CONTEMPORARY DATA SCIENCE FOR FINANCE STUDENTS. ESSENTIAL FEATURES OF COMMONLY USED STATISTICAL SOFTWARE - A COMPARATIVE STUDY

Mădălina Viorica ION (MANU)
Bucharest University of Economic Studies, Romania
mvmadalina@yahoo.com

Ilie VASILE
Bucharest University of Economic Studies, Romania
vasile.ilie@fin.ase.ro

Abstract
This paper inventories some of the essential traits of the software preferred by researchers, students and professors, such as R or RStudio, or Matlab and also their possible utilizations. In order to fill the gap in the Romanian literature and help finance students in choosing proper tools according to the research purpose, this comparative study aims at bringing a fresh, useful perspective in the relevant literature. In Romania, the use of R was the focus of several international conferences on official statistics held in Bucharest, and others having business excellence, innovation and sustainability as purpose. In this time, at global scale, R and Python programming languages are considered the lingua franca of data science, as common statistical software used both in corporations and academia. In this paper, I analyze basic features of such software, with the purpose of application in finance.

Keywords: software, programming language, data science, R, RStudio, features, Matlab, statistical.

1. INTRODUCTION

This paper is a comparative study of the most used software instruments for modeling, in an attempt to understand their main features and then define a model for the enterprise value.

Comparisons of finance problem solving using MATLAB[1], R or Eviews[2] represent an area of research that is not yet developed in Romania, although small steps have been made in this direction (Simionescu 2014); the use of such software in applications in Romania is way behind the current researches. In Romania, the use of R was the central theme of several international conferences on

[1] Students use MATLAB – to analyse data develop algorithms and modeling - and Simulink for projects and competitions while solving practical challenges in robotics, fuel-efficient cars and autonomous vehicles, and for research (Mathworks n.d.).
[2] EViews 6 Student Version is the premier forecasting and analysis package for Windows-based computers, with a wide range of statistical and graphical techniques (Quantitative Micro Software 2007).

official statistics, organized by R-omania Team in Bucharest[3] since 2013 (R Project R-omania Team n.d.).

This comparative study aims to identify some of the particularities of the software tools for modeling currently used by researchers or students, such as R language / RStudio, or MATLAB[4], in order to understand some of the advantages and disadvantages of using them in financial applications depending on the scope.

This research can / should be extended as there are books and many tutorials and courses for an in-depth study of software instruments for modeling and analyzing finance phenomena.

2. LITERATURE REVIEW

Statistics is, in short, the science of learning from data. Modern statistics allow the analyst to fit and assess models (Benjamin S. Baumer 2017). For contemporary data science, integrative knowledge from statistics, computer science, mathematics, and a domain of application is required and also requiring highly skilled/ trained personnel (Benjamin S. Baumer 2017).

A research design indicates how to plan and conduct empirical research (both quantitatively and qualitatively), including: descriptive statistics, mathematics, and popular software tools for analysis (N.J.Salkind 2010).

Automation changes the activities of all sectors, including finance, as robots and computers can perform routine physical work activities better and more cheaply than humans, being increasingly capable of accomplishing activities that include cognitive capabilities (James Manyika 2017).

The analysis of the impact of automation, covering 78% of the global labor market, the high percentage of time spent on activities with the technical potential for automation by adapting currently demonstrated technology shows a very high potential for automation, in many countries. In figure 1, we can see the high potential for automation in the European countries[5], in the case of the Finance and insurance sector; when all the sectors all considered, this potential is still high, reaching 50% in several countries, such as Russia or Italy (McKinsey Global Institute 2017).

[3] http://www.r-project.ro/conference2017/

[4] With MATLAB®, financial risk models for banking, insurance, asset management, or supervisor institutions, stress testing, or decision-making optimization under uncertainty is possible, in the context of complex regulatory requirements (such as IFRS 9, Basel III, and Solvency II). In MATLAB, quantitative methods can be applyed across risks (credit, market, operational, etc.) (The MathWorks, Inc. 2018)

[5] For some reason, the situation of Romania has not been included in this study

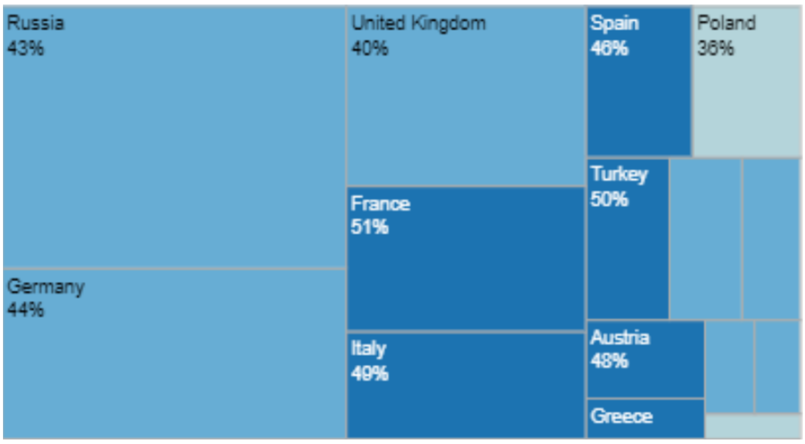


FIGURE 1. THE HIGH POTENTIAL FOR AUTOMATION IN THE FINANCE AND INSURANCE SECTOR
Source: (McKinsey Global Institute 2017)

Computational econometrics is applied econometric practice, as the development of computational techniques for econometrics (estimation or numerical methods); computer studies (Monte Carlo experiments, genetic algorithms, network studies, or estimation methods like simulated annealing); studies in which the computer does nonlinear estimation of large-scale systems, massive simulations; development of computational environments for econometric studies (GAMS[6] or Stata[7]) (Edited by David A. Belsley 2009). The General Algebraic Modeling System (GAMS) is used by multinational companies, universities, research institutions and governments in energy and chemical industries, for economic modeling, or agricultural planning, or manufacturing (GAMS n.d.).

In practice, most data science teams, economists and corporations use a mix of languages, often, at least R and Python[8] (R for Data Science n.d.). In the top 5 of the TIOBE index for February 2018, the Python Programming Language ranks 4 (climbing one step in the last year), while R is in ascension, ranking 13 in the popularity of programming languages; however, MATLAB programming language is falling, ranking the 19th, after reaching its highest position in 2001, when it was #10 (TIOBE 2018).

[6] Started as a project at the World Bank in the 1970s, GAMS was the first software system to combine the language of mathematical algebra with traditional concepts of computer programming, in order to describe and solve optimization problems (GAMS n.d.).

[7] Stata is an integrated statistical software package with tools for data analysis, data management, and graphics (STATA n.d.)

[8] Python is a powerful programming language, with high-level data structures and object-oriented programming, for scripting and rapid application development. The Python interpreter and the extensive standard library are freely available in source or binary form from the site https://www.python.org/ that also includes distributions of free third party Python modules, programs and tools, and documentation (Python Software Foundation 2018).

3. METHOD

This qualitative research is based on discussions and inputs from PhD students[9], with respect to the advantages of using R system - a statistics system or, in other opinions, an environment with both classical and modern statistical techniques (R 2009). For example, data mash-ups in R can analyze mortgage foreclosures by extracting addresses from a public report, placing them on a map and grouping them by various criteria, such as valuation and other socio-economic factors (Loukides 2011). Others are starting to use R for data analysis, data manipulation from Oracle databases, reports, data editing, survey estimation and found some problems working with big data sets, also survey sampling as there are surveys with complex sampling designs (Rudys 2016). R has the advantage that allows extensions of functions developed by the user for econometric modeling (Simionescu 2014). Also, if we consider the output from a regression or discriminant analysis, R gives minimal output and stores the results in a fit object for subsequent interrogation by further R functions (Venables, Smith and Team 2016).

On the other hand, RStudio is recommended for beginners as no prior knowledge/ experience with R/ RStudio are required.

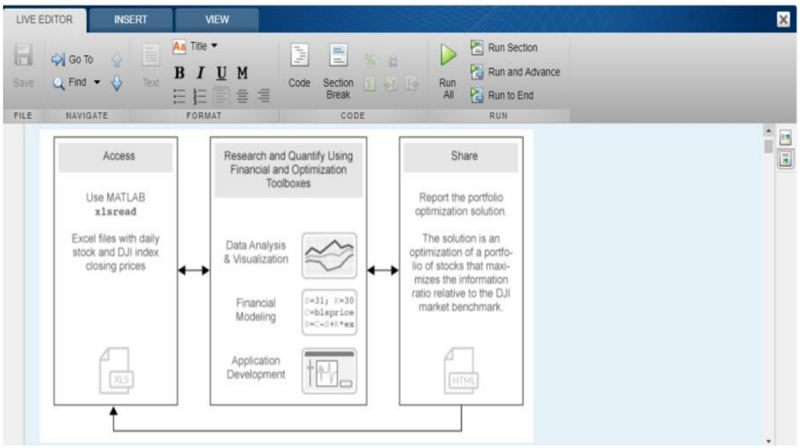
The features of RStudio include, besides the increased productivity of R users (Benjamin S. Baumer 2017):

- Integrates tightly reproducible analysis tools,
- Packages, such as the very useful and also very popular *ggplot2* for creating graphics (based on the data input, maps variables to aesthetics, etc.) (Hadley Wickham n.d.);
- Helps avoid error-prone "cut-and-paste" workflows;
- supports authoring HTML, PDF, Word Documents, and slide shows, and interactive graphics (with Shiny and ggvis);
- code-completion,
- The open Source Edition of RStudio Executes R code directly from the source editor, jumps to function definitions, has interactive debugger to diagnose and fix errors quickly (RStudio 2018).

However, for the students in Finance, using MATLAB' Financial Toolbox™ functions for mathematical modeling and statistical analysis of financial data, can help optimizing portfolios of financial instruments, optionally take into account turnover and transaction costs, estimate risk, analyze interest rate levels, price equity and interest rate derivatives, and measure investment performance. Time series analysis functions help the user perform transformations or regressions with missing data and convert between

[9] participating in EUNIT2017 conference, http://asecib.ase.ro/simpozion/2017/simpozion.htm

different trading calendars and day-count conventions (The MathWorks, Inc. 2018). In figure 2 below, we can see the steps of portfolio optimization using the Portfolio object in Financial Toolbox™, in an example of MATLAB Portfolio Optimization against a benchmark.



Use MATLAB **xlsread**

Excel files with daily stock and DJI index closing prices

Report the portfolio optimization solution.

The solution is an optimization of a portfolio of stocks that maximizes the information ratio relative to the DJI market benchmark.

FIGURE 2. PERFORMING PORTFOLIO OPTIMIZATION AGAINST A BENCHMARK IN MATLAB
Source: (The MathWorks, Inc. 2018)

In the annex, table 1 summarizes several traits of these systems.

4. RESULTS

There are many options for the finance students with respect to software instruments for modeling and analyzing finance phenomena. However, he/ she can choose the proper tools according to the tasks performed, accessibility of the soft, and knowledge of it, to name a few. One common feature of the programming languages studied is that the information input can be done from all kinds of files, including Excel.

5. CONCLUSIONS

Data analytics help organizations harness their data, identify opportunities or reduce costs through automation. In this era of “big data”, data from sales transactions, web entries, etc. are amassed by networks of instruments and computers, by skillful analysts, researches, or economists, etc. (Benjamin S. Baumer 2017).

In such context of machines getting smarter, there is a strong need for continuous, lifelong learning in general and even more, in finance. We should keep in mind that the ability to adapt skills to the changing needs of the workplace will be critical for the employees and lifelong learning, a must. However, the reality falls far short of the necessity (McKinsey & Company 2017).

REFERENCES

Benjamin, S. Baumer., Daniel, T. Kaplan., Nicholas, J. Horton. Modern Data Science with R. 2017. https://books.google.ro/books?id=NrddDgAAQBAJ&dq=data+science&hl=ro&source=gbs_navlinks_s, (accessed 2018).

Briand, Genevieve. & R., Carter, Hill. Using Excel for Principles of Econometrics. New York: John Wiley and Sons, 2011.

Brooks, Chris. Introductory Econometrics for Finance. Cambridge: Cambridge University Press, 2014. Edited by David, A. Belsley., E. J. Kontoghiorghes. "Preface." In Handbook of Computational Econometrics. Chippenham, Wiltshire: John Wiley & Sons, 2009.

Essnet Validat Foundation. "Methodology for data validation 1.0." 06 2016. https://ec.europa.eu/eurostat/cros/system/files/methodology_for_data_validation_v1.o_rev_2016_06_fial.pdf (accessed 01 2018).

Fernandez, Pablo. Company valuation methods. Madrid, 2017. GAMS. About the Company. GAMS. n.d. https://www.gams.com/about-the-company/ (accessed 12 2017).

Hadley, Wickham., Winston, Chang. Overview. n.d. http://ggplot2.tidyverse.org/ (accessed 01 2018).

Huber, Chuck. Data management made easy. STATA. 11 15, 2017. https://blog.stata.com/ (accessed 01 2018).

James, Manyika., Michael, Chui., Mehdi, Miremadi., Jacques, Bughin., Katy, George., Paul, Willmott. & Martin, Dewhurst. Harnessing automation for a future that works. 01 2017. https://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works (accessed 2018).

Loo, Mark van der. "Statistical Data Cleaning with R." Use of R. Bucharest, 2017.

Loukides, Mike. What Is Data Science? 04 2011.

Mathworks. n.d. https://de.mathworks.com/videos/matlab-student-overview-89557.html (accessed 01 2018).

McKinsey & Company. McKinsey Quarterly. New York, 2017.

McKinsey Global Institute. Where machines could replace humans — and where they can't (yet). 01 2017. https://public.tableau.com/profile/mckinsey.analytics#!/vizhome/InternationalAutomation/WhereMachin sCanReplaceHumans (accessed 2018).

N. J. Salkind, editor. Encyclopedia of Research Design. California, USA: SAGE Publications, 2010.

Python Software Foundation. 2018. https://docs.python.org/3/tutorial/index.html (accessed 01 2018).

Quantitative Micro Software, LLC. EViews 6 Student Version. 2007.

R for Data Science. n.d. http://r4ds.had.co.nz/introduction.html (accessed 01 2018).

R Project R-omania Team. n.d. http://www.r-project.ro/.

R. What is R? 2009. https://www.r-statistics.com/page/16/ (accessed 01 2018).

RStudio. n.d. https://www.rstudio.com/products/rpackages/ (accessed 01 2018).

RStudio Desktop. 2018. https://www.rstudio.com/products/RStudio/ (accessed 01 2018).

Rudys, Tomas. "Use Of R in Statistics Lithuania." (Romanian Statistical Review), no. 2 (2016).

SAS. About SAS. SAS. n.d. https://www.sas.com/en_us/company-information.html#history (accessed 01 2018).

Simionescu, Mihaela. Modelare economică în Matlab, R și Eviews. București: A.S.E., 2014.

STATA . Why Stata. n.d. https://www.stata.com/why-use-stata/ (accessed 01 2018).

The MathWorks, Inc. Financial Risk Management: Improving Model Governance with MATLAB. 2018. https://www.mathworks.com/campaigns/products/offer/financial-risk-improve-model-governance-white-paper.html (accessed 2018).

The MathWorks, Inc., Financial Toolbox. 2018. https://www.mathworks.com/help/finance/getting-started-with-financial-toolbox.html (accessed 2018).

The MathWorks, Inc., Portfolio Optimization Against a Benchmark. 2018. https://www.mathworks.com/help/finance/examples/portfolio-optimization-against-dow-benchmark.html (accessed 2018).

TIOBE. TIOBE Index for February 2018. 02 2018. https://www.tiobe.com/tiobe-index/ (accessed 02 2018).

Venables, W. N., D. M. Smith & R Core Team. "An Introduction to R." 2016. https://cran.rproject.org/doc/manuals/r-release/R-intro.pdf.

Webb, Allen. McKinsey Quarterly 2017 Number 4: Overview and full issue. 2017. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/mckinsey-quarterly-2017number-4-overview-and-full-issue (accessed 2018).