model carries over to the multiple regression models.

Let *y* be a response variable and let x_1, x_2,…, x_m be *m* variables that we will use for predictors. For each variable we have *n* values recorded. The multiple regression model we discuss here is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{m1i} + \varepsilon_i$$

There are $m+1$ parameters in the model labeled β_0, $\beta_1, \ldots, \beta_m$. They appear in a linear manner, just like a slope or intercept in the equation of a line. The x_i are predictor variables, or covariates. They may be random; they may be related, such as powers of each other; or they may be correlated. It is assumed that the ε_i values have a normal distribution with mean 0 and unknown variance σ^2. If the x_i variables are random, this is true after conditioning on their values.

Multilevel models, or mixed effect models, can easily be estimated in R. Several packages are available. Here, the lme() function from the nlme-package is described. The specification of several types of models will be shown, using a fictive example.

The lme() function fits a linear mixed-effects model in the formulation described in Laird and Ware (1982) but allowing for nested random effects. The within-group errors are allowed to be correlated and/or have unequal variances.

```
fit.lme <- lme(y ~ x1+x2+x3+x4+x5+x6, data = dataset, random = ~1 | id_number))
```

```
Linear mixed-effects model fit by REML
 Data: dataset
       AIC      BIC    logLik
  36839.58 36894.15 -18410.79

Random effects:
 Formula: ~1 | id_number
        (Intercept) Residual
StdDev:    7.412778 78.26072

Fixed effects: y ~ x1 + x2 + x3 + x4 + x5 + x6
                 Value Std.Error   DF  t-value p-value
(Intercept)  1.8467629 1.8745149 3133  0.98520  0.3246
X1          -0.0606849 0.0196211 3133 -3.09284  0.0020
X2           1.0739386 0.0194280 3133 55.27781  0.0000
X3           0.1848692 0.0221620 3133  8.34170  0.0000
X4          -0.0559870 0.0106834 3133 -5.24055  0.0000
X5           1.0602459 0.0110052 3133 96.34071  0.0000
X6           0.1208277 0.0193596 3133  6.24124  0.0000
 Correlation:
         (Intr)    x1     x2     x3     x4      x5
X1       -0.045
X2        0.000 -0.793
X3       -0.028 -0.016 -0.395
X4        0.031 -0.760  0.594  0.143
X5       -0.043  0.566 -0.634  0.068 -0.861
X6        0.001  0.122  0.048 -0.351  0.037 -0.417

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-5.90073521 -0.07554414 -0.02012171  0.01091294 20.15701935

Number of Observations: 318
Number of Groups: 24
```

The output from the summary method for lme objects consists of several parts:

- The first part of the output gives the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), which can be used for model selection, along with the log of the maximized restricted likelihood.
- The second one displays estimates of the variance and covariance parameters for the random effects, in the form of standard deviations. The term labeled Residual is the estimate of σ.
- The table of fixed effects is similar to output from lm; to interpret the coefficients in this table, in particular:
    - The fixed-effect intercept coefficient $\hat{\beta}_1 = 1.8467629$ represents an estimate of the average level of dependent variable y when all factors take null values.
    - Given the parameterization of the model, the coefficient for x1, $\hat{\beta}_2 = -0.0606849$, represents the relationship of average y and

the factor x1. The negative sign explain the inverse correlation between variables.

- The part of the output labeled Correlation gives the estimated sampling correlations among the fixed-effect coefficient estimates, which are not usually of direct interest. Very large correlations, however, are indicative of an ill-conditioned model.
- Some information about the standardized within-group residuals, the number of observations, and the number of groups, appears at the end of the output.

Along this part of paper it is presented a relatively brief description of how to conduct statistical analyses using linear regressions in R. The paper could be a starting point to provide students and researchers in many disciplines means of using R to analyze their data.

5. CONCLUSION

The free available, powerful and convenient computing of data using R packages has revolutionized the practice of statistical data analysis worldwide. Except of some official statistical systems like those in Italy, Austria, Australia, Canada the international using R as a main tool, there are many companies that are using it, including Pfizer, Shell, Facebook, Google, Mozilla, Times, The New York Times, The Economist, NewScientist, Lloyd's, Bing, Johnson&Johnson [20].

As an extra accreditation for R, it is appreciated the statement of Norman Nie, co-founder of SPSS in the in the late 1960's: "R is the most powerful and flexible statistical programming language in the world". Currently, Nie is CEO [21] and president of Revolution Analytics, a company that provides commercialized versions of R programs [22]. Revolution Analytics is willing to grow the R community through sponsorship of the Inside-R.org community website, funding worldwide R user groups and offering free licenses of Revolution R Enterprise to everyone in academia. Recently, Revolution Analytics won the DataWeek Awards for category of "Data Science Technology", being selected as a top innovator in this field [23].

The blogging R phenomenon is so vast that it exists even a specialized website with R news and tutorials contributed by 393 [24] R bloggers: http://www.r-bloggers.com/.

R allows users and experts in specific fields of statistical computing 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*.

Unfortunately, Romania it is not on the current available users list worldwide [25], but it is not late to make this happen. Our country has very good and competitive computer scientists, which could become useRs. For the moment, there is a small group in the official statistic involved in small area estimation based on R technique.

References

[1] Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New **S** Language: A Programming Environment for Data Analysis and Graphics*, Pacific Grove, CA: Wadsworth and Brooks Cole.

[2] Chambers, J.M, and Hastie, T.J. (1992), *Statistical Models in **S***, Pacific Grove, CA: Wadsworth and Brooks Cole.

[3] Hodgess, E. (2004), "A Computer Evolution in Teaching Undergraduate Time Series", *Journal of Statistics Education* Volume 12, Number 3 (2004), www.amstat.org/publications/jse/v12n3/hodgess.html

[4] Ihaka, R., and Gentleman, R. (1996), ``**R**: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299-314.

[5] Krause, A., and Olson, M. (2000), *The Basics of **S** and **S-Plus***, New York: Springer.

[6] Muenchen, R., (2012), The Popularity of Data Analysis Software, http:// http://r4stats.com/articles/popularity/

[7] Pinheiro, J.C., and Bates, D.M. (2000), *Mixed-Effects Models in **S** and **S-Plus***, New York: Springer. Spector, P.C. (1994), *An Introduction to **S** and **S-Plus***, Belmont, CA: Duxbury.

[8] Venables, W.N., and Ripley, B.D. (2000), ***S** Programming*, New York: Springer.

[9] Venables, W.N., and Ripley, B.D. (2002), *Modern Applied Statistics with **S-Plus*** (4^{th} ed.), New York: Springer.

[10] Venables, W.N., Smith, D.M. and the **R** Development Core Team (2003), *An Introduction to **R***, London: Network Theory Limited.

[11] Zivot, E., and Wang J. (2002), *Modeling Financial Time Series With **S-Plus***, New York: Springer-Verlag.

[12] http://r4stats.com/2012/05/09/beginning-of-the-end/

[13] There is a specific procedure to approve the implementation of the package in R's environment

[14] http://r4stats.com/articles/popularity/

[15] CRAN (The Comprehensive R Archive Network) is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R

[16] http://r4stats.com/articles/popularity/

[17] Integrated Development Environment

[18] http://www.inside-r.org/pretty-r

[19] http://www.dataweek.co/index/winners

[20] http://www.revolutionanalytics.com/what-is-open-source-r/companies-using-r.php

[21] Chief Executive Officer

[22] Smith D., *R is Hot.* (2010) from www.revolutionanalytics.com/R-is-Hot/

[23] http://www.dataweek.co/index/winners

[24] This number is available for the current date 30^{th} of August 2012

[25] http://rwiki.sciviews.org/doku.php?id=rugs:r_user_groups