Overcoming R angst. The tools that help statisticians learn and use R effectively

Adrian Oţoiu (adrian.otoiu@csie.ase.ro) - corresponding author
Bucharest University of Economic Studies, Department of Statistics and Econometrics

Emilia Ţiţan (emilia.titan@csie.ase.ro)
Bucharest University of Economic Studies, Department of Statistics and Econometrics

ABSTRACT

This paper makes a review of the most popular graphical user interfaces (GUIs) in order to help new and unexperienced users learn and use R. Although R is free and benefits from state-of-the-art implementations of statistical algorithms and methods, it is not the first choice among statistical software users. This can be attributable to a lack of user-friendly interfaces similar to the ones available for other statistical packages. However, there are several GUIs that enable R users to use it quickly and effectively without having to use interactive programming and learn R code. For standard statistical analysis, RCommander proves to be the best solution that not only offers standard functionalities that are similar to commercial statistical software, but allows users to visualize, use, and customize the codes that perform standard data and statistical procedures, thus helping them to effectively learn R. For more advanced users, there are two main specialized GUIs RStudio, which facilitates the use of R through the use of scripts and generation of interactive applications and reports, and Rattle, specialized for data mining procedures and algorithms.

Keywords*: software, open source, R, GUI.*
JEL classification*: C87*

1. INTRODUCTION

R has distinguished itself as one of the most powerful statistical packages available. R has become the dominant analytics package as the share of analytics professionals using it has increased from 20% in 2007 to over 70% in 2015 (Rexer, 2015), being also the primary tool for 36% of the analytics professionals, far ahead from SAS, IBM SPSS Modeler and Knime, with 7% each (Rexer, 2015).

Its versatility and advanced capabilities are often superior to mainstream tools such as SPSS, SAS, Stata, EViews, not only in terms of being able to accommodate different data formats and statistical methods, but also to produce superior graphs. However, its steep learning curve is preventing

many users to adopt it and makes them revert, or keep on using, other user-friendly, commercial software.

While having these qualities and being available free of charge recommends R as the ideal choice for statistical analysis, its basic interface, a console with minimal options, prevents most users from learning and using it. The Rexer Analytics 2015 Data Science Survey shows that, although R is by far the most popular analytics software, and over 70% of analytics professionals use R, more than half of them do not use it as their primary analysis tool (Rexer, 2015).

However, the widespread use of several Graphic User Interfaces (GUI), which offer several alternatives to the then dominant use of R command line, have reduced the share of the latter from 51% in 2011 to only 17% in 2015, with the top R interface, R Studio, being used by 68% of R users in 2015.

Today, several R GUIs offer extensive point-and-click functionalities, which can help flatten, or even skip, the learning curve and allow users to take advantage of Rs diverse powerful capabilities in much the same way as in other commercial statistical packages. In this paper, we attempt to discuss the most popular R interfaces available, and focus on some powerful capabilities that are often not available outside the chosen interfaces.

2. LITERATURE REVIEW AND GENERAL FRAMEWORK

A survey of data miners prepared annually by Rexer Analytics puts R and its popular interfaces at the top of the most used analytics tools available, with 76% of the analytics professionals using it, also being the primary tool for 36% of them (Rexer, 2015). Also, the same source shows that, between 2011 and 2015, about 1/3rd of the users have switched to a GUI tool and abandoned using R console in their work in favor of one of its most popular interfaces (Rexer, 2015).

The primary graphical interface that is the most commonly used R environment is RStudio. Except for the command line, the next most popular interfaces are RCommander, Emacs and Knime. Among graphical user interfaces in R, we distinguish general GUI interfaces, RCommander, Deducer and RKWard (Valero-Mora and Ledesma, 2012), and specialized interfaces that are either for programming (e.g. gWidgets, ESS/Emacs) or extensions of other software packages into R, among which Knime, Statistica, IBM SPSS Suite (Rexer, 2015).

Among the most popular analytics software, R is perceived as one of the most difficult ones, with a steep learning curve (Muenchen, 2014), due to a variety of reasons, among which: 1) Help files are not always designed for beginners, 2) Misleading function or parameter names, 3) Inconsistent syntax, 4) Difficult manipulations of variables and data management, 5) Unformatted output and different output formats, 6) Lack of GUI.

Also, while R is the most used analytics software, 64% of analytics professionals still use other software as their primary work tool (Rexer, 2015).

3. A STOCK-TAKING OF R INTERFACES

Based on our review, there are three types of GUI's for R: interfaces to use R from different programs, interfaces for completing programming tasks in R, and, the most useful type, dedicated GUI interfaces, a classification based mostly on the work of Valero-Mora and Ledesma (2012).

As the scope of the current paper is to examine the interfaces that enable analysts and statisticians to use R in much the same way as commercial software packages are (e.g. SPSS, SAS, Stata, EViews), we will focus on the dedicated GUI Interfaces that can help harness the power of R, make the R learning curve smoother, and are available for free.

The interfaces to be examined are RStudio, the most popular R environment, RCommander, an interface similar to commercial software packages, and Rattle, an interface dedicated for performing data mining tasks.

3.1. Essential capabilities of interfaces

All GUIs have basic capabilities built in that support customary operations. First of all, there are file manipulation capabilities, that enable importing, exporting, visualization and editing of data files, and allow the use of different formats, among which text, CSV, SAS, SPSS and Stata.

The modular structure of R, which require installing and loading specialized packages to perform specific operations, is integrated in these GUIs. RStudio and Rcommander have quick and easy package management functionalities. Rattle makes an exception as it does not offer explicit loading and dropping packages.

In addition, all of them feature windows which display the code used to perform needed commands, and a console which displays the results of commands and operations performed.

Data used can be visualized in convenient ways in all GUIs, allowing for quick examination of data, finding out their properties, changing them into different types (e.g. text to numeric, numeric to factor levels, etc.) and convenient editing if needed.

3.2. Distinctive features and specialisations

What makes the difference between the three interfaces, RStudio, RCommander and Rattle, are their specific capabilities, the targeting for specific needs, and focus for different user types and levels of proficiency.

RStudio is essentially programming-oriented, providing the options of creating scripts, projects and documents. It also has useful capabilities for importing and exporting files, managing datasets, display of the workspace structure, plots, packages and command history (see Figure 1).

This fits the initial purpose of RStudio, designed as an interface to improve the use of R through several features aimed at enhancing productivity through coding and debugging facilities. Everything else that RStudio offers revolves around these features, which aim at providing enhanced functionalities to easy manage the workspace, see available files and variables, packages that are available and can be activated or deactivated, and create documents in HTML, pdf and Word, and interactive applications using the package Shiny.

So far RStudio has emerged as the most popular GUI for R. Its share among the R tools has jumped from 8% in 2011 to 62% in 2015 and has become the dominant environment to use R (Rexer 2015).

RCommander is an point-and-click interface (Fox, 2017), which renders some of the essential capabilities of R and allows users focus more on statistical work without having to learn the specific R command for performing statistical analysis (Fox and Bouchet-Valat, 2016).

File and data management capabilities enable point-and-click data imports from a variety of formats (text, csv, Excel, Stata, SPSS, Minitab and SAS), merge datasets, subsetting datasets based on user-defined criteria, transforming and ordering variables, exclude cases with missing data, etc.

RStudio interface overview

Figure 1
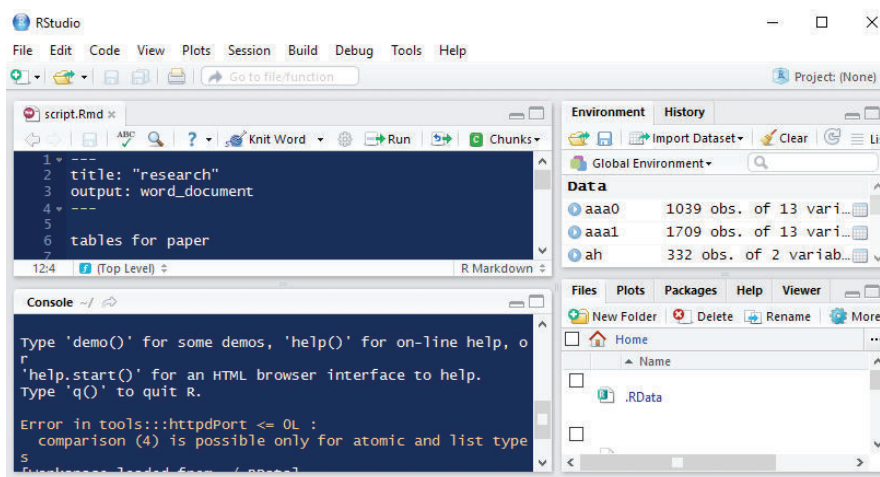
Source: RStudio software.

Beside the essential file and data management capabilities, and package management capabilities, RCommander has a collection of statistical functions, graphs, modelling options, statistical tests, and distributions. They cover most of the statistical tools and procedures required for regular quantitative work: summaries, tests for means, variances, and normality of distributions, ANOVA, non-parametric tests, correlations, and contingency tables. Among the advanced statistical methods, RCommander offers convenient implementations of regression analysis, GLM models, ordinal regression models, multinomial logit, factor analysis, cluster analysis and principal components analysis. Scale reliability for testing composite index components is also available following an implementation that is unique to RCommander. Besides a complete set of graphs, it also has a strong modelling set of options, which includes coefficients testing, model selection tools which implements stepwise algorithms, specification and diagnostic tests, among which Durbin-Watson, Breusch-Pagan, VIF factors, and diagnostic plots.

To sum up, RCommander is the closest GUI to established commercial software, offering comparable tools and options for data and statistical analysis. Another feature that is not easily accessible is the possibility to customize menu options to include commands that are not currently available, or new procedures. There are also some plugins, such as RcmdrPlugin.BCA, which reconfigures the menu to make it amenable for performing business analytics tasks, and RcmdrPlugin.epack, which implements time series analysis, added as two main menu options in addition to the regular RCommander plugin.

RCommander Interface
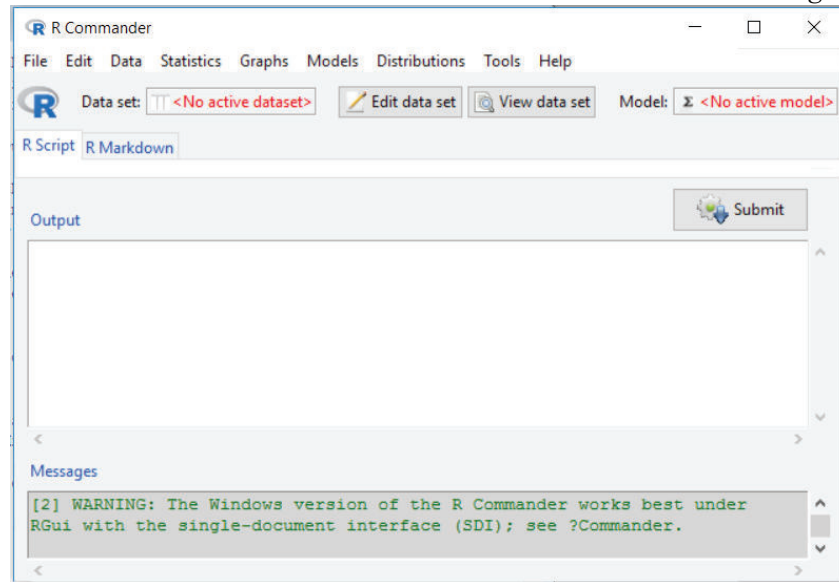
Source: RCommander package

Rattle is a data mining interface, which, aside from automating data mining tasks through a simple, easy-to-use interface, also provides a stepping stone for the use of R for advanced data analysis (Williams, 2009).

Its purpose is somewhat similar to RCommander, as it aims to provide a user-friendly environment for data mining and modelling. However it appears to be a lot more automated, customized for data mining, starting from data import, and continuing with data exploration, testing and transformation. Three tabs, cluster, association and model provide a wide range of modelling options, which are tested for their performance in Evaluate tab. All operations appear in logs with comments, and scripts can be exported so as to be used later on (Williams, 2011), either in R Console or in RStudio (Togawave, 2015).

Although it makes available a lot of point-and-click capabilities, Rattle is not very flexible in terms of customization; thus, besides data miners and modelers, Rattle remains off limits, or presents little interest for regular data analysis or statistical work. Another drawback are the recent issues concerning its recent smooth installation and operation, which can be successfully solved by power users but remain off-limits for regular researchers. Also, for more advanced work, the company that produced Rattle, Togawave, recommends

the use of RStudio to create and run data mining tasks that go beyond the capabilities offered by Rattle (Togawave, 2015).

Rattle interface

Figure 3
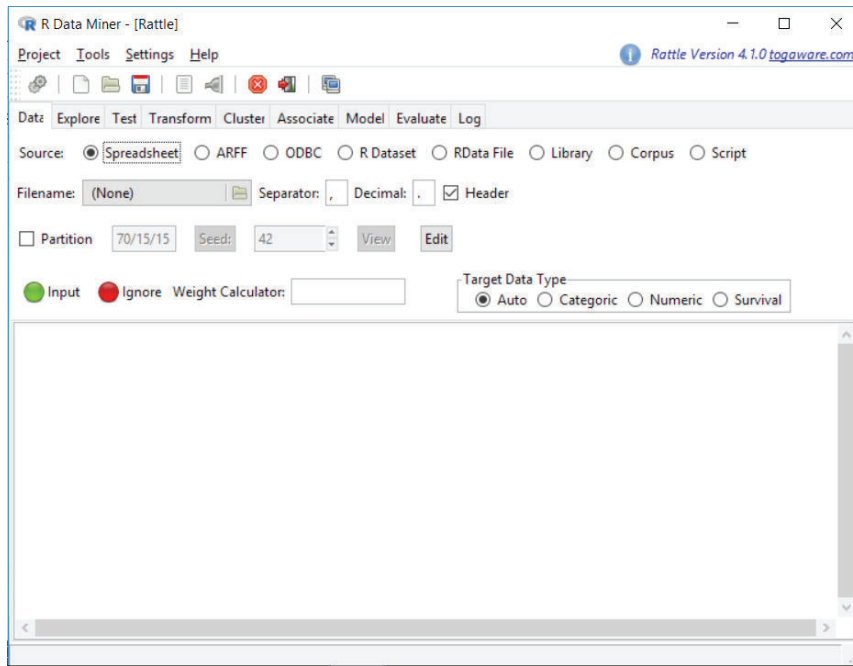
Source: RCommander package

3.3. Strenghts and weaknesses for each interface

A SWOT analysis can shed light on each of the examined interfaces so far, putting in perspective all advantages and shortcomings of each of the examined interfaces.

SWOT analysis of the analytics GUI for R

Table 1

	Strengths		Weaknesses
RStudio	One of the best development environments Allows production of applications and documents in different formats (Rpubs, Word, Pdf, Html) Allows free posts of R output and files in RPubs Quick manipulation and selection of active files and packages in right-hand windows	RStudio	Except basic file and package manipulation, requires knowledge of R code. Issues with graphs as size of graph window prevents their proper display Less user-friendly than most commercial packages, RCommander or Rattle
RCommander	Does not require prior knowledge of R Allows performing most of the operations available in commercial software packages Model selection tools and capabilities Customizable interface for specific tasks (e.g. time series, business analytics) Parameter changes of menu functions	RCommander	Changes of active files difficult Undloading packages is done using RCode Poor integration of standard R commands which are not always functional Customization of interface unproperly documented Lack of development effort comparable to the other two GUI's
Rattle	Integrated layout for performing data mining Detailed data mining framework Implementation of social analysis and natural language processing solutions	Rattle	Poor flexibility for regular data analysis and statistical work Cumbersome development of advanced implementation Issues governing proper installation and functionalities

SWOT analysis of the analytics GUI for R

Table 1 (continued)

	Opportunities		Threaths
RStudio	Development excellence Integration with other R tools and other analytics software Addition of user-friendly functionalities to enlarge user base	RStudio	Other user-friendly interfaces and proprietary solutions
RCommander	Becoming the standard entry-level R environment Having a larger collection of plug-ins that can automate a larger number of R commands, and enable performance of latest analysis tools (e.g. natural language processing, social media analysis) Support-based solutions for less tech-savvy users Capabilities available for advanced users such as document and application production	RCommander	Lack of developmental advances Features and evolution of other open-source software offering similar user-friendly analytics solutions Lack of appropriate popularization of its features and capabilities
Rattle	Improvements in flexibility of uses for data analytics and statistical work Widening the target users range to include statisticians and data analysts which do not need the standard data mining framework Constant acquisition and integration of latest analytics methods	Rattle	Other open-source data mining solutions offering Poor support for bugs for regular users may undermine its usability for less tech-savvy users to which it addresses Improved basic functionality to make installation and use more easy and bug-free

Source: authors' compilation

Some important considerations must also be made, to complement the essentially concise nature of the SWOT analysis.

RStudio requires knowledge and a good mastery of R commands, and can be thought of as a programming environment rather than a tool to learn R. Although it helps learn R through its interactive script window, the user still has to learn each and every command used using R reference sources, among

which the most important are: package documentations, help files, books, among which UseR! Springer series, and specialized forums. Thus, it is not much help for R beginners and less tech-savvy users, and its emergence as the main interface for R does not ensure that R will become the preferred software used for data analysts and data scientists.

Rattle is a point-and-click interface with powerful data mining capabilities. Its substantial tools for data processing and manipulations can be of help for intensive data mining and modelling work, which is well supported by a wide range of modelling options, among which: cluster analysis, association analysis, decision trees, linear regression, SVM, neural networks. Also, it allows testing of the obtained models on different splits of the original data, using advanced tools (e.g. ROC curves, lift, etc.). However, for more advanced work, commands have to be exported and modified/used separately, either in the R console or in RStudio. Also, its recent issues are likely to affect its popularity and chase away users towards similar open-source data mining solutions (e.g. RapidMiner, Weka).

RCommander is also a point and click interface, which exhibits the highest degree of flexibility and versatility compared to the other two interfaces. Its two-window interface allows the user to see immediately the script generated by the point-and-click commands, and the corresponding output in the window below it. Based on the R documentation, the user can modify the parameters of the GUI-generated commands to suit her needs. Its commands range from data processing commands (importing, variable transformation, merging datasets, deleting observations, etc.), to statistical tests and models, graphs, model testing, selection and comparison, and statistical distributions. RCommander also benefits for dedicated built-in commands, such as Scale Reliability, and from plugins that change the menu options to make it suitable for particular types of analyses. Among the most well-known plugins are epack, with time series modeling options, and BCA for customer and business analytics, which has data mining capabilities similar to Rattle. Other specialized plugins are available at http://www.rcommander.com. So far it is the best environment that can smoothen the steep learning curve for R and make it accessible to wide range of users.

4. CONCLUSION

Although R is a programming-oriented language, there are tools that make its use fairly simple and similar to commercial software packages such as SPSS, STATA, etc. These tools offer a wide range of options when it comes to how and how fast much one wants to learn R and, at the same time, do regular statistical and data analytics work in an easy way comparable with dedicated statistical packages.

For learning R, the most helpful interface is RCommander. With a lot of import and data manipulation options, it allows the user to see the scripts used with point-and-click commands and modify them based on options available, or to use pre-defines scripts loaded with the file menu. The specific plug-ins, and options to customize RCommander makes it the best starting point for any user without a programming background, who is accustomed to using commercial software packages, to do statistical work and at the same time learn R. However, its layout is not meant to run larger scripts, however, this can be overcome by using it under RStudio.

For data mining professionals, Rattle may be a better option since it offers an integrated environment that allows the use of selected R commands with no knowledge of R required. However, the transition from point-and-click to script-based work is not as straightforward as in RCommander, therefore the learning curve is still steep and cumbersome for those without any programming background and skills, and it may actually induce a gap that prevents the average user make the transition. Its occasional issues with respect to its installation and use may pose a threat to its future as a viable open-source data mining environment.

RStudio is an interface for users who already have knowledge of R, or want first to learn R and perform statistical/analytical work as they gain the necessary knowledge. Its capabilities make it the best choice for using R for experienced users. RStudio can also integrate RCommander or Rattle, facilitating the transition to advanced use for those who learned to use R with either of, or both of, these two interfaces.

Based on our review, we believe that the three GUIs are sufficient for overcoming the difficulties of using R, depending on the types of users and on the specific tasks to be performed. In the case of RCommander, its functionalities and features are very similar to other commercial statistical packages. However, for advanced users, these interfaces are not a perfect substitute for actually learning R and become familiar with its commands and options. Nevertheless, learning commands and code for statistical software is also needed for other leading commercial software (e.g. SAS, SPSS) for the purpose of carrying out advanced statistical and analytics work. Therefore, the limitations of the three GUIs for R should not be seen as a major disadvantage, but rather as a state of affairs that is consistent with making the transition to advanced use straightforward and accessible, perhaps more so than other commercial packages that do not benefit from specialized information disseminated via online documentation and forums, and an user community as helpful as the one for R.

References

1. **Fox, J**., 2017, U*sing the R Commander: A Point-and-Click Interface for R*. Chapman & Hall/CRC Press
2. **John Fox**, **Milan Bouchet-Valat**, 2016, *Getting Started With the R Commander*, available at http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf, accessed on February 3, 2017
3. **Pedro M. Valero-Mora, Ruben Ledesma**, 2012, *Graphical User Interfaces for R*, Journal of Statistical Software vol. 49, issue 1, pg. 1-8.
4. **Rexer, K.**, 2015, *Rexer Analytics 2015 Data Miner Survey*. Retrieved from www.RexerAnalytics.com on January 31, 2017
5. **Muenchen, R**., 2014, *Why R is hard to learn*. Available at http://r4stats.com/articles/why-r-is-hard-to-learn/ accessed on February 1, 2017.
6. **Togawave**, 2015, *Rattle: Data Mining by Example*, available at , accessed on February 4, 2017.
7. **Williams, Graham J**, 2009, *Rattle: A Data Mining GUI for R*, The R Journal Vol. 1/2, pg. 45-55.
8. **Williams, G. J.**, 2011, *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*, Use R!, Springer.