esa

ECOSPHERE

Evaluating the popularity of R in ecology

JIANGSHAN LAI [1,2,†], CHRISTOPHER J. LORTIE [3,4], ROBERT A. MUENCHEN [5], JIAN YANG [6], AND KEPING MA [1]

[1]State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093 China
[2]University of Chinese Academy of Sciences, Beijing 100049 China
[3]Department of Biology, York University, Toronto, Ontario M3J1P3 Canada
[4]The National Center for Ecological Analysis and Synthesis, Santa Barbara, California 93101 USA
[5]OIT Research Computing Support, University of Tennessee, Knoxville, Tennessee 37996 USA
[6]Key State Laboratory for Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093 China

Citation: Lai, J., C. J. Lortie, R. A. Muenchen, J. Yang, and K. Ma. 2019. Evaluating the popularity of R in ecology. Ecosphere 10(1):e02567. 10.1002/ecs2.2567

Abstract. The programming language R is widely used in many fields. We explored the extent of reported R use in the field of ecology using the Web of Science and text mining. We analyzed the frequencies of R packages reported in more than 60,000 peer-reviewed articles published in 30 ecology journals during a 10-yr period ending in 2017. The number of studies reported using R as their primary tool in data analysis increased linearly from 11.4% in 2008 to 58.0% in 2017. The top 10 packages reported were lme4, vegan, nlme, ape, MuMIn, MASS, mgcv, ade4, multcomp, and car. The increasing popularity of R has most likely furthered open science in ecological research because it can improve reproducibility of analyses and captures workflows when scripts and codes are included and shared. These findings may not be entirely unique to R because there are other programming languages used by ecologists, but they do strongly suggest that given the relatively high frequency of reported use of R, it is a significant component of contemporary analytics in the field of ecology.

Key words: code; ecology journal; open science; packages; R; reproducibility; statistical programming; Web of Science.

Received 10 August 2018; revised 2 November 2018; accepted 5 December 2018. Corresponding Editor: Tanya Berger-Wolf.

† **E-mail:** lai@ibcas.ac.cn

INTRODUCTION

The rapidly increasing volume and variety of ecological data has been driven by many factors. Automated collection by sensors, remote sensing, bioinformatic sequencing, easier data transfer, the emphasis of journals on data reporting, data repositories for ecology, and a change in the culture of collaboration have contributed to a collective opportunity for the discipline (Hampton et al. 2017). Facing these big data, many ecologists likely need programming skills to check quality, manipulate, analyze, and visualize data. Most ecologists often need to write their own code as part of their research (Mislan et al. 2016). Many statistical programming tools are suitable for ecological data analysis including R, Python, MATLAB, and SAS to name the most common.

At present, R is among the world most powerful statistical languages, and it is generally very popular in science (Bollmann et al. 2017). It provides interactive data analysis tools as well as high-quality graphics for different fields in scientific research (Mair et al. 2015). It has a thriving global community of users, developers, and contributors, and over 13,000 contributed packages are available on CRAN (Comprehensive R Archive Network, https://cran.r-project.org/). Many of these packages were contributed by experts in various areas of applied statistics to enhance its functionality.

R is both an environment and programming language appropriate for ecological research because R packages include a broad range of methods employed in ecological analysis as well as numerous routines for data exploration. More than 100 packages are often used in ecological data analysis as highlighted in the CRAN Task View: Analysis of Ecological and Environmental Data (https://CRAN.R-project.org/view=Environmetrics). Packages included in other Task Views (e.g., spatial, multivariate, cluster, and phylogenetics) are also useful for ecological data analysis. It is likely that ecologists will continue to use R for research because of the immediate benefits and growing availability of relevant R packages for ecological data analysis. There are already numerous books, articles, and online resources available demonstrating both power and practical use of R for ecological research, and many offerings are open and freely available. Thus far, no empirical study has examined the frequency of R use and associated packages for the reported analyses of ecological data in publications.

Academic journals are one of the primary vehicles for ecologists to disseminate findings. The popularity of R reported in publications in ecological journals can thus potentially function as a proxy measure of general use within the discipline. In this study, over 60,000 peer-reviewed articles published in the 30 main ecology journals during the past 10 yr (2008–2017) were evaluated for reported use of R and R packages to identify key trends in general reported use and patterns in common packages.

Methods

We selected a total of 30 journals based on the rank of impact factor (IF > 3.0, ISI 2017, www.webofknowledge.com) using the Web of Science (WoS) Journal Citation Reports' (JCR) category "Ecology" (Appendix S1). Review journals and relatively new journals published after 2008 were excluded excepting *Methods in Ecology and Evolution* (first published on February 2010). This journal publishes new methods in ecology and evolution and seemed ideal to include because of this relatively more technical focus. Articles published from 2008 to 2017 within the selected journals were retrieved. Research articles were filtered by the option "document types: articles" in the Web of Science excluding other non-research articles (e.g., reviews or editorials) in these journals. Some articles use R as the means for data analysis in the main text but only cite R packages in the reference section. To avoid this potential omission, the Methods section of each article was also manually inspected in every publication. The number of articles explicitly using R or R packages was counted, and if a specific R package was listed, the identity was also recorded.

Non-metric multidimensional scaling analysis (NMDS) based on Bray–Curtis dissimilarity matrix among journals of usage records for package within each journal was used to test whether journals with similar scope tended to use a similar set of packages using the metaMDS function in the R vegan package (Oksanen et al. 2018). All calculations in this paper were conducted using R statistical language (R Development Core Team 2017).

Results

General patterns in reported R usage

Collectively, a total of 60,902 research articles were published in the 30 selected ecology journals from 2008 to 2017 (Appendix S1). Among them, 20,395 articles (33.5%) explicitly listed R as the statistical software for analyses. The percentage of articles reporting R use increased linearly across time from 11.4% in 2008 to 58.0% in 2017 (Fig. 1, Correlation $r = 0.99$, $P < 0.0001$), regardless of difference among journals (Appendices S2, S3). In the most recent two full years, half of the papers reported use of R (52.4% for 2016 and 58.0% for 2017) in these ecological publications.

All journals examined showed a consistent increase in the percentage of articles using R from 2008 to 2017 (Appendices S2, S3). The average percentage among all journals (Appendix S4) show *Methods in Ecology and Evolution* had the highest mean percent of reported R use (53%). This suggests that the majority of contemporary methods in ecology and evolution are supported by or done using the programming language R. A group of journals related to spatial ecology have particularly high percentages: *Global Ecology and Biogeography* (2nd position with 51.6%),

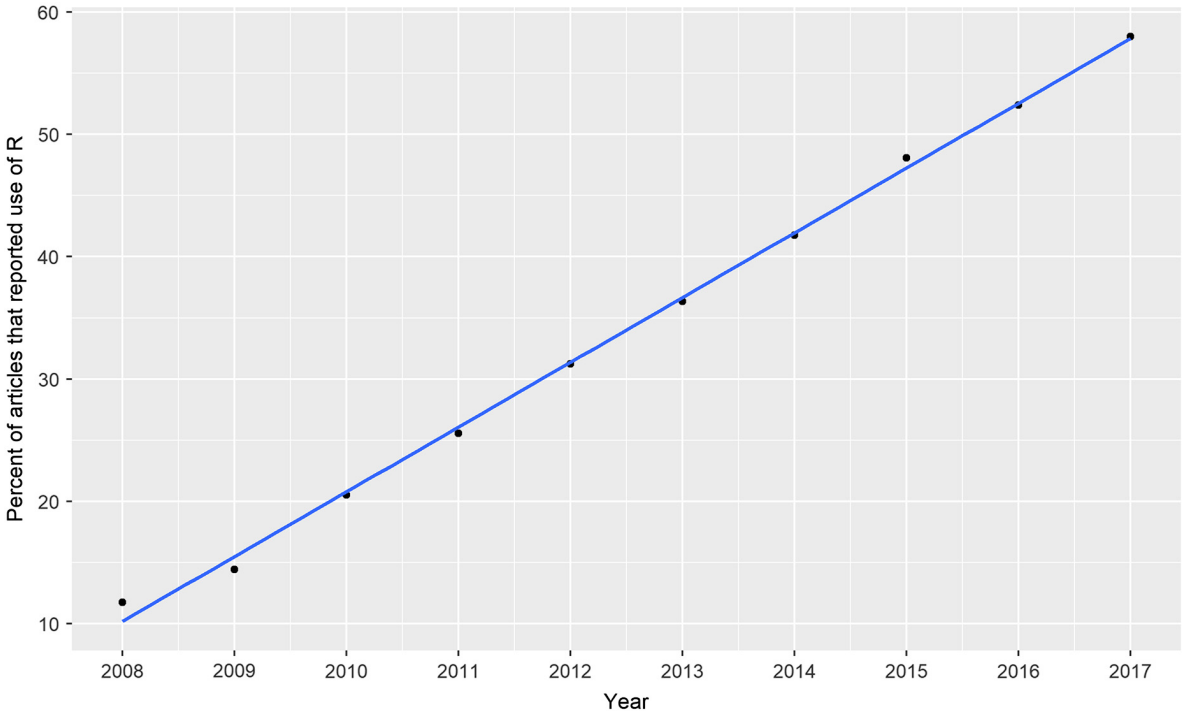

Fig. 1. The percentage of research articles using R in the 30 main ecology journals from 2008 to 2017. Data were from over 60,000 articles with methods inspected for reported R use.

Ecography (3rd position with 50.9%), and *Diversity and Distributions* (5th position with 46.3%; Appendix S4).

Journal of *Ecology* and *Journal* of *Animal Ecology*, two flagship journals for plant ecology and animal ecology published by the British Ecological Society, place 4th with 48.1% of all the articles using R and 6th with 45%, respectively. A group of ecology journals focused on synthesis also had relatively high percentages of R use including *Ecological Monographs* (7th, 43.5%), *Oikos* (8th, 39.3%), *Ecology* (9th, 39.1%), *Oecologia* (10th, 38%), *Journal* of *Applied Ecology* (11th, 37.3%), and *Ecology Letters* (12th, 37%; Appendix S4).

R package usage patterns

More than 2400 R packages were used in these articles, and 31 packages were used in more than 100 articles (Fig. 2). The package with highest frequency was lme4 which is used to fit and analyze linear mixed models. The runner-up was vegan a package widely used for multivariate analysis in community ecology. These two packages were far more popular than other packages (Fig. 2). The third was nlme which is used to fit linear and nonlinear mixed models. The fourth was ape a package for analyses of phylogenetics and evolution. The details of 10 most frequently used packages listed in Appendix S5.

Non-metric multidimensional scaling analysis suggests that journals with similar scope tend to use a similar set of packages (Fig. 3 and Appendix S6). Seventeen journals shared lme4 as their most frequently used packages and most locate on bottom part of NMDS plot (Fig. 3). Vegan was the most frequently used packages for 12 journals that distribute on upper part of NMDS plot (Fig. 3). *Evolution* was the only journal that neither lme4 nor vegan were the most frequently used packages, replaced by ape package (Fig. 3 and Appendix S6). Most journals in the first quadrant focus on macroecology (Fig. 3), for example, *Global Ecology and Biogeography*, *Ecography*, *Diversity and Distributions*, *Journal of Biogeography*, *Ecosystems*, and *Landscape Ecology*. A group of synthesizing ecology journals are close to the center; these include *Ecology Letters*, *Ecology*, *American Naturalist*, *Oikos*, *Oecologia*,

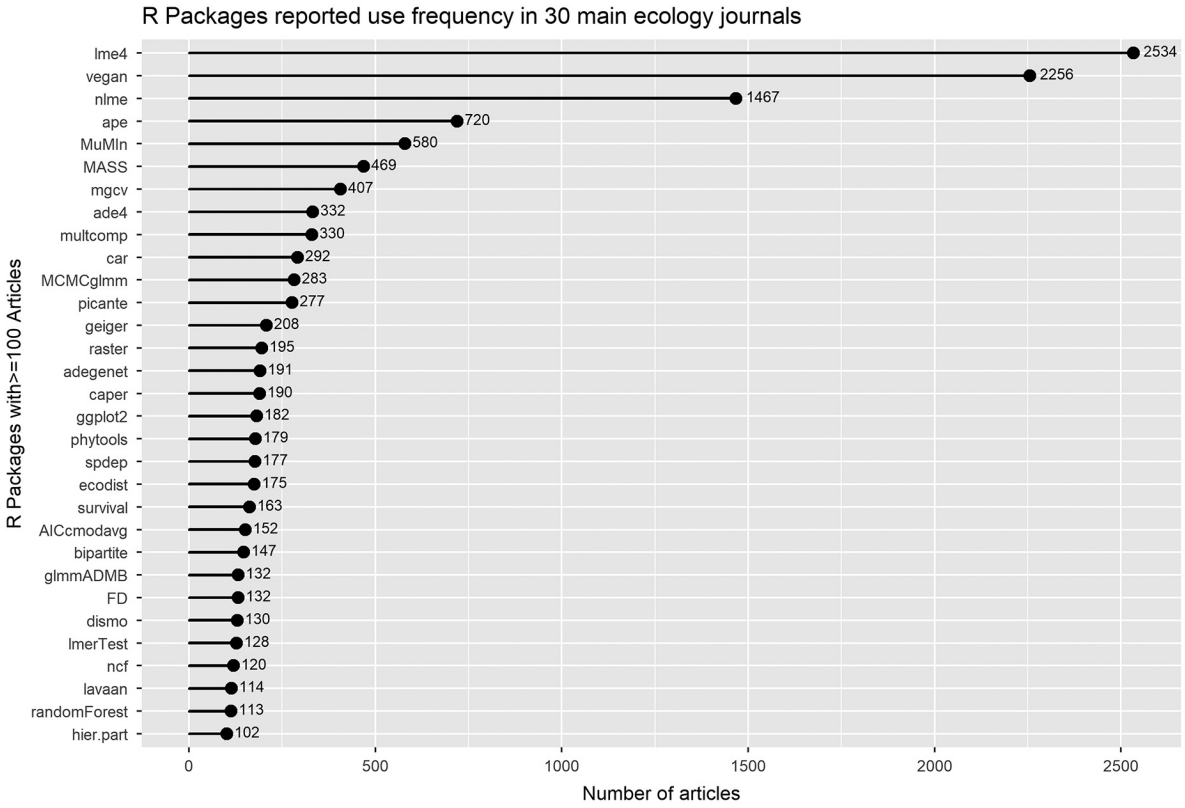

Fig. 2. The citations of the most popular R packages (>100 articles) in the 30 main ecology journals from 2008 to 2017. Citations were determined by inspecting the methods of over 60,000 research articles in these journals—not from Web of Science Citation index.

Journal of Ecology, Journal of Animal Ecology, Functional Ecology, and *P ROY SOC B-BIOL SCI* (Fig. 3). These journals all shared lme4 as their most frequently used packages. The *ISME Journal*, *Fungal Ecology,* and *Microbial Ecology* publish articles on the interaction between microorganisms and their environments, and vegan was the most frequently used packages among them. Lme4 was much less popular and falls behind the fifth position in the *ISME Journal* and *Microbial Ecology* (Appendix S6). The small sample size for the citation records for R packages in *Ecological Monographs* resulted in greater Bray–Curtis distance from the center as shown in the NMDS plot (Fig. 3), due to its limited volumes.

Discussion

Computation and programming are important components of modern ecology. There are many tools available to ecologists for data analysis and visualization—each with relative costs and benefits. Nonetheless, the percentage of research articles using R as the statistical tool within the 30 main ecology journals has increased from 11.4% (2008) to over 50% by 2016. This popularity can be the product of at least several of the following explanations.

First, R has an enormous amount of resources. Ecological data and questions are becoming increasingly sophisticated and as a result ecology has become an increasingly quantitative, statistical field (Higgs 2015, Touchon and McCoy 2016). Through CRAN and over 13,000 packages, R offers almost all existing statistical models for ecological analysis (R CRAN Task Views: Environmetrics; Phylogenetics; Spatial). There are also at least 3000 functions specific to ecological research (http://finzi.psych.upenn.edu/cgi-bin/namazu.cgi?query=ecology&max=100&result=

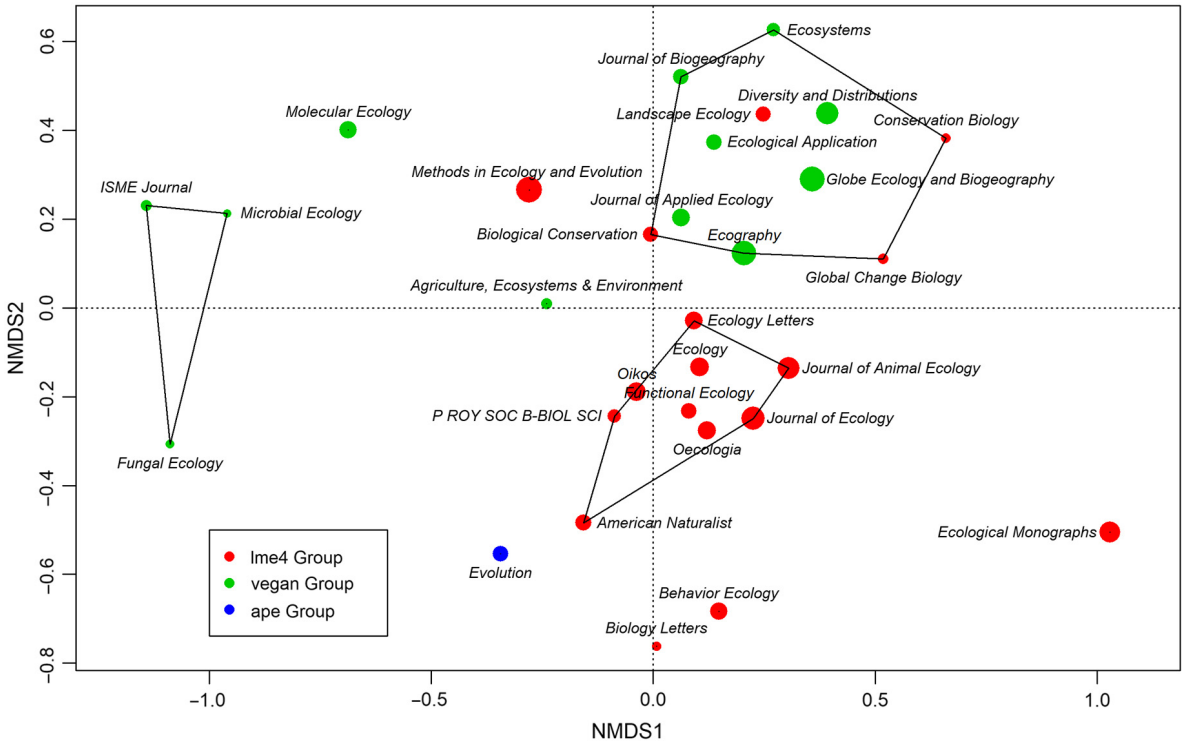

Fig. 3. Non-metric multidimensional scaling plot of a Bray–Curtis dissimilarity matrix of usage records for package within 30 main ecology journals. The diameter of each point for at each journal is proportional to the mean percentage of articles listing R within publications from that journal between 2008 and 2017.

normal&sort=score&idxname=functions&idxname=views). R also offers significant flexibility in data visualization and can produce high-quality graphics for academic publications. Many statisticians and quantitative ecologists use R as their working language so many new statistical methods are first developed in R (Legendre and Legendre 2012, Bivand et al. 2013, Swenson 2014, Borcard et al. 2018). This is shown by the fact that that the highest percentage of R usage occurs in *Methods in Ecology and Evolution*, a journal that publishes new methods in various fields of ecology and evolution.

Second, R is free and open source. Source code for all the functions is freely available. Sharing R code also became a culture for ecologists (i.e., ROpenSci). While similar sharing is possible with code-based tools such as SAS or MATLAB, the great expense of those packages limits access. Because our search was constrained to only consider the R use, we were unable to quantify the trends of other computer programs in ecology journals. However, across a broad range of academic journals, the use of R has been growing while the use of proprietary tools such as SAS and SPSS has declined from 2008 to 2016 (http://r4stats.com/articles/popularity/; Muenchen 2017). Another earlier study that focused on ecology also indicated that the usage of SAS and SPSS has been decreasing from 2006 to 2013, based on the survey of seven main ecology journals which were included in the current research (Touchon and McCoy 2016). It is also likely that ecologists are reducing their reliance on expensive commercial software (Tippmann 2015). Of Course, although R is open source, most R functions are written by C, C++, and FORTRAN for efficiency and speed; it is possible to examine the C, C++, and FORTRAN source code of packages, but few Windows users ever see that code.

Finally, the online R community (e.g., R-sig emailing system, Stack Overflow, Blogs) facilitates the exchange of ideas, problem-solving, and code sharing for specific challenges. Dryad,

GitHub, and Zenodo have made code sharing and archiving possible. Its active user community and rich set of resources make R a convenient and quick tool for data analysis by getting relevant code quickly through various search engines, such as Google.

Although the usage frequency of an R package depends on many factors, for example, release date, the number of functions, and its scope, the popularity of R packages still reflects the main trends of statistical methods for ecological data analysis. Ecological models often need to consider random effects to quantify the variation among units, space, and time (Bolker et al. 2009). Therefore, it is understandable that two packages for fitting mixed effect models: lme4 and nlme are first and third most frequently used in ecological articles (Fig. 2). The vegan package is powerful, and it contains a large number of functions for multivariate analysis and biodiversity as well as other functions based on dissimilarity matrices (e.g., Mantel test and principal coordinate analysis) are useful for molecular biology and microbial ecology (Excoffier et al. 1992, Dixon 2003, Phipson and Smyth 2010, Diniz et al. 2013). So journals focusing on macro-scale ecology (*Global Ecology and Biogeography, Ecography, Diversity and Distributions, Ecosystems,* and *Agriculture, Ecosystems & Environment*) and micro-scale ecology (*Molecular Ecology, ISME Journal, Fungal Ecology,* and *Microbial Ecology*) share vegan as the most frequently used package, as well as in applied ecology (*Journal of Applied Ecology* and *Ecological Application*). During the last decade, there was a burst in the number of studies incorporating phylogenetic information in community ecology (Swenson 2014, Cadotte et al. 2017), and the ape package became the fourth most popular package among ecologists.

In ecology, reproducibility is not always adequate perhaps in part due to difficulties in open access to primary data and records for analysis (Hampton et al. 2015, Roche et al. 2015, Lortie 2017, Lowndes et al. 2017) and in part due to challenges in replicating the methods used to handle diverse ecological data. R is an ideal mechanism to break through these limitations because it can be used to describe workflows and decisions in treating and modeling data in ecology. Increasingly, ecology journals request that authors make their code publicly available with data (Barnes 2010, Hampton et al. 2015, Nosek et al. 2015, Mislan et al. 2016). This decision will enhance the capacity to review and replicated the scientific process in ecology because it illuminates some of the major components associated with the workflows that support interpretation and discussion of evidence described in publications.

The number of R packages for ecological data analysis has grown rapidly in the past years, allowing ecologists to process and visualize data. The R language has accelerated the growth and development of ecology by connecting us with data scientists and computational thinking (Visser et al. 2015) and by opening up new biostatistical and modeling tools that were previously largely inaccessible (Simpson 2018). R also promotes communication between ecologists and statisticians (Andy 2015, Higgs 2015). Almost surely, the future will see additional exciting new developments and possibilities for R to benefit ecology, and hopefully for ecologists, in turn, to advance R and open computation and analyses.

Acknowledgments

We are grateful to hundreds of graduate students who participated in "R course" taught by Dr. Jiangshan Lai at University of Chinese Academy of Sciences and Graduate School of Chinese Academy of Agricultural Sciences for their work to survey literature which provided the empirical basis for this research. We also thank Jinlong Zhang, Xiao Xiao, and Hongjun Wang for their comments on the manuscript. The research was supported by grants from National Key Research and Development Project of China (2016YFC0500202) and National Natural Science Foundation of China (41471044). CJL was supported by NSERC DG and NCEAS. LJS collected the data. JSL and CJL analyzed the data. All authors wrote the manuscript.

Literature Cited

Andy, H. 2015. The new statistics with R: an introduction for biologists. Oxford University Press, Oxford, UK.

Barnes, N. 2010. Publish your computer code: It is good enough. Nature 467:753.

Bivand, R. S., E. Pebesma, and V. Gómez-Rubio. 2013. Applied spatial data analysis with R. Second edition. Springer, New York, New York, USA.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. S. White.

2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127–135.

Bollmann, S., D. Cook, J. Dumas, J. Fox, J. Josse, O. Keyes, C. Strobl, H. Turner, and R. Debelak. 2017. A first survey on the diversity of the R community. R Journal 9:541–552.

Borcard, D., F. Gillet, and P. Legendre. 2018. Numerical ecology with R. Second edition. Springer, New York, New York, USA.

Cadotte, M. W., T. J. Davies, and P. R. Peres-Neto. 2017. Why phylogenies do not always predict ecological differences. Ecological Monographs 87:535–551.

Diniz, J. A. F., T. N. Soares, J. S. Lima, R. Dobrovolski, V. L. Landeiro, M. P. D. Telles, T. F. Rangel, and L. M. Bini. 2013. Mantel test in population genetics. Genetics and Molecular Biology 36:475–485.

Dixon, P. 2003. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science 14:927–930.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes – application to human mitochondrial-DNA restriction data. Genetics 131:479–491.

Hampton, S. E., et al. 2015. The Tao of open science for ecology. Ecosphere 6:120.

Hampton, S. E., et al. 2017. Skills and knowledge for Data-Intensive environmental research. BioScience 67:546–557.

Higgs, D. M. 2015. Ecology and statistics: A healthy union? BioScience 65:1021–1025.

Legendre, P., and L. Legendre. 2012. Numerical ecology. Third edition. Elsevier, Amsterdam, The Netherlands.

Lortie, C. J. 2017. Open sesame: R for data science is open science. Ideas in Ecology and Evolution 10:1–5.

Lowndes, J. S. S., B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O'Hara, N. Jiang, and B. S. Halpern. 2017. Our path to better science in less time using open data science tools. Nature Ecology & Evolution 1:160.

Mair, P., E. Hofmann, K. Gruber, R. Hatzinger, A. Zeileis, and K. Hornik. 2015. Motivation, values, and work design as drivers of participation in the R open source project for statistical computing. Proceedings of the National Academy of Sciences USA 112:14788–14792.

Mislan, K. A. S., J. M. Heer, and E. P. White. 2016. Elevating the status of code in ecology. Trends in Ecology & Evolution 31:4–7.

Muenchen, R. A. 2017. The popularity of data science software. http://r4stats.com/articles/popularity

Nosek, B. A., et al. 2015. Promoting an open research culture. Science 348:1422–1425.

Oksanen, J., et al. 2018. vegan: community ecology package. R package. version 2.5-2. https://CRAN.R-project.org/package=vegan

Phipson, B., and G. K. Smyth. 2010. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly Drawn. Statistical Applications in Genetics and Molecular Biology 9. https://doi.org/10.2202/1544-6115.1585

R Development Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roche, D. G., L. E. B. Kruuk, R. Lanfear, and S. A. Binning. 2015. Public data archiving in ecology and evolution: How well are we doing? Plos Biology 13:e1002295.

Simpson, G. L. 2018. Modelling palaeoecological time series using generalized additive models. bioRxiv. https://doi.org/10.1101/322248

Swenson, N. G. 2014. Functional and phylogenetic ecology in R. Springer, New York, New York, USA.

Tippmann, S. 2015. Programming tools: adventures with R. Nature 517:109–110.

Touchon, J. C., and M. W. McCoy. 2016. The mismatch between current statistical practice and doctoral training in ecology. Ecosphere 7:e01394.

Visser, M. D., S. M. McMahon, C. Merow, P. M. Dixon, S. Record, and E. Jongejans. 2015. Speeding up ecological and evolutionary computations in R; essentials of high performance computing for biologists. Plos Computational Biology 11:e1004140.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.2567/full