PRACTICAL ASPECTS OF R IN FINANCE, MANAGEMENT INFORMATION AND DECISION SCIENCES

David E Allen, University of Sydney

ABSTRACT

The environment and programming language R is reviewed and its capabilities, attractiveness and limitations are assessed, with reference to its continuing potential use and application finance, management information and decision sciences. A brief review of the top packages and programmers reveals further information about the relative strengths and weaknesses of R. Its adoption by major technology companies such as Microsoft and their augmentation of the core R library seems likely to further promote its future growth and popularity.

Keywords: R, Open-Source, S, Cran Views.

INTRODUCTION

R is a powerful environment and programming language for the manipulation, analysis and visualization of numerical data. New technology and ideas often appear first in one of the multitude of R package libraries. R is available under an open source license, which means that anyone can download and modify the code. Furthermore, as the source code is directly visible it can be modified and improved. R is easily extensible which means that developers can write their own software and distribute it in the form of new R packages. R compiles and runs on a wide variety of UNIX platforms, Windows and Mac OS.

R is similar to the S language and environment which was developed at Bell Laboratories in the 1970s by John Chambers. R was created by Ross Ihaka & Robert Gentleman at the University of Auckland, New Zealand and is currently developed by the R Development Core Team, of which Chambers is a member. The project was conceived in 1992, with an initial version released in 1995 and a stable beta version in 2000. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

For particularly demanding computer intensive tasks, C, C++ or FORTRAN code can be called up when required at run time. R is also a form of object programming. This means that the program is based on the concept of 'objects', which may contain data, in the form of yields, often known as attributes and code, in the form of procedures, often known as methods. This is strength in comparison to a number of other statistical computing methods. Another advantage is the ease with which publication quality plots can be produced, including mathematical symbols and formulae.

It suggests on the CRAN R website (https://www.r-project.org/about.html) that R is an integrated suite of software facilities for data manipulation, calculation and graphical display and that its attractive features include:

- An effective data handling and storage facility,
- A suite of operators for calculations on arrays, in particular matrices,
- A large, coherent, integrated collection of intermediate tools for data analysis,
- Graphical facilities for data analysis and display either on-screen or on hardcopy, and
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

It provides an 'environment' for statistical computing in that it comprises an integrated, coherent and fully-extendable system and set of capabilities. This contrasts with targeted software that has a very specific and inflexible set of capabilities, as is frequently the case with commercial software.

The closest commercial alternative would be Mat lab whose website suggests that it combines a desktop environment tuned for iterative analysis and design processes with a programming language that expresses matrix and array mathematics directly. However, it is particularly expensive and a single license and required 'tool boxes' can be as much as $3000 plus for the base single license with extra expense for required toolboxes. By contrast similar tasks can be done with R at a zero cost.

There are also some good integrated development environments (IDE) for R, which are also open source and therefore free. R studio is a good example and provides a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management (https://www.rstudio.com).

The R Consortium is collaboration between the R Foundation, R Studio, Microsoft, TIBCO, Google, Oracle, HP and others. It is chartered to fund and inspire ideas that will enable R to become an even better platform for science, research and industry (https://www.r-consortium.org). The central mission of the R Consortium is to work with and provide support to the R Foundation and to the key organizations developing, maintaining, distributing and using R software through the identification, development and implementation of infrastructure projects. The consortium website suggests that: “The R language has enjoyed significant growth and now supports over 2 million users. A broad range of industries have adopted the R language, including biotech, finance, research and high technology industries. The R language is often integrated into third party analysis, visualization and reporting applications”.

The consortium's mission is to seek to work with and provide support to the R Foundation and key organizations and groups developing, maintaining, distributing and using R software. The growth and development of R as a leading platform for data science and statistical computing. The R Foundation maintains a permanent seat on the board of the R Consortium, as an open communication channel for R Consortium members. It seeks to fund projects to enhance R and support its users. Projects are proposed by the R community at large and selected for funding by the Infrastructure Steering Committee. R Consortium members nominate the selection committee and provide funds for project grants with their membership dues. The R Consortium sponsors R related conferences (including use R!), meetings (including Sat RDays and RLadies) and local user groups worldwide. Enabling the use of R in commercial environments and fostering collaboration between companies investing in R. R Consortium committees are developing programs for R language certification and training, consulting and employment.

There are a number of top-tier companies using R, such as: Facebook where it is used for behavior analysis related to status updates and profile pictures. Google applies it to analyses advertising effectiveness and for economic forecasting. Twitter employs it for data visualization and semantic clustering. Microsoft recently acquired Revolution R Company, now known as Revolution Analytics and uses it for a variety of purposes. Microsoft in early 2017 also announced R Tools for Visual Studio (RTVS), an add-on R environment that integrates into Visual Studio. Uber employs R for statistical analysis whilst Airbnb use it in scaled data science. IBM has joined the R Consortium Group, whilst the Australian ANZ bank group employs it for credit risk modeling.

For large data sets a lot of companies still use the commercial software SAS. However, SAS is no match for R in the availability of cutting edge tools, as R is frequently used by academics to develop new statistical techniques which are provided for the general user in new packages. These often take years to become available in commercial programs. Revolution Analytics is promoting the use of R for commercial users and it has a business model based on its service packages, which give customers access to the libraries the company develops in-house. These commercial libraries are aimed at corporate customers who deal with large amounts of data-big data. Revolution Analytics, does not confine itself exclusively to R, but also creates user interfaces and algorithms, often using C++ to write its algorithms.

The company has also developed libraries that become open source, such as R Hadoop. This allows users of R Hadoop to leverage the data-computing environment provided by Hadoop to manage their data.

ScaleR, which is available on the Microsoft server (https://www.microsoft.com/enus/sql-server/machinelearningserver), permits businesses to interrogate all of their data by scaling it to work on parallel processors. Standard R packages, access internal machine memor and will run out of memory when dealing with large amounts of data. ScaleR manipulates the data to process chunks of it on different servers simultaneously.

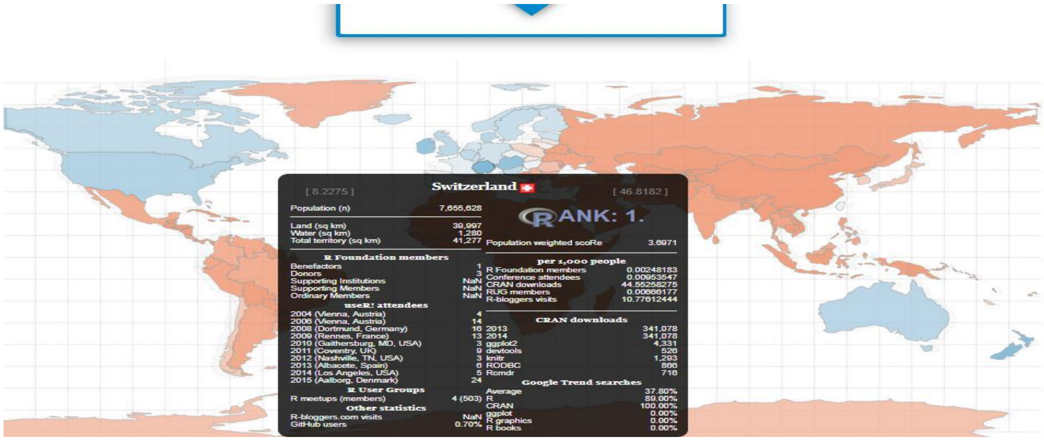
Facebook data scientists frequently use open-source R packages developed at R Studio by Hadley Wickham such as ggplot2, dplyr, plyr and reshape. Their use permits the team to explore new data through custom visualizations without having to develop new code, which is available in standard R packages.

Advantages and Dis-Advantages of R

As was indicated in the previous section, there are a number of advantages for using R in applications in Finance, Management Information and Decision Sciences, to summarize:

- It’s powerful and state of the art.
- It is used by statisticians, data-miners and programmers in some of the world's largest hi-tech companies.
- It is easy to improve, extend and modify its vast array of library packages. It has great data visualization and graphics capabilities.
- It runs on a variety of platforms including windows, mac and UNIX.
- It is programmable; if something is not available in R you can program it or add a new package with the required capabilities.
- There is a great deal of on-line support, in the way of dedicated websites, blogs, conferences, users groups and so forth.
- It is open source and free, which is a massive advantage for such a powerful programming environment.

- There are also disadvantages to R which include the following:
- There is a fairly steep learning curve initially and basic things, such as methods of importing data sets and data conventions and manipulation, take a little time to become familiar with.
- It is not interactive in the point and clicks sense, as some commercial programs are, though there is a point and click interface in the form of Rcmdr, which has an ever-expanding set of statistical capabilities.
- By and large it is command based and users have to become familiar with scripting code.



Population (n)	7,655,628
Land (sq km)	39,997
Water (sq km)	1,280
Total territory (sq km)	41,277
R Foundation members	
Benefactors	1
Donors	3
Supporting Institutions	NaN
Supporting Members	NaN
Ordinary Members	NaN
useR! attendees	
2004 (Vienna, Austria)	4
2006 (Vienna, Austria)	14
2008 (Dortmund, Germany)	16
2009 (Rennes, France)	13
2010 (Gaithersburg, MD, USA)	3
2011 (Coventry, UK)	9
2012 (Nashville, TN, USA)	3
2013 (Albacete, Spain)	6
2014 (Los Angeles, USA)	5
2015 (Aalborg, Denmark)	24
R User Groups	
R meetups (members)	4 (503)
Other statistics	
R-bloggers.com visits	NaN
GitHub users	0.70%

Population weighted scoRe	3.6971
per 1,000 people	
R Foundation members	0.00248183
Conference attendees	0.00953547
CRAN downloads	44.55258275
RUG members	0.00666177
R-bloggers visits	10.77612444
CRAN downloads	
2013	341,078
2014	341,078
ggplot2	4,331
devtools	526
knitr	1,293
RODBC	866
Rcmdr	716
Google Trend searches	
Average	37.80%
R	89.00%
CRAN	100.00%
ggplot	0.00%
R graphics	0.00%
R books	0.00%

Figure 1
GLOBAL USE OF R BY COUNTRY PER 1000 POPULATION

It is continually evolving and new versions of the base set of programs are released several times a year. Sometimes backward compatibility can be an issue.

However, the continuing expansion of the R user community shows that the advantages of R outweigh its disadvantages on a grand scale. The website http://rapporter.net/custom/R-activity/#score/3 provides an inter-active map of global R use on a scale related to the proportion of users per thousand populations. The map is shown in Figure 1. Switzerland is the number one user on this basis, followed by New Zealand, Austria is ranked three, Ireland four, US islands five, United States six, Australia seven, Singapore eight, Denmark nine, the UK ten and Canada eleven. So it truly is a global phenomenon.

Packages Available for Finance, Management Information and Decision Sciences

The Cran views website summarizes the various R packages available for use in various areas. For example https://cran.r-project.org/web/views/Finance.html the finance directory a list of packages useful for empirical work in Finance, grouped by topic. It notes that: “Besides these packages, a very wide variety of functions suitable for empirical work in Finance is provided by both the basic R system (and its set of recommended core packages) and a number of other packages on the Comprehensive R Archive Network (CRAN). Consequently, several of the other CRAN Task Views may contain suitable packages, in particular the Econometrics, Multivariate, Optimization, Robust, SocialSciences and TimeSeries Task Views”. Eleven packages are listed under the heading 'regression models' and nearly forty packages are listed under the 'time series' heading. There are roughly seventy packages under the heading 'finance', plus another twenty packages under the heading 'risk management'. Two books with dedicated packages are mentioned, plus another twenty-two packages under the heading 'data and date management'.

The CRAN task view web page on “Machine Learning & Statistical Learning', https://cran.r-project.org/web/views/MachineLearning.html lists packages to implement ideas and methods developed at the borderline between computer science and statistics-this field of research is usually referred to as machine learning. Under the heading 'Neural networks and deep learning', some sixty-one R packages are mentioned. The next heading 'boosting and gradient design', lists a further fourteen packages. 'Bayesian methods' comprises three packages, whilst a further three are included under the term 'optimization using genetic algorithms'. Two packages are listed under 'association rules' and a further two under 'fuzzy rule based systems'. Six packages are listed under 'model selection and validation', with a further three under 'other procedures'. 'Meta packages' contains three package references, whilst 'elements of statistical learning' contains a package relating to a book with the same title. The package ‘rattle’ provides a graphical user interface to data mining in R. A further eleven packages are listed under the heading 'visualization'.

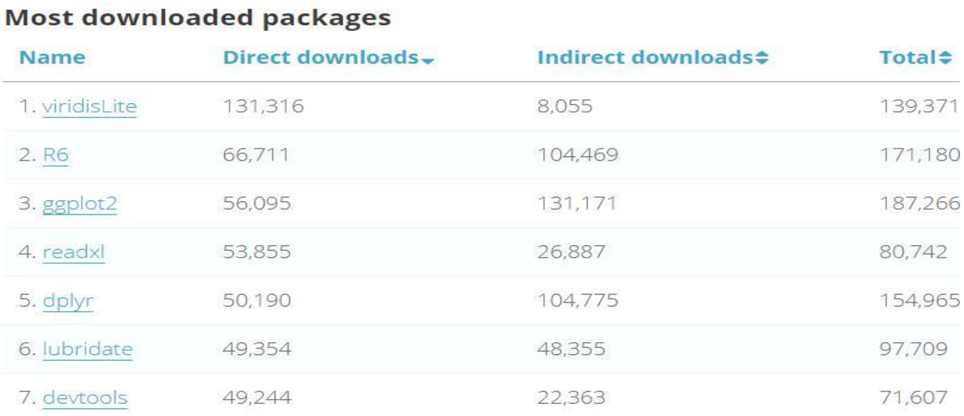
The above is only a partial list of R packages that may be of use in particular subject areas. It is far from the whole story, as its notable that in early 2017 CRAN, the R repository, reached the milestone of 10,000 packages. Even this does number not constitute the entire total. There are another 1294 packages for genomic analysis in the Bio Conductor repository, plus there are also hundreds of R packages available only on GitHub, commercial R packages from vendors such as Microsoft and Oracle and an unknown number of private, unpublished packages. The R leader board https://www.rdocumentation.org/trends currently mentions 14,805 total packages. This might be an embarrassment of riches, but there are a number of ways of sifting through these packages to find what’s most suitable for your purposes. We have already considered CRAN task views. MRAN (the Microsoft R Application Network) provides a search facility for R packages.

What are the Most Popular R Packages?

Figure 2, taken from https://www.rdocumentation.org/trends shows the current most-downloaded R packages.

The first package on the list, viridisLite, is part of the new 'matplotlib' color maps ('viridis'-the default-'magma', 'plasma' and 'inferno') to 'R'. 'mat-plotlib' <http://matplotlib.org/> is a popular plotting library for 'python'. These color maps are designed in such a way that they will analytically be perfectly perceptually-uniform, both in regular form and also when converted to black-and-white.

The second on the list, R6 package is also a technical package for the creation of classes with reference semantics, similar to R's built-in reference classes. Compared to reference classes, R6 classes are simpler and lighter-weight and they are not built on S4 classes so they do not require the methods package.



Most downloaded packages

Name	Direct downloads	Indirect downloads	Total
1. viridisLite	131,316	8,055	139,371
2. R6	66,711	104,469	171,180
3. ggplot2	56,095	131,171	187,266
4. readxl	53,855	26,887	80,742
5. dplyr	50,190	104,775	154,965
6. lubridate	49,354	48,355	97,709
7. devtools	49,244	22,363	71,607

Source: https://www.rdocumentation.org/trends

Figure 2
R MOST DOWN-LOADED PACKAGES

The third, ggplot2, is a comprehensive graphics package, which permits 'declaratively' creating graphics, based on "The Grammar of Graphics". The user provides the data, tells 'ggplot2' how to map variables to user requirements, what graphical parameters to use and the package takes care of the details.

The fourth, readxl, imports excel files into R and supports '.xls' via the embedded 'libxls' C library, whilst the fifth, dplyr, provides a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges, including:

- Mutate-adds new variables that are functions of existing variables
- Select-picks variables based on their names.
- Filter-picks cases based on their values.
- Summarize-reduces multiple values down to a single summary.
- Arrange-changes the ordering of the rows.

The sixth on the list, lubridate, provides convenient functions to work with date-times and time-spans: fast and user friendly parsing of date-time data. The seventh, devtools seeks to make R package development easier by providing R functions that simplify common tasks. The eighth, 'readr' provides a fast and friendly way to read rectangular data (like 'csv', 'tsv' and 'fwf').

The fact that these packages are the most downloaded provides information about issues commonly encountered by R users. The use of graphics is important and viridisLite and ggplot2 provide graphical enhancements.

Top Programmers-1

Development of R and R packages is important and R6 and devtools provide convenient assistance in this area. Imputing data in the correct format into R is a commonly encountered hurdle faced by users and assistance is provided by readxl and dplyr. Another common challenge is the formatting of dates and times correctly for time-series analysis and this is assisted by lubricate.

Top Programmers-2

The website https://www.rdocumentation.org/trends also indicates who the authors of the most downloaded packages are. Top of the list is Hadley Wickham who currently has 105 packages on CRAN plus 16 packages on GitHub. He is Chief Scientist at R Studio and an Adjunct Professor of Statistics at the University of Auckland, Stanford University and Rice University. His most downloaded package is ggplot2, which is a system for 'declaratively' creating graphics, based on "The Grammar of Graphics".

The second on the list is Kirill Müller who currently has 27 packages on CRAN and 3 packages on GitHub. His most downloaded package is 'tibble' which provides a 'tbl_df' class (the 'tibble') that provides stricter checking and better formatting than the traditional data frame.

Third on the list Yihui Xie who has 37 packages on CRAN and 8 packages on GitHub. Yihui Xie is a software engineer at R Studio and he earned his PhD from the Department of Statistics, Iowa State University. His interests include interactive statistical graphics and statistical computing. The most downloaded package he has written is MIME, which is a specialist package for interpreting "Multipurpose Internet Mail Extensions”. The second on the list, Knitr, provides a general-purpose tool for dynamic report generation in R using Literate Programming techniques.

The fourth most downloaded programmer on the list is Dirk Eddelbuettel, who is a Canadian statistician, data scientist and researcher. He is the author of the open-source software package Rcpp and has also written the textbook Seamless R and C++ Integration with Rcpp. This is his most downloads package which provides R functions as well as C++ classes offering a seamless integration of R and C++. Many R data types and objects can be mapped back and forth to C++ equivalents which facilitates both writing of new code as well as easier integration of third-party libraries.

The fifth is Jeroen Ooms who is a postdoctoral researcher at UC Berkeley. He has 54 packages on CRAN and 4 packages on GitHub. His most downloaded package JsonLite is a fast JSON parser and generator optimized for statistical data and the web. The package offers exible, robust, high performance tools for working with JSON in R and is particularly powerful for building pipelines and interacting with a web API. JSON or JavaScript Object Notation is a minimal, readable format for structuring data. It is used primarily to transmit data between a server and a web application.

The fact that these are the most downloaded packages suggests that enhanced graphics, coding of data and securing of web data, internet facilitation and interpretation of mail and java data sources as well as convenience of porting and interfacing R with calculations in C++ are major themes adopted by R.

Top Programmers-3

Users: It is also interesting that a number of these programmers are associated with R Studio which is the producer of a popular R IDE but which also has commercial interests in R development.

This focus of these developers can be contrasted with comments about the future of R by Ihaca (2010), who together with Robert Gentleman was one of the original developers of R. He has expressed a number of reservations about the future of R in various presentations he has given in recent years. He notes that the S-like appearance of R was not an original feature but was added incrementally. This similarity to S was driven by the desire to access already existing programming expertise and code. Once R became more S-like the move towards making it S compatible became irresistible. This ultimately produced the current mature and widely-used system. Now that R now has a large number of users who require a stable platform for getting work done, it is no longer particularly suitable as a base for experimentation and development.

He suggests that the strengths of R include the fact that it is an interactive, extensible, vectorized language. It possesses a large run-time environment which provides a good deal of statistical functionality. It has good (static) graphics capabilities which are enhanced by the creation of new packages such as ggplot2. There are a wide variety of community support mechanisms such as websites, blogs, dedicated conferences and so forth. An additional strength is the fact that users have the freedom to inspect, modify and redistribute the source code.

He points out that R is not very good at handling large-scale problems and lists some issues causing difficulties. He suggests that the execution of large amounts of R code causes problems and that scalar (element-by-element) computation are slow. R encounters difficulties with computations on large volumes of data and he notes that some computational problems involve a mix of all of these. He suggests that statistical techniques are increasingly computationally intensive and that large datasets in the order of magnitude of petabytes are increasingly becoming available.

However, in section one of this paper I mentioned some of the major technology companies involved in the commercial development of R plus the fact that they are already using it for analytical and statistical purposes. A number of the packages developed facilitate parallel computing plus the interfacing of R with C++.

It is significant that in 2017 Microsoft released Microsoft R Open (https://mran.microsoft.com). Microsoft R Open is the enhanced distribution of R from Microsoft Corporation. The current release, Microsoft R Open 3.4.2, is based the statistical language R-3.4.2 and includes additional capabilities for improved performance, reproducibility and platform support. At the moment this is open-source and therefore free. The benefit to users is that a set of specialized packages released by Microsoft Corporation are added to further enhance the performance of base R.

- Multi-threaded math libraries that brings multi-threaded computations to R.
- A high-performance default CRAN repository that provide a consistent and static set of packages to all Microsoft R Open users.
- The checkpoint package that make it easy to share R code and replicate results using specific R package versions.

The multi-threaded math libraries improve the performance of the base R installation. R was originally designed to use only a single thread (processor) at a time. R continues to work that way unless linked with multi-threaded BLAS/LAPACK libraries. Today's multicore machines offer parallel processing power. Microsoft R Open provides access to these by including multi-threaded math libraries. These libraries make it possible for many common R operations, such as matrix multiply/inverse, matrix decomposition and some higher-level matrix operations, to compute in parallel and use all of the processing power available to reduce computation times.

It seems to the writer that R has become something of a juggernaut and that there are now so many users and developers that new capabilities and means for speeding up its scope and computational efficiency seem to follow automatically and novel fixes are being developed for avoiding inherent limitations.

Practical Introductions to the Use of R in Finance, Management Information and Decision Sciences

There are a number of useful introductory primers on the use of R on the Cran website (https://cran.r-project.org, 'The R Manuals'). The manuals include 'An Introduction to R', which is based on the former 'Notes on R' and provides an introduction to the language and how to use R for doing statistical analysis and graphics. 'R Data Import/Export', describes the import and export facilities available either in R itself or via packages which are available from CRAN. 'R Installation and Administration', describes how to install R on the various operating systems, UNIX, Windows and mac OS, plus descriptions of how to run it and various internal features. “Writing R Extensions”. Covers how to create your own packages, write R help files and the foreign language (C, C++, FORTRAN...) interfaces. 'The R language definition' provides documentation of the R language per se. 'R Internals' provides a guide to the internal structures of R and coding standards for and was conceived for the benefit of the core team working on R itself. Finally, 'The R Reference Index' provides all the help files of the R standard and recommended packages in a printable format.

On the Cran website under the 'Documentation heading there is a section featuring “FAQ”, frequently asked questions, plus several dozen contributed documents, both about the general use of R, are primers on more specific statistical and graphical uses of R. There are also numerous published primers on R.

Palgrave Macmillan in association with Springer has a 'Use R' series edited by Gentleman, Hornik & Parmigiani, which at the time of writing featured some 65 titles. (https://www.palgrave.com/la/series/6991). These feature specialized topics in various subject area applications, such as a guide to network analysis with R, quality control with R, genomic data analysis with R, modern optimization in R, Bayesian networks in R, to mention but a few.

In relation to Finance, Management Information and Decision Sciences, Singh and Allen (2017) have published R in Finance and Economics: A Beginner's Guide, which provides an introduction to the use of R in finance and economics research. Pfaff (2016) has the second edition of an excellent guide to financial risk modeling and portfolio optimization with R and has written an R package 'FRAPO' to accompany the book. Pfaff (2008) also has an excellent guide to integrated and co-integrated time series using R, which is part of the 'Use R' series. Rupert and Matteson (2015) have a book titled Statistics and Data Analysis for Financial Engineering which features R examples. Gilli et al. (2011) consider numerical methods and optimization in finance and have written an R package 'NMOF', which provides functions, examples and data from their text.

The late Diethelm Wurz founded Rmetrics (https://www.rmetrics.org), which has an excellent series of eBooks, some of which are free, but segments of code are now likely to be outdated, given that R has moved on from where it stood in 2016.

The Cran task view https://cran.r-project.org/web/views/MachineLearning.html provides an overview of R packages that implement ideas and methods developed. At the borderline between computer science and statistics, a field of research is usually referred to as machine learning. Hastie et al.'s (2009) primer on elements of statistical learning has an accompanying R package 'ElemStatLearn'.

Rforge also has some excellent R packages (https://r-forge.r-project.org). These include the previously mentioned 'NMOF', plus 'Opefimor', which is a companion package to the Iacus (2011) text on option pricing using R. Other packages in the machine learning area include 'Applied Predictive Modeling', a companion package to the text by Kuhn and Johnson (2013).

CONCLUSION

In this short review I have provided an introduction and an overview of the programming language R, which is beginning to dominate numerous areas of data analytics, graphical analysis and statistical computing, across multiple discipline areas. The fact that is free and open source, greatly adds to its attractiveness and impetus.

I have also mentioned the critique of R's capabilities by one of its inventors, Ihaca. These weaknesses appear likely to be partially circumvented by the adoption of R by major technology companies, such as Microsoft, who have recently released Microsoft R Open, to bring multi-threaded computations to R. Given the various attractive attributes of R that I have briefly touched upon it seems likely that its further rapid expansion and global adoption is likely to continue apace.

REFERENCES

Singh, A.K. & Allen, D.E. (2017). *R in finance and economics: A beginner's guide*. World Scientific: Singapore.

Gilli, M., Maringer, D. & Schumann, E. (2011). *Numerical methods and optimization in finance*. Elsevier.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction (2nd Edn).* Springer.

Iacus, S.M. (2011). Option pricing and estimation of financial models with R. Wiley.

Ihaca, R. (2010). *R: Lessons Learned, Directions for the Future: In JSM Proceedings, Statistical Computing Section*. Alexandria, VA: American Statistical Association.

Kuhn, M. & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Pfaff, B. (2008). *Analysis of integrated and co-integrated time series with R*. Springer.

Pfaff, B. (2016). *Financial risk modeling and portfolio optimization with R (2nd Edn)*. Wiley.

Ruppeert, D. & Matteson, D.S. (2015). *Statistics and data analysis for financial engineering*. Springer.