FORUM

The R Software Environment in Reproducible Geoscientific Research

PAGE 163

Reproducibility is an important aspect of scientific research, because the credibility of science is at stake when research is not reproducible. Like science, the development of good, reliable scientific software is a social process. A mature and growing community relies on the R software environment for carrying out geoscientific research. Here we describe why people use R and how it helps in communicating and reproducing research.

R is a multiplatform open-source software environment [*R Development Core Team*, 2012] that implements S, a language designed for data analysis. It provides low-level routines for data management tasks and linear algebra and high-level routines for fitting statistical models or creating complex graphs. The R engine is being developed by a small team called R core. R is extensible, and a set of more than 3500 add-on packages are being actively maintained by a similar number of developers.

Why R Works: The Extension Packages

One of the main reasons for the success of R is the possibility for developers to write extension packages, combined with the ease with which R users can access packages once they are available from the Comprehensive R Archive Network (CRAN). The step from being a user to becoming a developer is small with R. Generic components can be reused when written as functions or as classes and methods.

Many of the people who use the software actually go on to develop it themselves as well. R provides tools to journal projects, to edit and run scripts, and to organize extension code as add-on packages. R packages usually contain R functions and/or classes and methods and/or data and may contain C, C++, Fortran, or Java code. CRAN packages need to be consistent and complete. Complete means that each function exported has documented each function argument. Packages may contain R code examples, demonstration scripts, testing code, and tutorials that include R code. Consistency checks include checks on the correspondence of R code and documentation, correct evaluation of examples, running tests, and regenerating the tutorials by running the R code in them. When on CRAN, packages are automatically built for Windows, Mac OS-X, and many Unix flavors, relieving developers of the burden of multiplatform support. Package verification mechanisms make CRAN a reliable and powerful resource for users and developers.

When writing R packages, one assumes that R works in a certain way and will continue doing so. When the working of R changes, there is a chance that this change will break a package, i.e., stop an extension package from working, especially when the package was using R syntax sloppily. Although R core uses the 3500 extension packages on CRAN to verify the impact of planned changes, improvements that break some packages are at times needed. In such a case, maintainers of packages affected are notified in a timely way so that they can take action before the change in R is released and users would become affected. Keeping to these quality control procedures and improving them over time have created a strong bond of trust between R core and the thousands of R package developers around them.

Any R package necessarily depends on R itself, but a second level of dependency occurs when packages depend on other packages. Dependencies of one R package on one or more other CRAN packages may be further reasons for packages to break. Authorship of CRAN packages is clear, as is who maintains it. When an author writes a paper and wants to refer to the software used for a particular analysis, the author may cite authors of key packages, thus helping increase the package authors' citation impact. Publications about scientific software are becoming more and more common and more and more frequently cited.

Why People Use R for Geoscientific Research

A mailing list and forum called r-sig-geo (special interest group on geospatial data handling and analysis) has been active since 2003. Of all mailing lists monitored by markmail.org (which excludes the main bioinformatics list), in 2011 r-sig-geo was the most active special interest group of the R project. The number of r-sig-geo subscribers exceeds 2400, and the monthly traffic averages around 250 e-mails. Through this mailing list and forum, researchers can reach developers and contact other users. An archive with more than 13,000 messages is available to search for previous questions and answers. Software developers use the list to learn from the user community what users need, where the software needs improvement, or where better documentation or tutorials are needed.

Since 2003, there has been a concerted effort by a number of R developers to establish a set of classes and methods for spatial data (points, lines, polygons, grids), which resulted in the sp package [*Bivand et al.*, 2008]. Today, nearly 100 CRAN packages depend, directly or indirectly, on the sp package, and a further 250 suggest it. From 2010 on, there have been similar efforts for handling various types of spatiotemporal data. The package rgdal allows reading and writing grid or polygon data in any of the more than 180 formats supported by the Geospatial Data Abstraction Library (GDAL) and the OGR Simple Features library.

A typical work flow might consist of (1) reading data, (2) manipulating data, (3) analyzing data, computing statistics, and generating figures, maps, and/or graphs, and (4) exporting all results to a manuscript. A strong reason R is the environment preferred by many is that it offers all needed functionality, and so it provides a fast solution in terms of research effort. Recent CRAN packages that provide topology operations for vector geometry or that analyze large raster files out-of-memory (see http://CRAN.R-project.org/package=raster) have decreased the need for external (geographic information system (GIS)) tools. Nevertheless, interfaces to external GIS (e.g., Grass, SAGA, ArcGIS; see http://cran.r-project.org/package=RpyGeo) exist.

Having a scripting environment with full documentation and source code available also makes R an excellent tool for teaching. With increasing use, the trust of users in the software increases. In addition, when researchers use R, research becomes easier to reproduce.

R for Reproducible Research

An often heard argument from modelers and developers who implement or reimplement complex work flows into an R work flow is the wish to have the whole process in a single environment. With their myriad of file formats and data models, geoscientific data come in many complex forms. When modeling strategies consist of several relatively complicated tools (databases, GIS, image analysis tools, statistical software, stand-alone models) and contain several conversions, the work flow becomes hard to maintain and hard to port to new architectures, and research findings become hard to reproduce. When interactive, manual operation ("mouse clicks") is part of the work flow, guaranteeing that reproducibility becomes even more difficult. In a scripting environment, one can follow and analyze the subsequent analysis steps taken and, when needed, find help in the documentation or, if that is not sufficient, in the source code.

How many researchers can pick up a paper they published more than 10 years ago and reproduce the analysis? R provides this in principle if data and script are kept. Versions of R and packages used to run a particular script can be documented, and all versions of R and contributed packages are archived on CRAN. This implies that while the software might move on, the software needed to rerun older scripts is, and will remain, available to the research community.

Package Sweave [*Leisch*, 2002] provides the possibility to automate the integration of generated results and text into final documents (e.g., pdf), from documents that mix text (LaTeX) and R analysis script sections. Complete books like that of *Bivand et al.* [2008] are written in Sweave. This allows one to guarantee that nothing gets mixed up when integrating results (numbers, tables, graphs) into a report or paper and that procedures remain accessible and reproducible.

One might argue that should many researchers use R, programing errors would affect equally large numbers of users. However, when many people use a particular piece of software, programing errors should be found earlier, leading to fewer people being affected by the resolved errors. When software is open source, software errors can be found and quickly corrected. Finally, being able to identify errors in hindsight as software errors is an important aspect of reproducible research. The position of the R Foundation on software development quality control is well documented [*R Foundation for Statistical Computing*, 2008].

References

Bivand, R. S., E. J. Pebesma, and V. Gomez-Rubio (2008), *Applied Spatial Data Analysis With R*, 378 pp., Springer, New York. [Available at http://www.asdar-book.org/.]

Leisch, F. (2002), Sweave: Dynamic generation of statistical reports using literate data analysis, in *Compstat 2002: Proceedings in Computational Statistics*, edited by W. Härdle and B. Rönz, pp. 575–580, Phys. Verlag, Heidelberg, Germany.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna. [Available at http://www.r-project.org/.]

R Foundation for Statistical Computing (2008), R: Regulatory compliance and validation issues—A guidance document for the use of R in regulated clinical trial environments, position paper, Vienna. [Available at http://www.r-project.org/doc/R-FDA.pdf.]

—EDZER PEBESMA, Institute for Geoinformatics, University of Münster, Münster, Germany; E-mail: edzer.pebesma@uni-muenster.de; DANIEL NÜST, 52°North Initiative for Geospatial Open Source Software GmbH, Münster, Germany; and ROGER BIVAND, Department of Economics, Norwegian School of Economics, Bergen, Norway