RESEARCH ARTICLE

Approaches to R education in Canadian universities [version 1; referees: 1 approved, 2 approved with reservations]

Michael A. Carson , Nathan Basiliko

Department of Biology and the Vale Living with Lakes Centre, Laurentian University, Sudbury, Canada

v1 **First published:** 30 Nov 2016, **5**:2802 (doi: 10.12688/f1000research.10232.1)
Latest published: 30 Nov 2016, **5**:2802 (doi: 10.12688/f1000research.10232.1)

Abstract

Introduction: R language is a powerful tool used in a wide array of research disciplines and owes a large amount of its success to its open source and adaptable nature. The popularity of R has grown rapidly over the past two decades and the number of users and packages is increasing at a near exponential rate. This rapid growth has prompted a number of formal and informal online and text resources, the volume of which is beginning to present challenges to novices learning R. Students are often first exposed to R in upper division undergraduate classes or during their graduate studies. The way R is presented likely has consequences for the fundamental understanding of the program and language itself; user comprehension of R may be better if learning the language itself followed by conducting analyses, compared to someone who is learning another subject (e.g. statistics) using R for the first time. Consequently, an understanding of the approaches to R education is critical. *Methods:* To establish how students are exposed to R, we used a survey to evaluate the current use in Canadian university courses, including the context in which R is presented and the types of uses of R in the classroom. Additionally, we looked at the reasons professors either do or don't use/teach R. *Results:* We found that R is used in a broad range of course disciplines beyond statistics (e.g. ecology) and just over one half of Canadian universities have at least one course that uses R. *Discussion and Conclusions:* Developing programming-literate students is of utmost importance and our hope is that this benchmark study will influence how post-secondary educators, as well as other programmers, approach R, specifically when developing educational and supplemental content in online, text, and package-specific formats aiding in student's comprehension of the R language.

This article is included in the RPackage channel.

Open Peer Review

Referee Status: ✓ ? ?

	Invited Referees		
	1	**2**	**3**
version 1 published 30 Nov 2016	✓ report	? report	? report

1 **Colin W. Rundel**, Duke University USA

2 **Eliezer Gurarie** , University of Washington USA

3 **Luc F. Bussiere**, University of Stirling UK

Discuss this article

Comments (0)

Corresponding author: Michael A. Carson (mcarson@laurentian.ca)

How to cite this article: Carson MA and Basiliko N. **Approaches to R education in Canadian universities [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2016, **5**:2802 (doi: 10.12688/f1000research.10232.1)

Grant information: MC is supported by the Ontario Trillium Foundation through an Ontario Trillium Scholarship for International PhD Students, by fellowships from Laurentian University, and by NSERC Discovery (#341381) and a Canada Research Chair Grant (#228457) held by NB.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: NB is an active professor at a Canadian university and had the opportunity to participate in the survey. The authors do not feel his participation, or lack of participation, compete with the interest or conclusions of the study.

First published: 30 Nov 2016, **5**:2802 (doi: 10.12688/f1000research.10232.1)

Introduction

The R language was developed in the early 1990s by Ross Ihaka and Robert Gentleman in an attempt to write a statistical computing language that combined desirable aspects of two other languages, Scheme[1] and S[2]. For all non-developer user purposes, R is an interpreted object-oriented language that relies heavily on packages, which contain functions that users apply to their data (see Ihaka and Gentleman, 1996[3] for a more through explanation of the details and thought process behind the development of R). It could be argued that the success of R was by luck or maybe design, but the choice to target usage at statisticians meant that it had a reasonably large and dedicated user base from its inception, and subsequently, it has gained attention across academic and professional disciplines[4]. In a general sense, the concept of user-developed packages is the reason R has gained a lot of ground over other statistical software, as the broader community is given the tools and freedom to write specific code for their disciplines and research questions, which is formatted into functions and grouped into a package. These packages are then vetted by the R Core Team and made available through the CRAN repository[5]. This flexibility and R's social organization has led to a rapid growth of R use and the R community, which is reflected in a number of areas, including the expansion of the Core Team, an exponential increase in the number of packages in CRAN (ca. 100 in 2001 vs. ca. 7,000 in 2016), the rise of email list traffic[6], the number of downloads per year, and general R activity[7]. Additionally, based on download history from CRAN, there are millions of current R users[8], R has had a consistent rise in Google scholar hits (SAS and SPSS are declining)[9], and there have been more packages added in 2015 than have existed in all of the SAS institute's history[9]. Taken together, these metrics indicate the rise in popularity of R, and highlight the importance of teaching the next generation of students and researchers the most applicable skills.

We are living in a time of rapid technological advancement and an age where the free sharing of ideas is becoming a standard practice[10,11]. Evidence of this is seen in the proven effectiveness of the open source framework, within which R is developed[12]. For R users, open source means not reinventing the wheel every time a new problem arises. Instead they can search for packages to address specific analyses that others have written and made publically available on CRAN or through sources like GitHub[13] and Omegahat[14]. The open source nature not only means that primary R resources are freely available, but that the R community at large is also willing to provide troubleshooting support, as evidenced by the multiple independent support websites (e.g. Quick-R and Cookbook for R) and community forums that address user questions and problems (e.g. Stackoverflow and R-bloggers). Thus, this means that an average user has a diverse toolset to pull from, and an even larger support community to help them accomplish their task at hand. The open source nature of R and its sharing community are two important reasons that R is gaining popularity so rapidly in many business, research, and educational sectors.

While the R language is not specifically limited to data analysis, in science, technology, engineering, and math (STEM) disciplines it is commonly used for this purpose. For example, there were approximately 35,000 scholarly articles published across all disciplines (STEM and others) in 2015 with R as the primary analysis tool, second only to SPSS, which had decreased by 25% from the previous year[9]. This is most likely because unlike other analysis tools, R is adaptable to specific problems, while remaining versatile enough to address more common data management, analysis, and graphing needs as well; users can easily write new code or adapt other users' code to address their specific needs. In this way R promotes an active learning process, which is proven to increase students' performance in STEM education[15]. Additionally R is an "all in one" environment that streamlines data analysis workflow from data management and analysis to graphical data presentation and text processing. The concept of packages is also in line with many STEM disciplines and the nature of the scientific process and dissemination, where a reader can find the exact package used by others and do a similar analysis for their study. Finally, R gives STEM users multiple options with many packages that do nearly the same thing in slightly different ways. For example if a user wants to create a general plot, that capability is in the *base*[16] package, but there are also options to use an array of other packages that generate plots in slightly different ways (e.g. *lattice*[17] and *ggplot2*[18]). In short, R gives users options and is easily adaptable to exact tasks at hand, greatly benefiting STEM users as well as the R community at large.

The importance of programming education is becoming evident and universities have a significant role to play[19]. R is a prime language to use in undergraduate classrooms because it is extremely versatile, free, has a large user community, is relatively easy to learn in terms of programming (see Fox 2009[6]), and is supported across multiple computing platforms. This means that a student could encounter R in a wide array of classes ranging from traditional statistics to, for example, an ecological modeling or bioinformatics course. The programming skills learned in one course would easily transfer to other courses, and departments could benefit by coordinating course content to better capitalize on this continuum. Along this line, R allows students to preform practical applications rapidly upon learning the language, whereas languages geared more towards software development require more base knowledge before writing more meaningful code. This means that R is a compelling language to learn for novice programmers. Furthermore a solid foundation in R better prepares undergraduate students for postgraduate education or for seeking employment in a broad range of sectors. While there are other programing languages, the overall versatility and open source nature of R means that many research institutions and cooperate entities are using R at an increasing rate. Even if R were not the primary coding language used later in a career, learning any programming language often means that a student is better equipped to enter the job market[20]; however, most other data management and statistical programs an undergraduate is likely to encounter are a point-and-click format (e.g. Excel and SPSS), so they gain little practical coding experience.

The goal of this survey-informed study was to highlight R usage at Canadian universities, shedding light on which types of courses use R, as well as overall R training offerings at the institutional level. Additionally, we look at some of the benefits and challenges professors encounter teaching R to their students, and motivations for using R in their research programs and teaching R in the classroom. To our knowledge this is the first study to look specifically at R usage in an educational context, and thus may also help serve as a benchmark for future characterization of R usage in universities in general and Canadian universities through time.

Methods

Survey methods

A survey of 70 Canadian universities was conducted using Google Forms (https://www.google.com/forms/about/) from June 1, 2016 to June 15, 2016 to estimate the number of universities offering courses that either use or teach the R. Universities were identified as recognized institutions of higher education in Canada that offer four-year degree programs. The survey was developed to specifically address how many universities offered (a) course(s) using R and in what capacity the program was used within courses. Following research ethics approval, the survey was sent to ca. 2,500 professors in Biology, Ecology, Chemistry, Statistics, Mathematics, and Computer Science departments (considered to be the most likely sources of R usage in a university). Contact information for individual professors was obtained from departmental websites at each university in May, 2016. Only full time active faculty were sent the initial request (i.e. the survey was not sent to adjunct/emeritus professors, graduate students, or technologists). Additionally, a request was made to forward the survey to any other faculty or departments that a respondent thought appropriate or had knowledge of R usage at their particular university. The survey was formatted with conditional responses and ranged from 10 to 22 questions depending on the respondents' answers. For example, if a respondent answered "yes" to teaching R they were taken to a different section than if they answered "no" to the same question. Survey questions and a figure diagraming conditional response layout is available in Supplementary File 1 and Supplementary File 2, respectively. Following the response period, results were downloaded and analyzed to determine the extent of R usage across Canada and evaluate usage patterns.

Data analysis

Both individual question responses, as well as combined question information, were used to evaluate R usage. For example, the response rate of universities was simply calculated by taking the number of universities with at least one respondent divided by the number of universities surveyed, while the calculation of R usage at universities was reflected by the number of universities with at least one respondent that also had at least one class utilizing R divided by the number of respondent universities regardless of R usage. All data are expressed as counts and formal statistical tests were not preformed. As with any voluntary surveying method, it must be noted that positive sampling bias is potentially a factor; meaning it is probable that respondents were at least familiar with what R is and people unfamiliar with the program were less likely to take the time to respond. All analysis and plotting was carried out in R version 3.3.1[16].

Ethics and consent

Ethics approval was granted on May 20, 2016 from the Laurentian University Research Ethics Board (REB) under REB file number 2016-04-14. Consent was obtained through a participant consent statement (Supplementary File 3) and electronic approval, which lead participants to the survey. This information is available in Dataset 1[21], and only one participant opted out of taking part in the survey.

Conditional requirements of the REB were to retain the anonymity of individual participants. To ensure this, but preserve the ability to analyze and deposit data, university name information has been removed from the dataset and replaced with number designations. Additionally, all comments or other potential individual level identifiers have been removed.

Results

Overall results

Of the 2,500 professors from 70 Canadian universities invited to participate, 157 responded. Of these only one participant elected not take the survey giving a total of 156 respondents. At least one response was recorded from 61 of the universities for an 87% response rate (i.e. at least one key informant per institution). Of the 61 responding universities, 65% (40) had at least one course that used R in some manner, while 36% (22) of responding universities had courses that were either specific to the R language or used it as the primary data analysis tool. Of respondents 51% used R in at least one course. Based on the courses taught by all respondents, R was used in 26% of courses in some capacity and of the courses that used R, 16% taught the R language.

Professors who teach R

Of courses using R, 60% were offered to both undergraduate and graduate students while only 8% were graduate-only, and the remaining 32% undergraduate-only (Figure 1). By far the most frequent use of R in the classroom was geared towards statistics, followed by courses explicitly focusing on the R language itself and ecological modeling, respectively, (Figure 2). Professors who

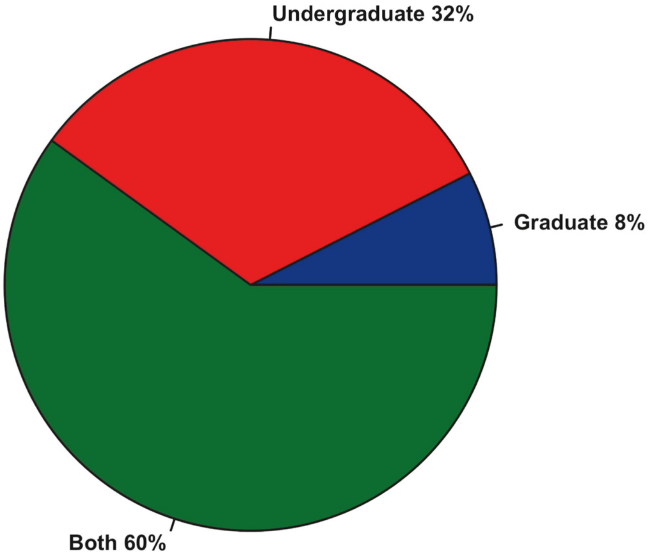

Figure 1. Distribution of R classes. Breakdown of course offerings for 80 professors who teach with R, where "both" means a class contains undergraduates and graduate students or the professor teaches both an undergraduate and graduate course using R.

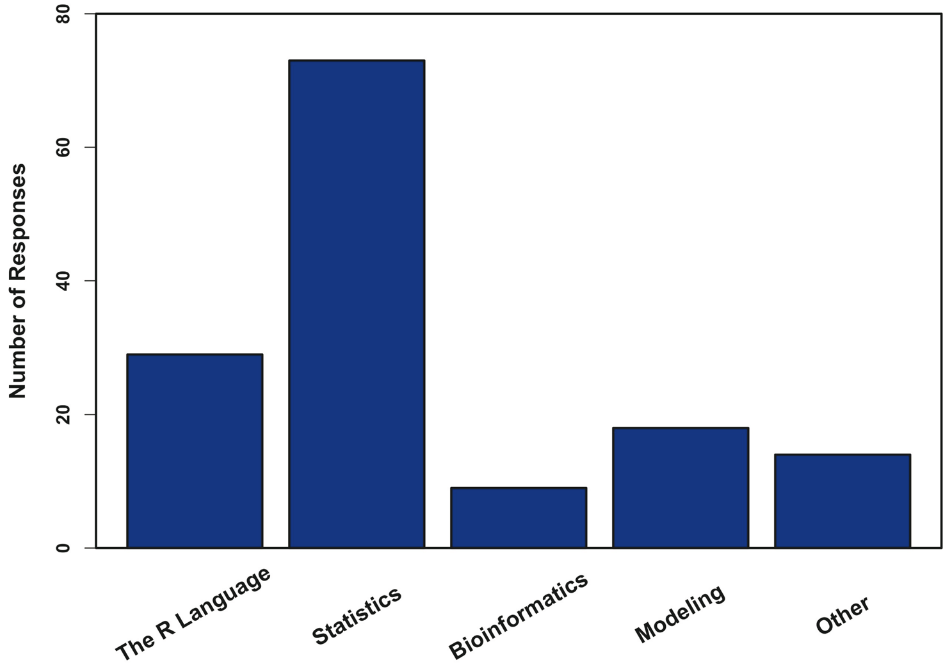

Figure 2. Subjects taught in R. Responses from 80 professors who teach R, regarding the subjects they teach in their courses that use R (multiple responses were allowed). Other includes climatology, population genetics, econoinformatics, and plotting.

taught R felt the biggest advantages included that it is free, followed multiple platform support, diverse packages, and being open source; the latter three were all weighted similarly (Figure 3). Cited disadvantages to teaching R were dominated by a steep learning curve, followed by the students not actually learning the language itself (e.g. using code that is "plug and play" and not written or altered by students; Figure 4).

Professors who do not teach R

A total of 76 professors did not teach with R at all. The most common reasons for not teaching with R are presented in Figure 5. Key reasons for not teaching R included teaching non-analytical courses or being unfamiliar with R. Many of the "other" responses included what could be classified as "departmental issues" (e.g. lack of time, perceived difficulty of learning R vs. programs like Excel, cooperation in coordinating between courses/professors). Professors who used R in their own research, but don't teach R, were more open to teaching R in the future when compared to professors who were unfamiliar with R (Figure 6). Overall, the majority of professors were open to teaching a class using R in the future.

Professors who use R themselves

Professor usage of R in research did not clearly reflect them teaching (with) it in the classroom. Figure 7 shows four groups based on whether professors taught and/or used R themselves in their research. The majority of professors (66%) used R themselves, while only 51% of professors actually taught R. In total, 19% of professors who used R themselves did not teach it. Professors who used R tended to use only R, but SAS/SPSS and MATLAB were also used along with an assortment of other programs (Figure 8). In comparison to reasons to teach with R, professors who used R still felt it being free was a good reason to use it, but also placed more emphasis on packages and it being a discipline standard (Figure 9). All professors who used R (100%) did so for descriptive statistical analyses, while modeling and figure generation were other common uses (Figure 10).

R by subject area

Of the 156 respondents 154 indicated a department affiliation grouped into biology/life sciences, math/statistics, and others, including professors who had multiple appointments in biology, math, stats, and/or were in completely unique departments, e.g. decision sciences. A total of 64% of respondents were in the biology/life sciences, and 48.5% taught R. Professors who identified with math and stats departments made up 27.5% of respondents and 56% taught R. Of professors who were in statistics alone, 100% taught R in at least one course. The remaining 8.5% of respondents were categorized as "others" and 54.5% taught at least one class using R.

Dataset 1. R survey results

http://dx.doi.org/10.5256/f1000research.10232.d144345

Raw data from the survey questions with the university names converted to numbers and other potential respondent identifiers removed.

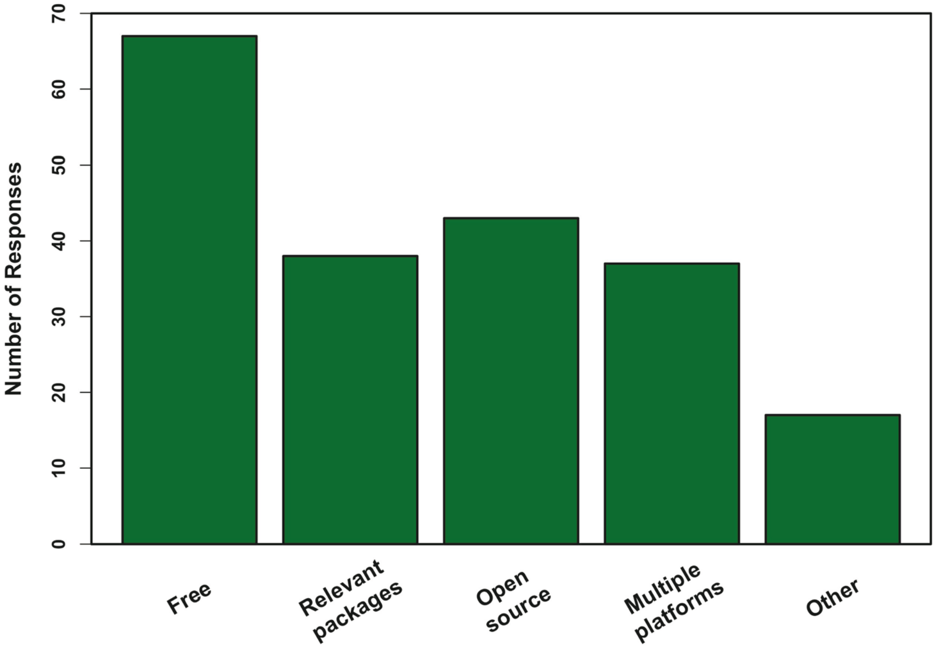

Figure 3. Advantages of R. Responses from 80 professors who teach R, regarding the biggest advantages to using R in the classroom (multiple responses were allowed). Other includes facilitates problem solving, teaches job applicable skills, the R community, graphics, flexibility, and reproducibility.

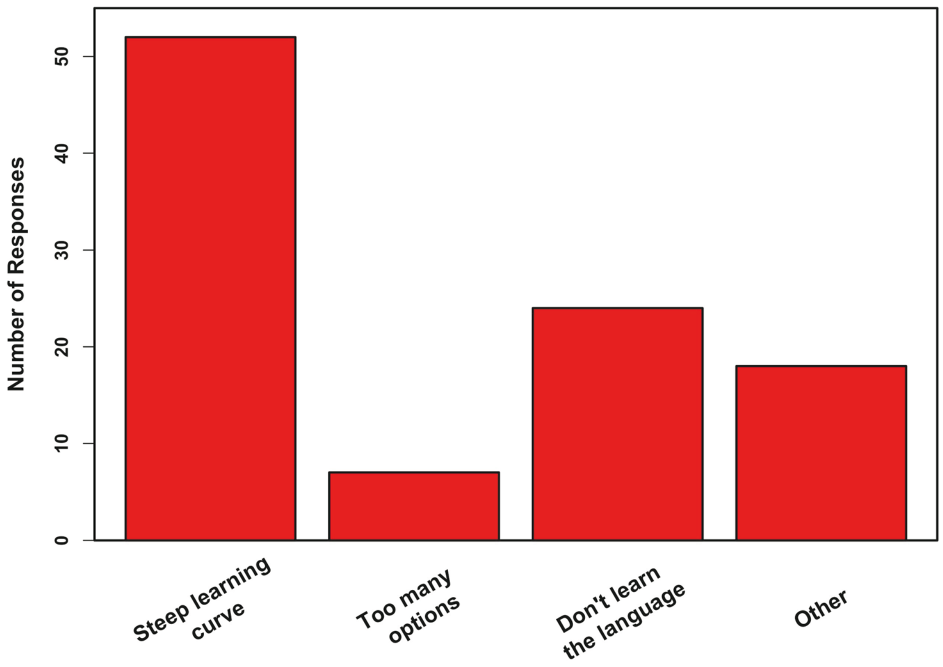

Figure 4. Disadvantages of R. Responses from 78 professors who teach R, regarding the biggest disadvantages to using R in the classroom (multiple responses were allowed). Other includes requires coding, colleagues cooperating, pushback from SAS users, students using multiple platforms in classroom, mainstream texts lack R examples, and R is used less in industry.

Discussion

The R language is beginning to make its way into Canadian universities with a wide range of courses spanning both graduate and undergraduate levels already in place. Over half of Canadian universities offer at least one course that uses R, but these courses are often not geared at the R language specifically, greatly diminishing the benefits to students. While a number of universities did offer multiple classes that use R, this was the exception and

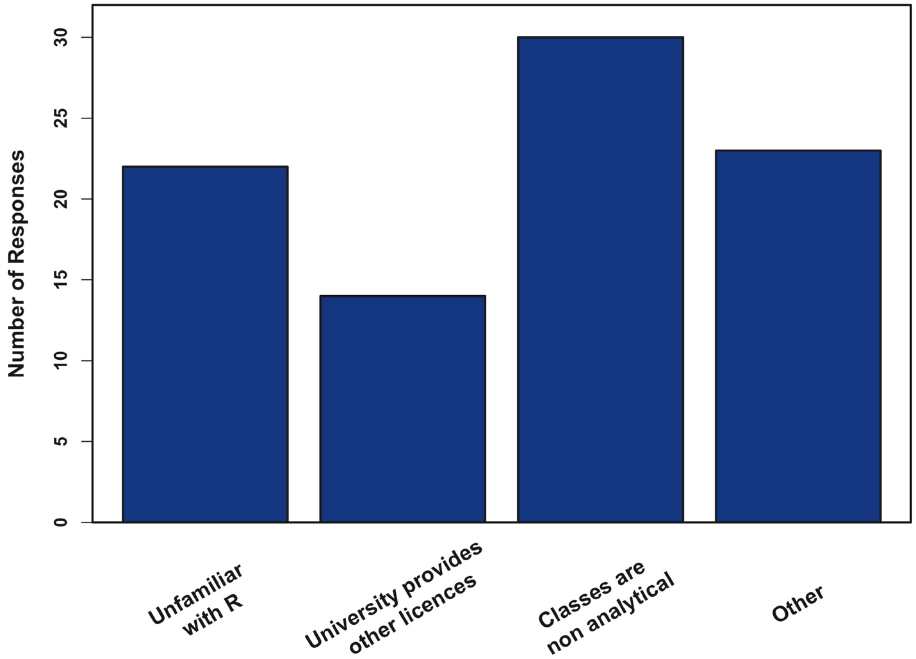

Figure 5. Reasons to not teach R. Responses of 76 professors who don't use R in any classes (multiple responses were allowed). Other includes time restrictions and classes that are already using other stats programs with limited departmental cooperation on switching over.

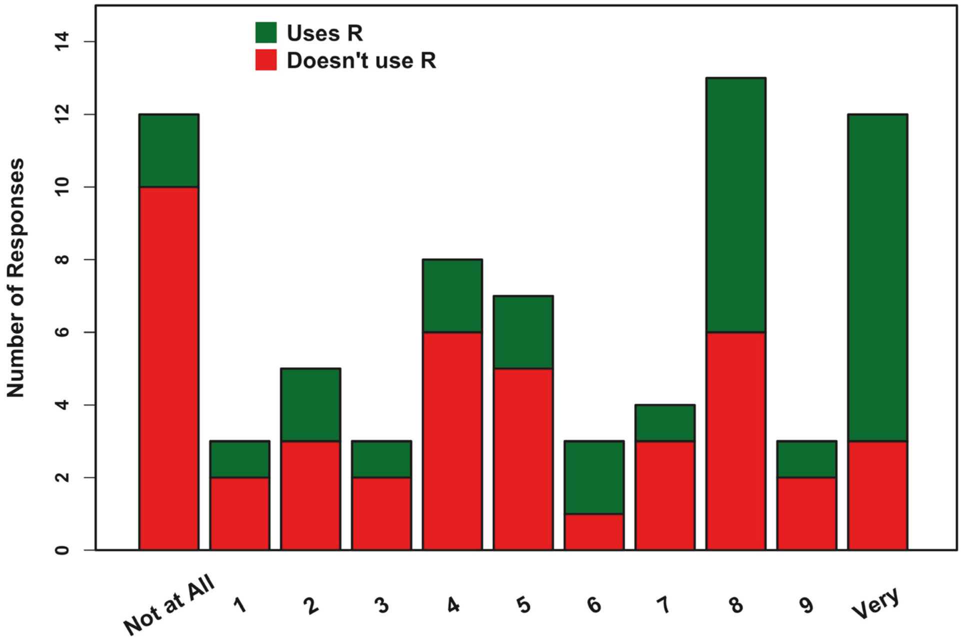

Figure 6. Openness to teaching R. Responses of 73 professors who don't use R in any classes, regarding their willingness to use R in future classes. Groupings are by professors who use R in their research, but don't teach it (green), and those who don't teach or use R themselves (red).

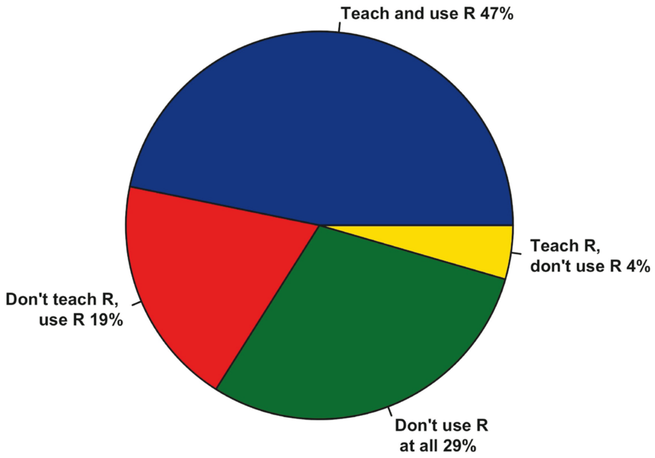

Figure 7. Professor R usage. Summary of how 156 professors interact with R. Note the large portion of professors who don't teach R, but use it in their own research.

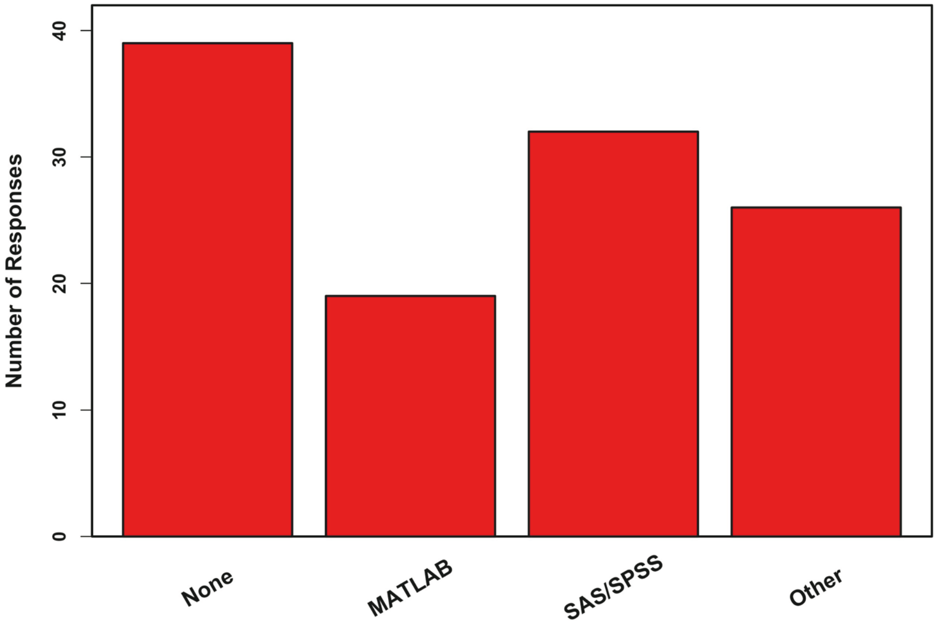

Figure 8. Use of other programs. Response of 100 professors who use R (could have multiple answers). Other includes Excel, LINDO, BMDP, Prism, PAST, MEGA, Statistica, Sigmaplot, Stata, JMP, DataDesk, Systat, STAN, OpenBUGS, Minitab, Mathematica.

not the norm, indicating that R is not being adopted by professors and expanding throughout Canadian universities as fast as it perhaps should be. There appeared to be a positive sampling bias towards people who use R themselves; meaning a professor who was unfamiliar with R was unlikely to respond to the survey, however this is not uncommon in surveys of this type[22]. Taking this into consideration, it is likely that these results represent the current state of R usage at Canadian universities relatively well. There was a diverse range of professor's experience with R as well as the subjects being taught using R. This reflects common trends in the R community where the language has been adapted beyond a statistical tool for use in an array of applications, for example interactive maps (rMaps package) and developing applications[23]. Taken together, both professors who currently teach and those who do not teach R need to consider new ways to adapt their coursework to include R in interactive and engaging ways.

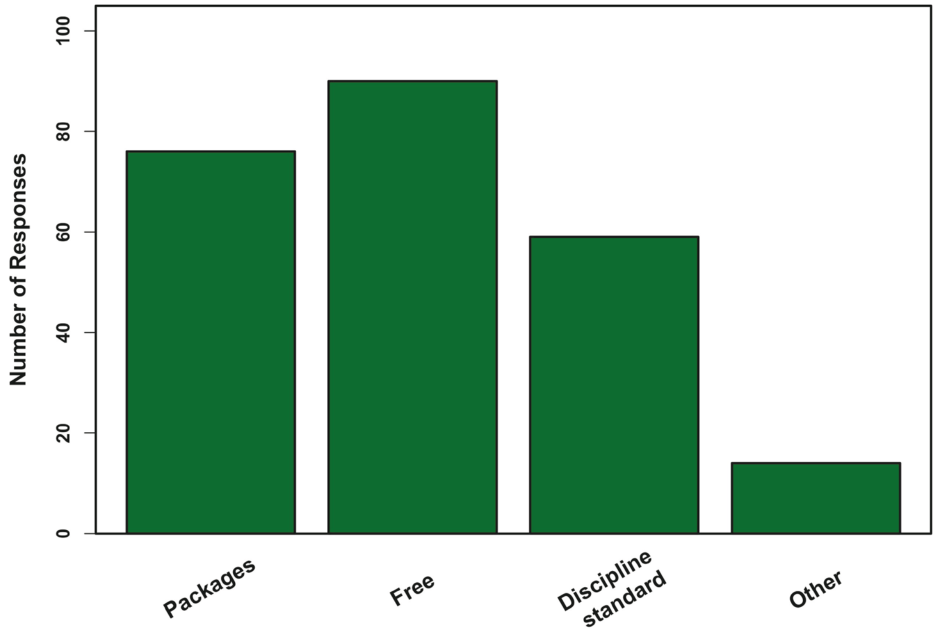

Figure 9. Why use R? Reasons 103 professors use R themselves (multiple responses allowed). Other includes new code/package development, multiple platform support, and user configuration/flexibility.

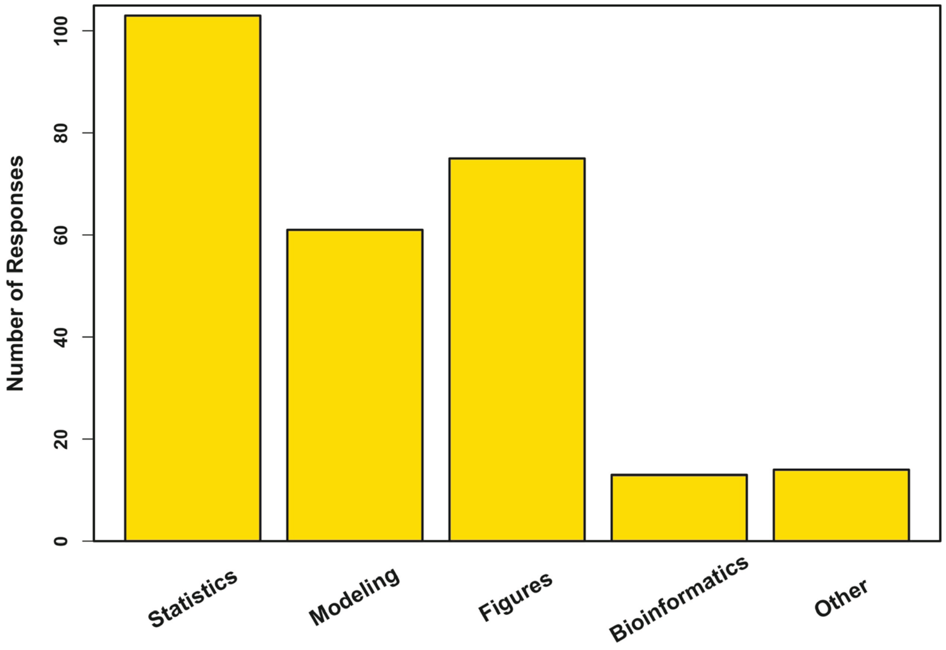

Figure 10. How R is used. Uses of R for 103 professors. Other includes data manipulation, simulations, and data exploration.

By far the most common application in the classroom was statistics, which is likely due to the origins of R being geared at that community[3]. Bioinformatics usage was a less common theme, but this is an area that will likely see significant growth in the coming years with a large amount of new package development promoted by Bioconductor (collection of packages specific to bioinformatics usage) prompted by the drop in DNA sequencing cost and rapid increase in sequence data being produced (NCBI). While the number of courses taught that explicitly teach the R language is perhaps lower than ideal, it must be noted that courses dedicated specifically to the R language may be a lofty goal, and incorporating R into courses in any manner is a useful learning exercise. This is also in line with the general need of more computer literate students regardless of academic discipline[24]. Overall, there weren't many professors that responded who didn't teach or use R themselves. The most common reason for not using R was that their classes were non-analytical. This appears justifiable, as some subjects rely less on data management and analysis. However, large portions of professors in this category were totally unfamiliar with the capabilities of R and it may be that they don't realize that R and packages within R are not exclusively focused on descriptive statistical analyses. For example, modeling transmission of a pathogen in a virology class or movement of animals across an ecosystem in ecology classes could both be incorporated in labs in these courses using R. The importance of adapting course material to match current trends in technology is highlighted in other research that retrospectively are easy to understand the importance of early adoption into the classroom[25–27]. For example, the broad movement from handwriting to typewriters to computers or the change from film to digital cameras and the finer resolution examples of software, which is updated on a more frequent basis. While preservation of older technologies is important, keeping students at the cutting edge of technologies and the programs/systems that operate them are key to current education. Along this same line the most concerning reasons for not teaching R included time restrictions and/or limited departmental cooperation, as well as general apathy towards adapting course material[28]. To us, these are potentially poor excuses for not altering courses to expose their students to a useful, widely accessible tool and emphasize a general lack of professor engagement, which is detrimental in the classroom[29].

Bringing R into the classroom has a number of advantages. First, it is free, so does not strain student or department budgets and is compatible with multiple platforms (Mac, PC, Linux) allowing students to download it on their personal computer instead of having to do assignments on university computers with restricted licenses. Second, it is also open source and has a large support community online with a number of forums to address virtually any sort of problem (e.g. Stackoverflow). Third, a major advantage to students is the current applicability of R in the classroom and beyond. The near exponential growth of R[6,7] highlights the importance of learning the language and is indicative of a desirable skillset across academic disciplines and career paths. This is due in part to the adoption of R in many areas outside of academia, but also because R (and coding languages in general) is a skillset that many employers look for in a potential employee. That is to say learning to code is desirable for today's students largely due to the fact that coding is a skill that is transferrable between languages and a process that teaches critical thinking and problem solving[20,30,31]. So even if a student never codes again, the process of learning to code may benefit the way they approach future work. It is worth noting that with the advantages come some disadvantages, the largest being a "steep learning curve". However, as sociologist John Fox[6] points out this is really in comparison to the point-and-click types of software that students are used to. In reality R is a relatively easy coding language to learn once the basic conventions are mastered, making it accessible to novice programmers.

The feasibility of introducing R into the classroom is highlighted in our study by the fact that many professors who don't teach R are open to teaching it in the future. Furthermore, it is possible to teach classes in R even if the teacher does'nt use it themselves, and we showed that a number of professors who don't use R themselves already teach R. After all there are numerous other skills professors pass on to students that they themselves don't use on a regular basis, if at all (e.g. a professor teaching an introductory course would typically only research on a very small subset of what they teach). Of particular interest are the professors who use R themselves, but don't teach R. This group could be a catalyst for universities and/or departments to introduce R into course material, greatly expanding the number of courses offering R and the subject areas using R. Willingness of adopting new technologies in the classroom is a common hurdle[32], but fortunately many professors in these positions are open to teaching R in the future, they just need to find the motivation to bring new material into their classroom[28]. Admittedly it takes time and effort to adapt a class that is already "refined" and it can be difficult to be the first to take that step within a department or institution[28,33], but professors should realize that the benefits greatly outweigh the costs, and can take the time to gradually begin to incorporate R into their course content. For example, a professor could promote R over other "less useful" programs (e.g. Excel), even if R will only be used for minor assignments, such as mean calculations and basic plotting. Then expansion of material could be done incrementally throughout the semester from the student's perspective and across multiple years of lectures from the professor's perspective. Additionally, professors should expand their own R knowledge and look for the new and exciting ways R is being used. R is no longer a purely "analytical" tool and lab courses could, for example, use R for lab report writing (markdown[34] is great for this), including all aspects of data management, plots, and text all in one file.

Individual comments provided valuable insights into problems with R in education, and the "learning curve" was a common theme amongst users and non-users both personally and for their students. As discussed before, this is in our view a misperception promoted by comparing R to "point and click" programs. While R is not as intuitive initially, once a foundation is established the subsequent adaptability and power over point and click platforms are large. Recently there has been an expansion of resources available to learn R in a fun and interactive way (e.g. Datacamp and swirl package[35]). These could serve as useful companions to professors looking to use R in their classroom as an effective way of "outsourcing" much of the initial learning process. Furthermore, it is our general thought that the R community needs to expand the currently available startup material to get people familiarized with R in a more interactive way. More specifically we feel that the R education community would greatly benefit from a more centralized location for material related to course content and examples of lesson plans that incorporate R. While some examples of this are available through sources like GitHub, these are collections of individual educators and there is no comprehensive location for educational material related to R. At an institutional level some professors suggested the idea of workshops, which are a great tool in university settings[36]. These can range from a weekend crash course to a semester long in depth introduction, which sets students and professors that are new to R on the right path from the beginning. From our personal experience, the lead author is in a trial period of teaching an R workshop, which is open to graduate, upper level undergraduate, and faculty, using hours normally devoted to teaching undergraduate labs, which is being met with positive reviews. Other comments indicated that these less formal forms of instruction may be a way to promote R in universities, ideally leading to a broader acceptance through time.

Conclusions

It is apparent that Canadian universities are beginning to put the R language to practice in classes with nearly 2/3 of the responding universities offering at least one course that uses R. However, fewer courses teach classes that are more specific to learning the language itself. While this is a good start to exposing students to R, it appears that Canadian universities in general are lacking R-based coursework. To our knowledge, there are no similar data for R usage at universities in other countries, but a comprehensive understanding of R usage in all levels of academics is necessary and would provide critical insights. Future work could use surveys to identify broad R usage trends as we did, but would benefit even more from obtaining detailed information from syllabi or course material itself. Surveys do depend on people's willingness to participate so perhaps individual case study reports from departments or individual teachers who have incorporated R might be of use, encouraging others to put forth the effort and use R in the classroom. Based off broad data on downloads and references to R, it is apparent that R is rapidly becoming a programming and data analysis language of choice for researchers, academics, and in industry. With this in mind it is in an institutions' and students' best interest to promote R in coursework among all of the STEM disciplines. Furthermore, the only "cost" to a university, department, or educator is the time required to rework course material into the R language. While this takes initial effort, we feel that the long-term benefit to students greatly outweighs this initial input. The R community is rapidly developing more "user friendly" graphical user interfaces and will continue to be at the forefront of data analysis and presentation for the foreseeable future. Without doubt, an understanding of R will benefit students beyond their coursework in postgraduate and professional settings.

Data availability

Dataset 1: R survey results. Raw data from the survey questions with the university names converted to numbers and other potential respondent identifiers removed. DOI: 10.5256/f1000research.10232.d144345[21]

Author contributions

MC conceived the study, wrote and distributed the survey, analyzed data, and wrote the manuscript. NB commented and revised the survey questions and design, guided analysis, and edited the manuscript. Both authors have agreed on the final content of the manuscript.

Competing interests

NB is an active professor at a Canadian university and had the opportunity to participate in the survey. The authors do not feel his participation, or lack of participation, compete with the interest or conclusions of the study.

Grant information

MC is supported by the Ontario Trillium Foundation through an Ontario Trillium Scholarship for International PhD Students, by fellowships from Laurentian University, and by NSERC Discovery (#341381) and a Canada Research Chair Grant (#228457) held by NB.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to thank all of the participants who took the time to complete the survey and those who responded with valuable additional information about R usage at their university and in their classrooms. A travel grant was provided by the Laurentian University Graduate Student Association to support a poster presentation of the results at the 2016 UseR! conference in Stanford, CA, USA, and we recognize the valuable input we received from attendees that helped develop this manuscript into a comprehensive and cohesive survey of R usage in the Canadian university setting.

Supplementary material

Supplementary File 1. Survey questions. A list of all questions that participants could have been asked. Some questions are repeated for the conditional response survey. See Supplementary File 2 for question pathways.

Click here to access the data.

Supplementary File 2. Question flow diagram. Diagram showing potential survey "paths" with conditional responses, numbers correspond to questions in Supplementary File 1.

Click here to access the data.

Supplementary File 3. Participant consent statement. The full consent statement that participants agreed to prior to taking the survey, approved by the REB.

Click here to access the data.

References

1. Sussman GJ, Steele GL Jr: **Scheme: A interpreter for extended lambda calculus.** *Higher-Order Symb Comput.* 1998; **11**(4): 405–439.
   Publisher Full Text
2. Becker RA, Chambers JM, Wilks AR: **The New S Language: a programming environment for data analysis and graphics.** Chapman and Hall, 1988.
   Reference Source
3. Ihaka R, Gentleman R: **R: A Language for Data Analysis and Graphics.** *J Comput Graph Stat.* 1996; **5**(3): 299–314.
   Publisher Full Text
4. Vance A: **Data Analysts Captivated by R's Power.** *New York Times.* 2009.
   Reference Source
5. CRAN: **The Comprehensive R Archive Network.** 2016.
   Reference Source
6. Fox J: **Aspects of the Social Organization and Trajectory of the R Project.** *The R Journal.* 2009; **1**: 5–13.
   Reference Source
7. Rapporter: **R activity around the world.** *R-bloggers.* 2014.
   Reference Source
8. James: **Where is the R Activity?** *R-bloggers.* 2013.
   Reference Source
9. Muenchen RA: **The Popularity of Data Analysis Software.** 2016.
   Reference Source
10. Heron MJ, Hanson VL, Ricketts I: **Open Source and Accessibility: Advantages and Limitations.** *J Interact Sci.* 2013; **1**: 2.
    Publisher Full Text
11. Khan MA, Urrehman F: **Free and Open Source Software: Evolution, Benefits and Characteristics.** 2012; **1**(3).
    Reference Source
12. Weber S: **The Success of Open Source.** Harvard University Press, 2004.
    Reference Source
13. Harrison E: **RStudio and GitHub.** *R-bloggers.* 2015.
    Reference Source
14. Chambers JM, Lang DT: **Omegahat Packages for R.** *R News.* 2001; **1**: 1–32.
    Reference Source
15. Freeman S, Eddy SL, McDonough M, *et al.*: **Active learning increases student performance in science, engineering, and mathematics.** *Proc Natl Acad Sci U S A.* 2014; **111**(23): 8410–5.
    PubMed Abstract | Publisher Full Text | Free Full Text
16. The R Core Team: **R: A language and environment for statistical computing.** 2016.
    Reference Source
17. Sarkar D: **Lattice: Multivariate Data Visualization with R.** Springer, 2008.
    Publisher Full Text
18. Whickham H: **ggplot2: Elegant Graphics for Data Analysis.** Sorubger-Verlag, 2009.
    Publisher Full Text
19. Board A: **Why Undergraduates Should Learn the Principles of Programming Languages.** *Language (Baltim).* 2010; 1–9.
    Reference Source
20. Pea RD, Kurland DM: **On the cognitive effects of learning computer programming.** *New Ideas Psychol.* 1984; **2**(2): 137–168.
    Publisher Full Text
21. Carson M, Basiliko N: **Dataset 1 in: Approaches to R education in Canadian universities.** *F1000Research.* 2016.
    Data Source
22. Sax LJ, Gilmartin SK, Bryant AN: **Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys.** *Res High Educ.* 2003; **44**(4): 409–432.
    Publisher Full Text
23. Baier T, Neuwirth E, De Meo M: **Creating and Deploying an Application with (R) Excel and R.** *The R Journal.* 2011; **3**(2): 5–11.
    Reference Source
24. McDonald DS: **Computer Literacy Skills for Computer Information Systems Majors: A Case Study.** *J Inf Syst Educ.* 2004; **15**(1): 19–33.
    Reference Source
25. Gillard S, Bailey D, Nolan E: **Ten Reasons for IT Educators to be Early Adopters of IT Innovations.** *J Inf Technol Educ.* 2008; **7**: 21–33.
    Reference Source
26. Guzey SS, Roehrig GH: **Teaching science with technology: Case studies of science teachers' development of technology, pedagogy, and content knowledge.** *Contemp Issues Technol Teach Educ.* 2009; **9**(1): 25–45.
    Reference Source
27. Rogers EM: **A prospective and retrospective look at the diffusion model.** *J Health Commun.* 2004; **9**(Suppl 1): 13–19.
    PubMed Abstract | Publisher Full Text
28. Hodas S: **Technology Refusal and the Organizational Culture of Schools.** *Educ Policy Anal Arch.* 1993; **1**(10).
    Publisher Full Text
29. Fink LD: **Creating Significant Learning Experiences: An integrated approach to designing college courses.** John Wiley & Sons Inc., 2013.
    Reference Source
30. Akcaoglu M: **Learning problem-solving through making games at the game design and learning summer program.** *Educ Technol Res Dev.* 2014; **62**(5): 583–600.
    Publisher Full Text
31. Robins A, Rountree J, Rountree N: **Learning and Teaching Programming: A review and discussion.** *Comput Sci Educ.* 2003; **13**(2): 137–172.
    Publisher Full Text
32. Gbomita V: **The adoption of microcomputers for instruction: Implications for emerging instructional media implementation.** *Br J Educ Technol.* 1997; **28**(2): 87–101.
    Publisher Full Text
33. Conroy CA, Bruening TH: **School subcultures as factors affecting technology refusal: An examination of applied academics implementation in Pennsylvania and resulting implications for agricultural teacher education.** 1994; **21**.
    Reference Source
34. Allaire JJ, Horner J, Marti V, *et al.*: **markdown: 'Markdown' Rendering for R.** 2015.
    Reference Source
35. Kross S, Carchedi N, Bauer B, *et al.*: **swirl: Learn R, in R.** 2016.
    Reference Source
36. Mjelde L: **The magical properties of worksop learning.** Peter Lang, 2006.
    Reference Source

Open Peer Review

Current Referee Status: ✓ ? ?

Version 1

Referee Report 15 December 2016

doi:10.5256/f1000research.11021.r18125

? **Luc F. Bussiere**
Biological and Environmental Sciences, University of Stirling, Stirling, UK

This article provides useful survey data on aspects of instruction using the R programming language at Canadian Universities. The authors report intriguing data on the numbers of respondents who use R for teaching and research, the subject areas in which the respondents work, and their willingness to teach future classes using R.

These data provide a useful glimpse of the adoption of R software in Canadian Universities, and the transparent inclusion of the survey and data makes this publication a valuable addition to the literature. My comments below are intended to provoke further critical analysis if possible.

Although I am sympathetic to the authors' opinions (as an instructor who uses R in my own research and teaching), I am not consistently convinced that these data support the authors' conclusions, even though those are made somewhat tentatively. My skepticism comes from a few sources, as detailed below. I think most of my concerns could be addressed though a follow up survey and additional analyses.

Much of the discussion is devoted to the argument that we need more teaching of R (especially in classes dedicated to the programming language itself, rather than its applications). I do not object to this assertion in principle (teaching with R has personally been a rewarding experience for me and most of my students), but the conclusion does not derive from the survey data, and the logic that underpins it is not always clear. The authors cite some pedagogical papers on the general importance of programming knowledge, but the relative value of programming per se (as opposed to its applications) for disciplines apart from computing science are not self-evident given the assumed cost to other portions of the curriculum. One could indeed use R markdown for lab report submissions, as the authors suggest on p.10, but I am not convinced that this would often be worth implementing if the main learning outcome sought is written communication skills. I think it would be useful if the authors could more clearly separate the discussion that derives directly from their survey findings from those that represent advocacy of a particular pedagogical opinion.

As the authors acknowledge, there is a risk of positive bias in their survey because respondents unfamiliar with R may have been less likely to respond. The importance of the bias could be estimated through attempts to contact nonrespondents, and contrasts of the scores with the original surveys, and methods for computing estimates of response survey quality seem to be reasonably well established and (of course) have been developed for analysis with R[1]. Such an effort could help clarify the importance of biases in this study.

For a paper about a language developed explicitly for conducting statistical analyses, the lack of statistics is quite jarring. The authors draw many conclusions about differences among categories of response based on apparent patterns, but it would be quite useful to know how much confidence we should place in the relative numbers of responses. Like the analyses of survey quality mentioned above, methods for conducting multinomial models and extracting multinomial CIs are readily available within R (e.g., see Villacorta 2012), and would allow the authors to both quantify uncertainty in their proportions and illustrate confidence limits for each response measure.

Some of the comparisons suffer from a lack of context. For example, Fig. 1 concerns the relative provision of R courses to undergraduates vs graduate students, but this contrast is difficult to interpret without more information on the number of courses in total that are offered to graduates and undergraduates. Is the rate of provision higher at the graduate level, given the smaller number of total courses on offer? I wonder if the authors can hint at the answer by assessing numbers of courses in each category at a few institutions.

In addition to a dissatisfying lack of measures of confidence in effects, the figures are not consistently laid out to permit effective consideration of the data. For example, in Figure 6, the key response variable is a scaled measure of willingness to teach R in future classes, but that variable appears on the x-axis instead of the y. Since the most meaningful contrast is between users and non-users of R, the authors could produce a plot that illustrates the numerical response scores in the two groups (e.g., in a strip chart) along with a measure of means and confidence limits: such a presentation would support the presumed difference much more persuasively, in my opinion, than the current layout.

Minor comments:

I spotted a few typographic errors, including the use of the word “preform” for perform on pp. 3 and 4, and “does’nt” on p. 10.

The Education Board that Authored citation 19 is incorrectly attributed as if it were a single author, whereas there are 8 individuals listed as authors on the report who could be acknowledged.

References

1. de Heij V, Schouten B, Shlomo N: RISQ manual: Tools in SAS and R for the computation of R-indicators and partial R-indicators. *Representativity Indicators for Survey Quality*. 2010.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 14 December 2016

doi:10.5256/f1000research.11021.r18120

?

Eliezer Gurarie

School of Environmental and Forest Sciences, University of Washington, Seattle, WA, USA

I feel odd reviewing this paper - since I have little technical expertise in assessing human survey-based studies. My interest in this topic is as a highly biased object of the study, specifically, as an enthusiastic user of R in the classroom, both for statistics and ecological modeling courses. [Is this what the animals I do research on feel like when they read my papers?]

Overall, the information presented is useful. The survey appears to have been conducted responsibly, with a reasonably high response rate (though with possible self-selection bias). The introduction provides a useful overview of the history and context of R's and the discussion is comprehensive and thoughtful. The availability of the survey results is a welcome contribution.

While the main purpose of the article seems to be to argue and advocate for the use of R in classrooms, it is not always clear how the survey results inform that argument. On the one hand, the fact that there are courses in 65% of institutions and across fields might encourage other professors to adopt R. On the other hand (given my bias) it is a shocking disservice to students that 35% of institutions use R in ZERO courses. There was a bit of a missed opportunity in that the article presents just a snapshot in time. It would have been interesting to see how the rate of R use has increased (which would have been possible by asking professors about their use of R 5 or 10 years ago). I imagine the rate of increase would have been very nearly explosive.* In any case, the claim that "R is not being adopted ... as fast as it perhaps should be" is more of an opinion (even if softly put, and one I completely agree with) than supported by the results. There are other slight disconnects between claims in the discussion and the survey, but then for this kind of pseudo-advocating article/essay this is perhaps more acceptable than what I am used to.

* - As a single datapoint: when I first proposed incorporating R into an introductory undergraduate statistics course at the Unversity of Washington in 2012 - not even five years ago - the idea was met with surprise and some scepticism by other instructors that the students could "hack it", The experiment ended up being an unequivocal success, with many students claiming that was the most useful portion of the course [certainly compared to looking up t-values in a table at the back of a textbook!] and I believe is now standard in the curriculum.

Results and Figures:

The weakest point in this paper is the ugliness of the figures (which is ironic, considering that one of the main selling points of R is the ability to make beautiful graphics). I understand that the results are simple counts, but the presentation could still be improved. Figure 1 is completely unnecessary, unless it were cross-tabulated against, e.g., subject of course (an important missing bit of information), for example sorted into statistics/ mathematics/ computer science vs. life/ social sciences.

In almost all of the bar plots, you could use horizontal bars, ordered top to bottom from highest count to lowest count, and go ahead and include all of the "Other" categories [e.g., in figure 2, climatology, population genetics, econoinformatics, and plotting]. Those results are interesting, and there's plenty of room if you abandon the fat vertical bars. These could also be cross-tabulated and color-stacked against subject, or at least "graduate" / "undergraduate".

Figure 6 (though seasonally appropriate) would be much improved if it were presented as a mosaic plot (i.e.: mosaicplot(table(R.Use, Willingness))), which is much better for comparing the relative shift across categories, while reflecting the sample sizes as well. It is, incidentally, interesting that so few people answer "9" compared to "8" (I guess 9/10 of "Very" is a more slippery concept than 4/5 of "Very"!) There's a psychological effect here somewhere, but in the meantime you might be better off pooling 1-3, 4-6, 7-9.

I must confess I would have no idea how to answer the question in Figure 10 - there is so much overlap. I really don't see how one can separate "Modelling" (and, often, "Statistics") from "Data manipulation" "Simulation", "Data Exploration", "Visualization", etc.

Discussion

Among the tools which facilitate the use of R in the classroom, one of the most important is the use of 'knitr' and 'Rmarkdown' to easily generate documents that combine text, math, code, figures and output. This is a very important omission. Perhaps the single most practical use of "knitting" documents is for teaching material - including lectures, labs and homework assignments - in particular for learning R. Also, report generation itself is a useful and totally accessible skill to teach, particularly considering the importance of reproducibility of analysis (another important advantage of R over point-and-click tools).

Minor:

Discussion - there's a "does'nt" that should be "doesn't".

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 12 December 2016

doi:10.5256/f1000research.11021.r18124

✓ **Colin W. Rundel**

Department of Statistical Science, Duke University, Durham, NC, USA

The authors describe the results of a survey of Canadian academics on their use of R in their courses and in their own work. While there is good information on the increasing popularity of R on the web, in industry, and in scholarly articles there is far less information on how R is being taught. It is in this important area that the paper provides some much needed insight.

In particular, I think many researchers would be surprised to find courses being taught using R in more than half of Canada's universities. Giving other educators this kind of information is tremendously valuable in inducing other educators to also decide to make the jump to R. In particular, being able to point to other universities and courses where R is being successfully taught is a strong argument against common complaints like the learning curve being too steep.

While a more systematic examination of R offerings across universities would have more reliable results than this survey, I believe that it still offers valuable (if potentially slightly biased) insights into the basic patterns of R education. The authors results are very encouraging to me as a educator interested in teaching R, but they also show that there is much more we can do in promoting R at other universities as well as growing our own course offerings locally.

Major Comments

Introduction

- I think the history as stated underplays the importance of R being free software (particularly as compared to S), this aspect also is clearly a hugely influential factor for professors based on the survey results.

- *there have been more packages added in 2015 than have existed in all of the SAS institute's history* - this comparison is based the number of packages added to CRAN vs a rough estimate of the number of procs contained in SAS 9.3 (from r4stats.com). This is a weak comparison originally which is then confusingly stated in the paper.

- While touched on tangentially I think reproducible research is worth mentioning explicitly, particularly in reference to the strength of programming languages vs. point and click tools.

Results

- *Based on the courses taught by all respondents, R was used in 26% of courses in some capacity and of the courses that used R, 16% taught the R language*. I believe that this would be more interesting to see broken down based on the respondent's field. In general, additional cross tabulation by discipline would give more insight into the data.

- Figure 1 is not really needed, giving the values in the text is sufficient in my mind. Also it is somewhat confusing about which subset of classes this breakdown applies to - this is generally true for many of the other figures. See my Figures comment below.

Discussion

- *indicating that R is not being adopted by professors and expanding throughout Canadian universities as fast as it perhaps should be*. I don't entirely follow the logic here, it is not clear how to establish how fast it is or should be expanding. While I don't disagree with the sentiment, this comes across as an unsupported opinion.

- *Taking this into consideration, it is likely that these results represent the current state of R usage at Canadian universities relatively well.* Again, this comes off more of an opinion than what is supported by the survey results and a claim like this needs additional support. The later conclusions are not invalided by removing this claim.

Figures

- Most figures could be shrunk considerably without negatively affecting readability, e.g. Figs 2-5. In some cases it might improve readability to combine plots into facets within a single figure (connecting figures to subsections).

Minor Comments

Introduction

- This is the first I've ever heard of omegahat, seems to be on a very different scale than github or even r-forge.

- While not without their issue, it seems worthwhile to also mention R's mailing lists and special interest groups.

Methods

- *A survey of 70 Canadian universities was conducted using Google Forms (https://www.google.com/forms/about/) from June 1, 2016 to June 15, 2016 to estimate the number of universities offering courses that either use or teach* ***the*** *R*

Results

- The basic organizational structure is based subsets of the respondents, it would be helpful to indicate the size of each of these subsets. For example, n=80 for professors who teach with R is only given explicitly in the Figure 2 label.

- *Professors who taught R felt the biggest advantages included that it is free, followed* ***by*** *multiple platform support, diverse packages, and being open source;*

Discussion

- *(e.g. Datacamp and* ***the*** *swirl package)*

- *Furthermore, it is our general thought that the R community needs to expand the currently available startup material****s*** *to get people familiarized with R in a more interactive way.*

Figures

- Bar plot labels are rotated, this is not needed - it makes the labels harder to read and takes up unnecessary space.

Conclusions

- *However, fewer* ***professors*** *teach classes that are more specific to learning the language itself.**

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.