Institution: University of Oxford

Unit of Assessment: 10 – Mathematical Sciences

Title of case study: R, a free software environment for statistical computing and graphics

1. Summary of the impact

R is a free and open-source software programming language and software environment for expressing and implementing statistical algorithms and graphics. It has become the *lingua franca* for developing and implementing new statistical methodologies - not just in statistics, but in applications across the whole spectrum of industry, from marketing and pharmaceuticals to finance. It is used by companies for research, analysis and production. Its power in analysing and visualising data helps organisations from charities to government. About one half of the core statistical modelling and graphics engine included in R builds on research carried out in Oxford.

2. Underpinning research

R is a collaborative research project, run by the R Development Core Team (20 people in 10 countries), which has been running in its current form since 1997. The output is a freely available software programming language and environment for statistical computing and graphics, which runs on all major computer platforms, including UNIX, Windows and Mac OS. It has a core set of libraries to which users can add packages. The central purpose of R is to build on mathematical and statistical research and to provide a toolbox of software that others can use.

In 1994, Brian Ripley, Professor of Applied Statistics at the University of Oxford (1990-present), published a seminal paper, [1], which sets up a general statistical framework for classification. Within this framework, data analysis methods from the Statistics, Pattern recognition and Machine Learning literatures can be compared. The paper showed how developments in Pattern Recognition could be identified with statistical ideas, to the benefit of both fields. For example, it introduced neural networks to the field of statistics as a classification tool, provided a means of fitting neural networks to statistical problems, and identified neural networks as a form of robust regression analysis. Ripley joined the R project in January 1999, and he implemented these newly introduced classification techniques, and others developed in his book [2], in R as a package called MASS [3]. A major research output in its own right, MASS was first released as part of R in 1999, and now forms an important part of the core statistical modelling and graphics engine, delivering many of the analysis and modelling tools used on a day-to-day basis by R-users.

Other substantial elements of the R-core package also result from Ripley's research. For example, the new methods in computational statistics which underlie the core time series package, in particular the development of methods for handling missing data and the identification of appropriate optimization methods, were first described in [4] and implemented by Ripley in a 2002 R-release.

At its core, R is itself a very substantial research output, [5], underpinned by fundamental statistical research such as [1], [2].

3. References to the research

* [1] Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion). *J. Roy. Statist*. Soc. B 56, 409-456

* [2] Ripley, Brian D. (1996) Pattern Recognition and Neural Networks, Cambridge University Press

[3] MASS http://cran.r-project.org/web/packages/MASS/MASS.pdf

[4] Brian D. Ripley. Time series in R 1.5.0. *R News* 2(2):2-7 June 2002.

* [5] The R Project for Statistical Computing http://www.r-project.org/

The three asterisked outputs best indicate the quality of the underpinning research. [1] is in a high quality internationally refereed journal, [2] is a key book and [5] is R itself.

4. Details of the impact

R is ubiquitous. It is found in any arena in which people analyse, visualise or manipulate data. It has huge economic impact, benefitting millions of users from every sector of industry, from pharmaceutical to finance. It has resulted in improvements in performance in existing companies, by providing a platform which combines state of the art data analysis tools with high quality data visualisation, and it has spawned new companies, most notably Revolution Analytics, whose purpose since 2009 has been to make impact through their REvolution software, designed specifically for commercial users. R is used by Government agencies, not-for-profit organisations, search giants like Google and even dating agencies [A].

How research underpins impact

With MASS included, the first non-beta version of R - 1.0.0 - was released on 29th February 2000. When Ripley joined the R project in January 1999 the code repository was 1.5Mb in size. By January 2008, it had grown to 15Mb and in May 2013 it stood at 25Mb. Analysis of SubVersion commits - the measurement used to assess which users are updating collaboratively developed software - (http://www.ohloh.net/p/rproject/commits/summary) shows that, from 1999 to 2013, just over 50% per cent were by Ripley and that this rate of contribution has been sustained over the entire period. This phenomenal output constitutes literally millions of lines of code.

R is a research output which is placed directly in the hands of the user. Some of the R core team, including Brian Ripley, run a repository (cran.r-project.org) of publicly available extension packages which currently contains over 4700 packages. The active community of users and developers enables R to constantly adapt to incorporate the very latest research in data analysis and meet the rapidly shifting needs of R-users.

The software, along with many original insights on its use in applied statistics, are laid out in *Modern applied statistics with S-plus* (R is an implementation of the S programming language combined with lexical scoping semantics). This classic book, written by Ripley and co-author Venables, was published by Springer in 1994 and now in its 4th edition. It has sold over 90,000 copies. It is complemented by more recent specialised texts, often aimed at practitioners, such as those in the Springer series `UseR!'.

Nature and extent of impact

Since it is a free, open-source project, the full impact of R is impossible to quantify. Nonetheless, there is compelling evidence of extremely widespread use. Perhaps most obviously, in 2009 it spawned its own, refereed, online journal, called *R Journal* (http://journal.r-project.org/), which contains both articles about the development of R itself and about its applications. Consulting company Revolution Analytics (formerly known Revolution Computing until 2009 when a $9m injection of capital by Intel and North Bridge Venture Partners was accompanied by a new strategy aimed at R end users) undertook a survey in August 2011 which reported over 2 million active users of R, of whom 1.2 million were based in industry [B]. Their website provides links to 55 Local R User Groups, whose members meet up to brainstorm, network and listen to guest speakers. Groups span Australia, Asia, Europe, Israel and North and South America [C]. Some, such as the Chinese Financial R users group, focus on specific areas of application.

The Revolution Analytics website also provides a glimpse of the vast range of organisations using

R and how the continually evolving R library enables them to stay at the cutting edge of data analysis [A]. It also links to powerful testimonies, such as that of Antonio Possolo of the US National Institute of Science and Technology, who was charged with making sense of the conflicting estimates on the rate of oil flowing from the BP Oil spill at Deepwater Horizon; *"The quality that you have built into R, through public open examination, is the greatest strength and source of confidence that I could have asked for"* [C]. Possolo's use of R to run uncertainty analysis and harmonize estimates was crucial to decision makers in coordinating the scale and scope of the response to the emergency. Revolution Analytics' more than 60 non-academic customers include Merck, who use their software to collect and analyse massive data sets in clinical drug trials and Pfizer, who use it to analyze genetic data, perform predictive modelling, and carry out exploratory data analysis [D].

R is used in every sector of industry. As examples, the Bank of America uses R for capital adequacy modelling, decision systems design and predictive analytics [E]; and Google use R to help analyse its marketing data [E]. Use of R is so widespread at Google that it has its own R style guide, while there is an annual meeting R/Finance, devoted entirely to applied finance using R [F], and the first `R in insurance' event took place in 2013 [G].

Personal testaments received by the University of Oxford in 2013 also confirm the ubiquity of R (we include the maximum of 5 allowed):

- The Head of Discovery Informatics at e-Therapeutics Plc [H] states "*The R environment is now the de-facto standard for bioinformatics data analysis and forms a critical part of our computational toolbox at e-Therapeutics. The R environment allows us to 'stand on the shoulders of giants', utilise the vast statistical and bioinformatics knowledge embedded within the software and concentrate on solving our specific problems rather than reinventing the wheel by rewriting standard data analysis algorithms. I'd estimate that this alone saves us months of work on any project utilising such data.* […] *R is a critical part of the computational resources used by e-Therapeutics and personally I find it hard to believe that the impact of R needs to be justified.*"

- The Head of Data Science R&D at dunnhumby says [I] "*One tool we are finding increasingly useful for exploring advanced techniques, investigating new algorithms and rapidly prototyping new concepts, is the open source statistical language R. We have a core research and development team who are regular users of R*"

- The HR Manager from Tessella, a consulting firm, states [J] "*In life sciences, R is a popular tool / environment with many of our clients who value and use R over and above the traditional use of the SAS product that has previously been the mainstay of data analysis in clinical development*. We commonly use R on projects to support analysis/consultancy done for pharmaceutical clients. *Examples include analysing simulated data, modelling of clinical trials data and the development of causal reasoning algorithms. In these projects, R is typically used in the investigative development stages and sometimes pre development of faster codes for use in a more commercial environment*. *Another example would be work we have undertaken with a large agro chemical business on a project relating to understanding the lifecycle of pesticides within the ground. Tessella have developed a simple tool which implements the FOCUS Kinetics guidance to generate degradation kinetics. This was developed in C# with an R back-end. Further work in this sector involved developing a method to predict and understand fish survival - the development tool was R*."

- The Global Head of Risk and Performance at Investec Asset Management writes [K] "*A number of us at Investec Asset Management have been regular users of R for a number of years. R was perhaps first used seriously late 2006 when I joined the firm. Since that time, R has been integral to trading-model development in an initiative that I launched. More recently, my colleagues have used R as a building block within an equity investment process.* […] *Over the years, R has become our standard environment for statistical/mathematical problem solving.* […] *for rolling out pre-built solutions for non-mathematicians the ability to prototype solutions in*

R prior to programming in a compiled language has been invaluable. […] *representatives from Investec Asset Management have attended R/Rmetrics, R/Finance conferences and regular local meetings such as LondonR."*

- The Chief Analyst at the Implementation Unit at the Cabinet Office says [L] "*the following Government Departments and Agencies have told me that they use R: Department of Business, Innovation and Skills; Department for Communities and Local Government; Department for Environment, Food and Rural Affairs; Animal Health and Veterinary Laboratories Agency; Environment Agency; Food and Environment Research Agency; Food Standards Agency; Forestry Commission; Ministry of Justice; Marine Management Organisation; Natural England; Office for National Statistics. I can confirm, as TfL's former Director of Policy Analysis, that Transport for London have also used R."*

5. Sources to corroborate the impact

[A] Revolution Analytics' list of Companies Using R, demonstrating reach of the impact http://www.revolutionanalytics.com/what-is-open-source-r/companies-using-r.php

[B] Revolution Analytics user statistics, demonstrating reach of the impact http://prezi.com/s1qrgfm9ko4i/the-r-ecosystem

[C] Revolution Analytics Blog, demonstrating the reach of the impact:
user groups: http://blog.revolutionanalytics.com/local-r-groups.html
Deepwater: http://blog.revolutionanalytics.com/2010/08/rs-role-in-the-national-response-to-the-bp-oil-spill.html

[D] Use of R at Merck and Pfizer: http://www.revolutionanalytics.com/aboutus/our-customers.php

[E] R is Hot: Part 1 http://www.r-bloggers.com/r-is-hot-part-1/

[F] R/Finance http://www.rinfinance.com/

[G] "There is an R in Lloyds", Head of Exposure Management and Reinsurance, Lloyd's, R in Insurance Conference, London, 15 July 2013 http://lamages.blogspot.co.uk/p/the-first-conference-on-r-in-insurance.html?goback=.gde_4180165_member_197992140#!

[H] Letter from Head of Discovery Informatics at e-Therapeutics Plc. Copy held by University of Oxford

[I] Letter from Head of Data Science R&D at dunnhumby. Copy held by University of Oxford

[J] Letter from the HR Manager from Tessella. Copy held by University of Oxford

[K] Letter from the Global Head of Risk and Performance at Investec Asset Management. Copy held by University of Oxford

[L] Letter from the Chief Analyst at the Implementation Unit at the Cabinet Office. Copy held by University of Oxford

[D]- [L] exemplify the reach and demonstrate the significance and ubiquity of R.