R Coding and Modeling

BEN MARWICK
University of Washington, USA

Introduction

R is a programming language and computing environment useful for statistical and geospatial analysis, data visualization, and mapping. It is significant as the most widely used scientific programming language in archaeology. This entry will briefly describe the origins of R and survey its distinctive features, including how it enables reproducible research. It will also highlight some current uses of R in archaeology, and suggest some possible future directions.

R was originally developed in New Zealand in the early 1990s as an academic research project to create a language for introductory data analysis courses. It is closely related to the S language and S-PLUS software. Currently it is maintained by an international community of about 20 computer scientists and statisticians, and has six-monthly update cycles to the core libraries. It is widely used in academia, especially statistics and social sciences, and in industry. R has a vibrant online community that is friendly to novices, as well as numerous informal user groups, an annual conference (useR!), a non-profit organization (R Foundation), and a peer-reviewed journal (*The R Journal*).

Distinctive features of R

R differs from other software commonly used for data analysis in two important ways. First, R is an open-source software program. This means that, unlike commercial data analysis software (e.g., SAS, Stata, SPSS, JMP, SigmaStat, etc.), R is free for anyone to download and install (cf. PAST, JASP, and PSPP). The full details of all its algorithms are publicly available for inspection and modification (cf. JASP and PSPP), unlike for many other programs where these details are not available to the user (cf. PAST, and all commercial software). Second, for many archaeologists a spreadsheet program such as Microsoft Excel is their primary tool of data analysis and visualization. The primary mode of interaction is by manipulating cells in the spreadsheet, and through mouse-clicks to access commands in the drop-down menus. R differs from all other statistical software because it is not a spreadsheet program and has very few mouse-driven actions. Instead, R has a prompt to which typed commands are interactively sent to the R interpreter.

As a programming language, R gives the user great flexibility through access to a vast variety of methods. The user is not limited to built-in functions, but can easily create new ones. The R interpreter evaluates commands typed by the user, and the computed output is printed to the screen or stored for later use. These typed commands are saved in a plain-text R script file, which becomes a record of all the steps in a data analysis.

Unlike most other software, R also has an environment. Any type of data file (e.g., Excel, CSV, OBJ, TIFF, JPEG, shapefiles, netcdf, etc.) can be imported into the R environment and manipulated, analyzed and combined with other data using a coherent and integrated set of operations.

After a data file is imported into R, it is usually represented as a vector (a one-dimensional set of values of the same type, i.e., numeric, character, logical, etc.) or combinations of vectors. A frequently-used data structure in R is the data frame. This is analogous to a table of data in a spreadsheet, and formally in R is a list of different types of vectors of equal length. Many functions (named sections of code that perform a specific task) in R are optimized for working with data frames, and can perform summary statistics and complex operations very quickly on these structures. Related structures include the matrix (vectors of the same length but only one type, equivalent to a table of only numbers or only characters, cf. the standard structure for MATLAB software), and the list (a vector that

The Encyclopedia of Archaeological Sciences. Edited by Sandra L. López Varela.

DOI: 10.1002/9781119188230.saseas0631

combines elements of any data type). Lists are complex, hierarchical objects (a list can contain lists) that are frequently used during looping operations to automate complex actions efficiently for many items (e.g., hundreds of Excel files need to be read and computed on).

R is primarily a functional programming language, which means that the basic unit of activity when using R is typically the evaluation of mathematical or computational functions. The basic installation of R contains over a thousand functions. In addition to these built-in functions, an R user can install over 10,000 packages of more specialized functions freely contributed by R users. These packages are distributed via three major online code repositories: Comprehensive R Archive Network (CRAN), Bioconductor, and GitHub. Some examples of add-on packages specifically for archaeologists include archdata (contains 11 archaeological datasets from around the world reported in published studies), FedData (automates downloading of geospatial data from several US and international federated data sources), zooaRch (analytical tools to make inferences on zooarchaeological data), and recexcavAAR (functions for 3D reconstruction and analysis of archaeological excavations, including methods to reconstruct natural and artificial surfaces based on field measurements).

In recent years a collection of R packages known as the tidyverse has emerged as a major new direction for the R language (Wickham and Grolemund 2016). These packages share common philosophies of tidy data, which is an approach to representing and manipulating data where variables are in columns, observations are in rows, and values are in cells. Tidyverse packages are designed to work together, with common data representations and simple, consistent programming interfaces for fast, efficient data analysis and visualization. For an archaeologist learning R for the first time, the tidyverse approach and packages are an ideal starting point.

Reproducible research using R

In many areas of social science there is an emerging concern to improve the reproducibility of published results. Reproducibility is the cornerstone of science, and often defined as the ability to recompute the results of a data analysis, given a dataset and information about the data analysis pipeline (Marwick 2017). One contributor to this crisis is the use of mouse-driven software such as SPSS, where many steps of the analysis pipeline are not preserved in a convenient form, or at all. This makes it difficult to reproduce results obtained by pointing-and-clicking. Using R is one aspect of solving the problem of irreproducible results. When data are analyzed using a programming language such as R, the saved R script file contains a full record of every decision made during data analysis, and can be shared with other researchers.

In addition to scripting every step of the data analysis, R also integrates with several other technologies to enable reproducible research. R code can be written among narrative text using the R Markdown text formatting system. Markdown is a simple, easy-to-learn, open-source language for formatting documents. This means that entire journal articles and theses can be written in R Markdown, with code and text interwoven together in the same document (Figure 1). Using the Pandoc program (included with RStudio), the R Markdown document can be executed, or rendered into a standard format such as PDF or Microsoft Word, so that the R code is run and the figures and tables for the manuscript are generated and inserted in place in the output documents.

The significance of this use of R in manuscript preparation workflow is that it avoids the typical practice of copying and pasting output from a statistical software program (e.g., SPSS) into a word-processing program (e.g., Word). Copying and pasting separates a result from the steps that generated it, making it impossible to trace the decisions made during an analysis. Copying and pasting is also error-prone because mistakes or ad hoc, undocumented modifications to output can be introduced to the text during revision cycles. R Markdown provides a simple convenient workflow for writing code and text to easily reproduce the results of a data analysis.

Another advantage of using R and R Markdown is that the data analysis pipeline behind an article or thesis can easily be shared with other archaeologists by sharing the R Markdown documents that generated the output document. This

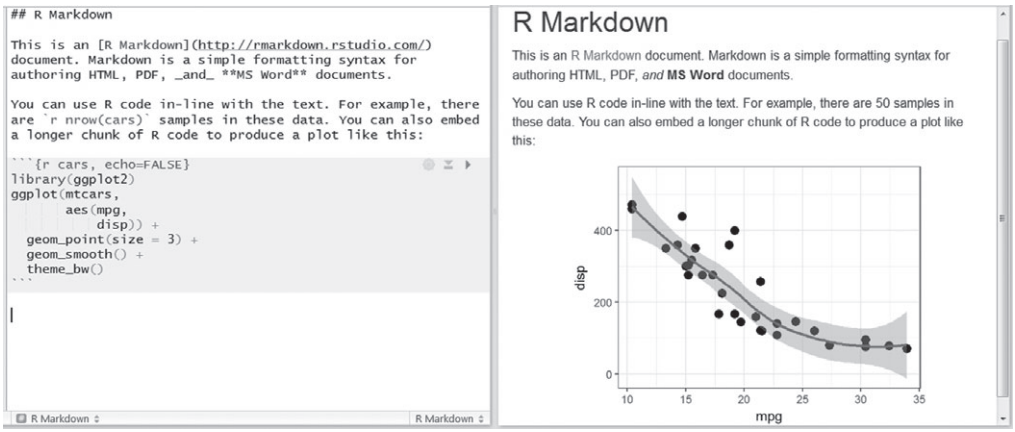

Figure 1 A screenshot from the RStudio program showing how R can be used for reproducible research. In the left panel is a text editor, where we write plain text and code in an R Markdown file (known as a Rmd file). In the right panel is the output that is produced when the Rmd file is "knit," or rendered, into a document. In this example, the Rmd has been knitted to produce a HTML file, but we could also produce a PDF or Microsoft Word document from the same Rmd file. The first paragraph of the text in the example demonstrates how to use markdown for basic text formatting (e.g., a heading, a URL, bold and italic text). The second paragraph shows how R code can be embedded inline in the text. The rendering process automatically runs the code and inserts the result in the text; here, it computes the number of rows in the "cars" dataset and inserts the result (50) in the rendered document. The text in the gray region on the left is a chunk of R code that produces the plot in the HTML file on the right. We use echo=FALSE in the code chunk to specify that the code chunk is not displayed in the HTML file; we see only the plot that the code generates. This method of writing text and code in the same document enhances reproducibility because the methods of data analysis (i.e., the R code) are explicitly included in the same document as the text, and the code can be easily and repeatedly run to generate results. This removes the need to copy and paste tables and plots from other software into the text, eliminating transcription errors and confusion about where a particular result came from.

is important for scientific archaeology because it enables more efficient adoption of new methods, and more rigorous checking of new results to establish their correctness. Furthermore, an archaeologist using R who does not make their code openly accessible upon publication is missing an important opportunity to increase the visibility and impact of their research.

A related tool that is often used with R to improve the transparency of data analysis is Git, the free and open-source version control system. Using Git with R means that researchers can track changes they make to their code and text files. Each decision point can be recorded, even if is later deleted from the final document. Git allows collaborators to compare versions of code and text, retrace errors, explore new approaches in a structured manner, while maintaining a full audit trail. Several free web services have user-friendly tools for collaborating in the open or privately with code and documents written with R and Markdown that are tracked using Git (e.g., GitHub, GitLab, BitBucket). The rrtools package at https://github.com/benmarwick/rrtools (Marwick, Boettiger, and Mullen 2017) includes instructions, templates, and functions for making a basic compendium suitable for doing reproducible research with R, Git and other related tools.

Current uses of R in archaeology

R's flexibility as an analytical tool means that it is currently used by archaeologists for a wide variety of applications. An online "task view" page at https://github.com/benmarwick/ctvarchaeology contains an annotated list of R

packages relevant to archaeological research. Specialist studies of typical archaeological items such as stone artifacts, animal bones, isotopes, and sediments have all been done to publication quality entirely with R. R has also emerged as the tool of choice for several computationally intensive subspecializations in archaeology.

Archaeological modeling in R

Archaeological modeling in R means representing archaeological data for a purpose. This means that the data stand for something in the real world, such as past human behaviors or site formation processes. The purpose is often to gain new insights into the relationships between observed variables and their possible behavioral and environmental correlates. Statistical modeling in R takes archaeological data and combines it with model formulae, equations, uncertainty and randomness to test hypotheses about how the system being modeled actually works. For any kind of modeling, from linear regression to exotic machine learning method, R is ideal for identifying patterns in data, classification, untangling multiple influences, and assessing the strength of the evidence. In R, a statistical model is often expressed with formula notation, for example $y \sim x$. This can be read as "y is modeled as a function of x," where y is a response variable whose content we are trying to model with x, called the explanatory variable. Following this notation, a typical linear regression model in R will look similar to this fit <− lm(y ∼ x), which is the R code for the traditional mathematical representation of $y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$.

This formula notation is the starting point for most models in R, including ones where explanatory variables are continuous, categorical, or a mixture of both, and when the response variable is a continuous measurement, a count, a proportion, or a category. The types of variables involved determine the type of model used. Typically where explanatory variables are continuous, regression models are used; when they are categorical, analysis of variance (Anova) is used; and when they are both continuous and categorical, analysis of covariance (Ancova) is used. For the response variable, when it is continuous we can use normal regression, Anova, and Ancova; when it is a proportion (i.e., continuous but within the range of 0–1) or binary (i.e., 0 or 1) we use logistic regression; and when it is a count we use log-linear models. For each of these variations, only minor modifications to the fundamental $y \sim x$ R code are necessary.

Working with radiocarbon dates is a common task for many archaeologists, and although several specialized software tools exist solely for this purpose (e.g., OxCal, BCal), R has established itself as part of the toolkit for innovative analyses of radiocarbon dates. New methods of analyzing sets of radiocarbon dates to model population dynamics have been developed by several R-using archaeologists. Another area where R has become established as the dominant tool is in studies of cultural evolution. In this context, R has been used to model transmission of cultural information, and test hypotheses about population structure and dynamics. Similarly, spatial analyses from hemispheric scales to intra-site scales often use R for modeling archaeological and climate data with a location component. R is also used for zooarchaeology, geoarchaeology, agent-based modeling, and predictive site location modeling in academic and commercial archaeological settings. These more specialized applications are enabled by R packages contributed by researchers and programmers in other disciplines, such as Quaternary science, ecology and evolution, and geography.

A unique characteristic of using R for specialized archaeological applications is that it provides a complete toolkit, not just for the exotic methods, but also for generic activities surrounding the core analysis, such as importing, cleaning, arranging and exploring raw data, testing hypotheses, and visualizing the output with maps and plots. Using R at every step of the research process improves efficiency because it minimizes context switching and manual handling, which may also accidentally introduce errors and confusion.

Future directions

For archaeologists, R has emerged as a universal toolbox with which any kind of data analysis is possible. In addition to the more than 10,000

packages of R functions contributed by users, we can freely write new functions and packages to implement new methods. However, perhaps the most valuable and unique attribute of R is that when writing R code to analyze data, we create an enduring record of all the decisions made during a research project. This record, that is, the R code files, can be archived at a DOI-issuing online repository (e.g., tdar.org, osf.io, figshare.com, zenodo.org). This means we can easily find and reuse it in the future, and we can cite our R code in our publications (Eglen et al. 2017). With our R code freely downloadable from a trustworthy repository, readers of our publications can download and use the R code that generated the results reported in the paper. This is important for the advancement of archaeology because it allows for more complete review of published research, and so a more robust evaluation of its reliability. It also allows for more efficient and rapid transmission of new methods through the research community.

SEE ALSO: Archaeological Sciences; Archaeology; Artificial Intelligence; Bayesian Statistics; Chi-Square Analysis; Cluster Analysis; Compositional Data Analysis; Computer Applications in Archaeology; Correspondence Analysis; Databases in Archaeology; Dating in Archaeology; Descriptive Statistics; Distance and Data Transformation; Exploratory Data Analysis (EDA); Factor Analysis; Geometric Morphometrics; Human Skeletal Data Quantification; Hypothesis Testing; Inferential Statistics; Information Modeling; Linear Models; Lithics Data Quantification; Mathematics; Multivariate Analysis; Neural Networks; Parallel and Distributed Computing; Predictive Modeling; Principal Component Analysis; Radiocarbon Calibration and Age Estimation; Regression and Correlation Analysis; Seriation; Simulation Modeling; Spatial Analysis; Statistics in Archaeology; Supervised Pattern Recognition

REFERENCES

Eglen, Stephen J., Ben Marwick, Yaroslav O. Halchenko, Michael Hanke, Shoaib Sufi, Padraig Gleeson, R. Angus Silver, et al. 2017. "Toward Standard Practices for Sharing Computer Code and Programs in Neuroscience." *Nature Neuroscience* 20 (6): 770–73. DOI:10.1038/nn.4550.

Marwick, Ben. 2017. "Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation." *Journal of Archaeological Method and Theory* 24 (2): 424–50. DOI:10.1007/s10816-015-9272-9.

Marwick, B., C. Boettiger, and L. Mullen. 2017. "Packaging Data Analytical Work Reproducibly using R (and Friends)." *The American Statistician*. DOI:10.1080/00031305.2017.1375986.

Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science*. Sebastopol, CA: O'Reilly. http://r4ds.had.co.nz.

FURTHER READINGS

Fox, John, and Harvey Sanford Weisberg. 2010. *An R Companion to Applied Regression* (2nd ed.). Thousand Oaks, CA: SAGE.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st ed.). Cambridge: Cambridge University Press.

Harrell, Frank E. 2010. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Hastie, Trevor, and Robert Tibshirani. 1990. *Generalized Additive Models*. Wiley Online Library.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R* (1st ed.). New York: Springer.