Ecography 39: 349–353, 2016
doi: 10.1111/ecog.02493

Towards a more reproducible ecology

Michael Krabbe Borregaard and Edmund M. Hart

M. K. Borregaard (mkborregaard@snm.ku.dk), Univ. of Copenhagen, Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, Copenhagen, Denmark. – E. M. Hart, Univ. of Vermont, Dept of Biology, Marsh Life Science Building, Burlington, VT, USA.

The workflow of ecological scientists is currently undergoing a quiet revolution. Recent years have witnessed a strong push towards a more open sharing of research data (Hampton et al. 2015), and the vast amount of field data being generated by individual ecologists is becoming available to the wider research community at an unprecedented rate. The wide availability of data has also pushed scientists to focus more intensively on the process of data analysis. Where computer programming ability was restricted to a very small subsets of researchers just a few years ago, the new generation of ecologists are trained programmers, developing novel software for analyses and exploring new ways to share and visualize data. This development is moving the field away from click-and-calculate (Graphical User Interface, GUI) statistical packages. A new paradigm has emerged, where individual scientists download, curate and share large amounts of data and analyse it using reproducible software packages and scripts written in languages such as R, Python and Julia.

This increased focus on analytical methods has led to a number of key developments in scientific sharing and publishing, one of which is the Software notes format here at *Ecography* that was first instigated in 2008 (Pettersson and Rahbek 2008). The purpose of Software notes was to create a platform for disseminating high-quality analytical tools for ecology, to increase scientific transparency by opening the possibility for researchers to subject their techniques to traditional rigorous peer-review, and finally to support a transition where authors would receive scientific credit for their intellectual contributions in developing the most widely used methods in ecology. In the following 8 years, this idea has set root, and today dedicated journals such as *Methods in Ecology and Evolution* and *Environmental Modelling and Software* have followed suit in publishing software and computational methods as independent intellectual contributions.

These developments in ecology have exciting implications – the availability of large amounts of data and the explosion in the analytical capabilities of ecologists, together with the potential for rapid dissemination of ideas in today's internet-based scientific community, means that ecology is moving forward rapidly, with a steep growth in the number of research papers. But this development also poses some important challenges. The large amount of project-specific software being generated for analytical studies means that analytical standards are harder to establish, potentially limiting the reproducibility of much of recently published science. Also, analytical and coding errors may escape detection, with potentially highly problematic results, such as when Geoffrey Chang and colleagues had to retract three Science papers after discovering an error in a homemade data-analysis program (Miller 2006). Substantial progress in our understanding of ecology rests on trustworthy, reproducible and transparent data analysis.

Reproducibility is the very hallmark of the scientific method. However, there is an increasing concern that many studies today might not be reproducible. The focus on novelty in 'high-impact' journals means that there is little incentive for researchers to directly replicate published studies, and this lack of replication of studies has come under increasing scrutiny (Iqbal et al. 2016, The Economist 2016). What is more worrying, efforts to systematically replicate published studies have often failed (Open Science Collaboration 2015). This is so concerning that it has lead to the founding of a new journal dedicated explicitly to the replication of published results (the *Preclinical Reproducibility and Robustness* channel of *F1000* opened 4 February 2016). A recent comment in Nature magazine (Allison et al. 2016) reported on the widespread problem of reproducibility in the natural sciences, and gave a sobering account of the obstacles to overcoming it, which includes the pressure on researchers to publish, the lack of established pathways for dealing with non-reproducible articles, and consistent issues with the statistical treatment of data (Krumholz 2016, The Economist 2016). Ecology faces particular challenges in reproducibility because data collection is often context dependent (Ellison 2010), and because there are few established standards for storing metadata and facilitating study replication.

The keys to a greater level of reproducibility in ecology are to establish analytical protocols that are robust and transparent, to faithfully document the analytical process including any failed attempts, and to ensure that the storage and acquisition of data is documented and includes the appropriate metadata. Fortunately, recent technological developments

promise to increase the reproducibility of ecological analyses, by establishing documentable and standardised workflows, where the process of data acquisition, analysis and graphical output is integrated and documented throughout, and collaborative work is integrated into the software itself. Such developments have thus far received some attention, mainly among younger scientists, and largely outside of the primary literature. The special issue 'Tools for Reproducibility in Ecology' seeks to promote the quest for a reproducible ecological science and highlight recent developments, while presenting a collection of software notes that aim to explicitly further scientific reproducibility in ecological data analysis.

The tools included in the current special issue are:

mangal – a standard for sharing network data, including a web service for accessing it and an R package front-end (Poisot et al. 2016a).

ENM – a tool for species distribution models with explicit workflow and structures for sharing and documentation of analytical methods (De Giovanni et al. 2016).

geoknife – a tool for acquiring geographical data from large data bases (Read et al. 2016).

macroeco – a Python environment for macroecological analysis, with a scripting GUI (Kitzes and Wilber 2016).

sdm – an extensible tool for species distribution modelling that provides a standardized and unified structure for handling species distribution data and for modelling distributions with correlative and mechanistic approaches (Naimi and Araújo 2016).

Biogeo – a tool for programmatically detecting and correcting errors in widely used species–occurrence databases (Robertson et al. 2016).

helminthR – a tool for downloading data on host–parasite interactions from online sources (Dallas 2016).

Included is also a guest editorial by Poisot et al. (2016b) that highlights workflows and methods for working with datasets synthesized across several sources.

These tools exemplify different aspects of a data analysis workflow with the potential to improve reproducibility of ecological research (Table 1). Such a workflow involves presenting and documenting standards for data and metadata storage and communication, documenting the process of data acquisition, relegating analytical steps to online facilities with well-documented protocols, documenting analytical work on GUI platforms, establishing clear analytical workflow protocols, and emphasizing unit-testing and quality control of analytical software. In the following, we describe how each of these approaches can play a role in ensuring scientific reproducibility.

Metadata and data standards

Not long ago, ecologists wanting to describe the natural world collected the data themselves. Some data would eventually be published, but much would remain in notebooks or, more recently, left in computer hard drives to eventually disappear at the retirement of the researcher, leading to an inevitable decay in data availability (Vines et al. 2014). Today's online platform allows data to be used for answering questions beyond the purpose they were collected for, and thus they become a shared resource for the global research community. Consequently, there is a push towards seeing data as a scientific product in itself, and there is ongoing work to develop a system that supplies the generation of data with suitable attribution (Mooney and Newton 2012, Data Citation Synthesis Group 2014).

The development towards a larger degree of data re-use and sharing has the potential to speed the pace of scientific discovery, but is not without problems, as reported by Poisot et al. (2016b) in the present special issue. The authors describe an approach to working with large-scale synthetic data sets, and discuss many of the pitfalls. One powerful tool to deal with such pitfalls is to agree on reproducible and standardised sampling methods across systems and localities (Nogués-Bravo et al. 2011); but even in the absence of standardized sampling a significant improvement can be gained by agreeing on standards for saving and sharing data.

Such a standard is described for network analyses in the software note describing *mangal* (Poisot et al. 2016a). The standard is explicitly formulated to make data acquisition, and as importantly, data deposition, as simple and straightforward as possible, while at the same time encouraging the deposition of as much useful metadata as possible. The *mangal* format is aimed at efficient parsing by machines, and comes with an associated web service and R package for easy download and deposition. A similar but smaller effort is made by the *helminthR* package (Dallas 2016), which presents a direct programmatic interface to the Natural History Museum of London's database of host–parasite interactions of helminth worms.

It is worth noting that the creation and adoption of data standards can be a long and arduous process (Edwards et al. 2011). Given that publications remain the primary currency of a scientific career, the painstaking curation of data,

Table 1. The software notes in the thematic issues and the elements of ecological reproducibility that they contribute to.

	Standards	Data acquisition	Documentation	GUI scripting	Exchangeable objects	Niche models	Calls online functionality	Explicit workflow	Workflow system
rmangal	x		x				x	x	R
ENM			x		x	x	x	x	Taverna/R
geoknife		x	x				x	x	R
macroeco			x	x					Python
sdm			x	x	x	x		x	R/shiny
helminthR	x	x	x				x		R
biogeo		x	x					x	R

including preparation of detailed metadata and establishment of standards, can be viewed as an unrewarded burden, especially for early career scientists. This perception might be gradually changing as journals begin to require data deposition with publication (Bloom 2014, Sandhu 2014) and allow data to become a stand-alone publication in journals such as *Nature*'s journal *Scientific Data*. Tools such an *mangal* (Poisot et al. 2016a) facilitate the usage of archives and the contribution of data in a standard, reducing the burden of proper data curation. The simultaneous increase in incentives from journal editors and the creation of software tools that facilitate standards go a long way towards increasing reproducibility.

Online computation and data acquisition

In addition to allowing for data re-use and synthetic data sets, well-developed standards play a crucial role in allowing for replicability of published studies: Any researcher should be able to reproduce all presented results in a network study using *mangal* by downloading the data directly through the package and following the analytical steps described in the methods section of the paper. The same approach to data accessibility applies to other types of data. Not all of the data used by ecologists can be characterized as observational ecological data: researchers in ecology use data on climate, geology, topography and other abiotic factors shaping the environment. Ecological research is increasingly reliant on data products from remote sensing, thus drawing on what can be rightfully called *big data* (Hampton et al. 2013), in which the acquisition, curation, and preprocessing of data products are an integral part of ecological analyses. While these data products are an important component of macroecological workflows, they often are only available at spatial scales much larger than a researcher might want (e.g. the Oregon PRISM project data) and require post-download preprocessing. However this post-download preprocessing can inhibit reproducibility on two fronts: First it may be 'ad-hoc' and not well documented, and second it may require a computational power that is not available for most users.

The *geoknife* R package (Read et al. 2016) offers a way to ensure reproducibility in these data acquisition and preprocessing steps, by delegating the analytical steps to online data providers, which generally implement well-documented and transparent procedures. *geoknife* offers a protocol to derive summary data (such as the monthly standard deviation of a high-resolution data set on temperatures) within an area exactly defined by the study area. The protocol makes the process of data acquisition easy for the ecologist, reduces the computational demands on local computer systems, reduces error rate in the data preprocessing step, and allows for easy reproducibility by allowing researchers to report a few simple lines of code that generate the input data in models.

Workflow tools and GUIs

A more encompassing approach to ensuring reproducible workflows is made possible by dedicated workflow systems like *Kepler* (Altintas et al 2004) or the open-source project *Taverna* (Wolstencroft et al. 2013). Such systems present graphical platforms for calling web services, accessing online data, and running established analytical steps on the data. The approach makes for very clear and highly reproducible science, and encourages the use of established protocols for analysis whenever they are available and feasible. *Taverna* was originally developed for molecular biology, but its use is not restricted to this field. In this issue, De Giovanni et al. (2016) provide *Taverna* components, scriptable in R, for environmental niche modelling (ENM), also known as bioclimatic envelope modelling or species distribution modelling (SDM)(Peterson et al. 2011). Whether the standardized protocols that are made feasible by such workflow tools will ever dominate the analytical toolbox of individual researchers is an open question. However, there is no doubt that dedicated workflow tools offer a very powerful platform for collaboration among larger groups or big field-based projects, a type of research organization that is in itself on the rise in ecology.

A slightly less encompassing approach based on scripting may be more attractive to individual researchers. In terms of supporting reproducibility, the increasing prevalence of scripting languages such as R and Python provides substantial improvements over point-and-click software packages, in that they allow analyses to be replicated exactly by re-running a script, which represents a more stringent representation of analytical choices than is possible within the methods section of a short-format research paper. However, scripts are not always shared along with the paper, they are often poorly annotated, they require special knowledge to read and are often difficult to read even for those who have that special knowledge. Also, scripting languages limit reproducible analyses to users with programming skills. However, programming skills are not common among large groups of ecologists, especially the important sector of ecologists based outside universities.

In this special issue, Kitzes and Wilber (2016) provide one innovative way of making reproducible documentation of analysis available to non-programmers. Along with *macroeco*, a package of macroecological tools programmed for the widespread and powerful Python language, they provide a windows-based GUI platform allowing analyses to be easily specified, and subsequently run by the underlying software. The GUI saves a small script file that exactly specifies the analyses performed and is succinct enough to include directly within the methods section of a paper. By linking the reporting of the data analysis so closely between the article text and the analytical process itself, the approach represents a powerful way of ensuring reproducibility.

Taking the graphical and accessible approach even further, Naimi and Araújo (2016) provide the *sdm* package for species distribution models that offers a fully fledged GUI interface, where models and analytical choices can be specified in a well-known and user-friendly format. The GUI, which is based on R's *shiny* package, converts the specified analytical settings directly into a standardized R script that can be shared along with the paper. In addition, the GUI offers the opportunity to save analytical settings as a binary data object, which can shared among collaborators and modified, ensuring that analyses can be reproduced directly from within the GUI. The package also enhances analytical reproducibility in two other ways that were discussed earlier: It allows data

preprocessing to be handled within the package itself, ensuring that it is standardized and reproducible; and it allows the application of practically the entire range of different techniques for SDM within the same framework, ensuring that differences between analytical results derive explicitly from the differences between methods, rather than ad-hoc assumptions made by different software.

Reliability of tools and data

A final, and crucial, aspect of reproducibility is to minimize the number of errors in published data. If studies cannot be replicated, it might be a sign that the analysis was flawed or that the reported results were untrustworthy, a situation that is highly detrimental to the quality and integrity of science. The few analyses that have been performed to quantify the prevalence of such errors (Simundic and Nikolac 2009, Gilbert et al. 2012, Open Science Collaboration 2015) indicate that they are more common in submitted and published papers than often anticipated. How serious these errors are for the advancement of science is still unknown, but they are potentially a very serious issue.

Relying on well-established analytical tools is one way to minimize the amount of errors, although computer programs, such as R packages, are rarely peer-reviewed and by no means exempt from errors. Journals publishing software-note formats can help reducing these errors by ensuring that published analytical software have a comprehensive test suite that ensures internal consistency and error catching within the software – e.g., the *macroeco* (Kitzes and Wilber 2016) package in this issue is covered by 135 internal unit tests and all are available in the package's *github* repository.

Errors of recording are another major complication, well-known among researchers extracting ecological data from field notebooks. The problem is greatly exacerbated by the reliance on large online databases based on such data, which pervade many modern ecological analyses. The thorny issue of errors finding their way into analyses has prompted the creation of automatic tools to detect and correct errors and inconsistencies, such as the widely used Taxonomic Name Resolution Service (Boyle et al. 2013), which resolves the identities of plant species based on taxonomic synonymy, and also corrects for spelling errors. Robertson et al. (2016) present the *biogeo* package, which offers facilities for correcting common errors and quality issues with occurrence records found in large data bases. Not only does the software highlight potential errors, it also provides probable suggestions for the correct entries and allows the user to correct them in an easy and reproducible manner. Likewise, *sdm* includes R functions to correct for spelling errors while coding and parameterizing species distribution models (Naimi and Araújo 2016).

The software notes in this special issue all contribute to a more reproducible ecology in which analyses rest on solid, error-checked software, without stymieing the free growth of creative analytical ideas; and where documentation and metadata support a solid foundation under today's fast-moving integrative ecological research field. The notes were chosen to highlight a breadth of topics and approaches that are required to ensure reproducibility. The sooner these considerations are integrated into our workflows and collaborations, the stronger the foundation of the ecology we build for the future.

References

Allison, D. B. et al. 2016. A tragedy of errors. – Nature 530: 27–29.

Altintas, I. et al. 2004. Kepler: an extensible system for design and execution of scientific workflows. – Sci. Stat. Database Manage. 2004, Proc. 16th Int. Conf., pp. 423–424.

Bloom, T. 2014. PLOS' new data policy: part 2. – <http://blogs.plos.org/everyone/2014/03/08/plos-new-data-policy-public-access-data/>.

Boyle, B. et al. 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. – BMC Bioinform. 14: 16.

Dallas, T. 2016. helminthR: an R interface to the London Natural History Museum's host–parasite database. – Ecography 39: 391–393.

Data Citation Synthesis Group 2014. Joint declaration of data citation principles. – Force 11.

De Giovanni, R. et al. 2016. ENM components: a new set of web service-based workflow components for ecological niche modelling. – Ecography 39: 376–383.

Edwards, P. N. et al. 2011. Science friction: data, metadata, and collaboration. – Soc. Stud. Sci. 41: 667–690.

Ellison, A. M. 2010. Repeatability and transparency in ecological research. – Ecology 91: 2536–2539.

Gilbert, K. J. et al. 2012. Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program Structure. – Mol. Ecol. 21: 4925–4930.

Hampton, S. E. et al. 2013. Big data and the future of ecology. – Front. Ecol. Environ. 11: 156–162.

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

Iqbal, S. A. et al. 2016. Reproducible research practices and transparency across the biomedical literature. – PLoS Biol. 14: e1002333.

Kitzes, J. and Wilber, M. 2016. macroeco: reproducible ecological pattern analysis in Python. – Ecography 39: 361–367.

Krumholz, H. 2016. Journal editors to researchers: show everyone your clinical data. – NPR, <www.npr.org/sections/health-shots/2016/01/26/464010931/journal-editors-to-researchers-show-everyone-your-clinical-data>.

Miller, G. 2006. A scientist's nightmare: software problem leads to five retractions. – Science 314: 1856–1857.

Mooney, H. and Newton, M. 2012. The anatomy of a data citation: discovery, reuse, and credit. – J. Librariansh. Sch. Commun. 1: eP1035.

Naimi, B. and Araújo, M. B. 2016. sdm: a reproducible and extensible R platform for species distribution modelling. – Ecography 39: 368–375.

Nogués-Bravo, D. et al. 2011. Communities under climate change. – Science 334: 1070–1071.

Open Science Collaboration 2015. Estimating the reproducibility of psychological science. – Science 349: 943.

Peterson, A. T. et al. 2011. Ecological niches and geographic distributions. – Princeton Univ. Press.

Pettersson, L. B. and Rahbek, C. 2008. Editorial: launching Software notes. – Ecography 31: 3.

Poisot, T. et al. 2016a. mangal – Making ecological network analysis simple. – Ecography 39: 384–390.

Poisot, T. et al. 2016b. Synthetic datasets and community tools for the rapid testing of ecological hypotheses. – Ecography 39: 402–408.
Read, J. et al. 2016. geoknife: Reproducible web-processing of large gridded datasets. – Ecography 39: 354–360.
Robertson, M. P. et al. 2016. Biogeo: an R package for assessing and improving data quality of occurrence record datasets. – Ecography 39: 394–401.
Sandhu, L. 2014. Journal of Ecology is part of new BES data archiving policy. – <https://jecologyblog.wordpress.com/2014/01/20/journal-of-ecology-is-part-of-new-bes-data-archiving-policy/>.
Simundic, A.-M. and Nikolac, N. 2009. Statistical errors in manuscripts submitted to Biochemia Medica journal. – Biochem. Med. 19: 294–300.
The Economist 2016. Let's just try that again. – The Economist, <www.economist.com/news/science-and-technology/21690020-reproducibility-should-be-sciences-heart-it-isnt-may-soon>.
Vines, T. H. et al. 2014. The availability of research data declines rapidly with article age. – Curr. Biol. 24: 94–97.
Wolstencroft, K. et al. 2013. The Taverna workflow suite: designing and executing workflows of Web services on the desktop, web or in the cloud. – Nucleic Acids Res. 41: 557–561.