Synthetic datasets and community tools for the rapid testing of ecological hypotheses

Abstract

. . .

Ecologists are tasked with providing information on issues involving a variety of organisational scales. Notable examples include, in addition to the global biodiversity crisis (**???**), predicting the consequences of the loss of trophic structure (Estes et al. 2011), rapid shifts in species distributions (Gilman et al. 2010), and increased anthropogenic stress on species and their environment (**???**). Most of these pressing topics require to be addressed (i) at the global scale, or at the very least across a variety of environments and landscapes, and (ii) through the integration of several types of data (Thuiller et al. 2013). Because of these requirements, new sampling is not always a viable solution on its own: there is no funding structure to finance it, and there are time and scale constraints involved that make it unrealistic. While it is obvious that data collection should continue, we argue that there are a (potentially large) number of macroecological questions that could be adressed (i) without additional data and (ii) at minimal cost, by making use of open data and community-developed software and platforms.

{>> Halpern et al. 2008. A Global Map of Human Impact on Marine Ecosystems. Science 319: 948-952. <<}

Existing data can, to an increasing extent, be used to *build* new datasets (henceforth synthetic datasets). There are several parallel advances that make this approach possible. First, the volume of data on ecological systems that are available *openly* increases on a daily basis. This includes point- occurrence data, as in *e.g.* GBIF, but also taxonomic knowledge (through ITIS, NCBI or EOL), and trait and interactions data. A vast treasure trove of ecological information is now available without having to contact and secure authorization from every contributor individually. Second, this data is often available in a *programmatic* way. As opposed to manual collection, identification, and maintenance of datasets, most of these services implement web APIs, *i.e.* services that allow users to query and/or upload data in a standard format. These services can be queried, either once or on a regular basis, to retrieve records with the desired properties. This ensures that the process is repeatable, testable, transparent, and nearly error proof. Finally, most of the heavy-lifting for these tasks can be done through a *burgeoning ecosystem of packages and software*, that handles query formatting, data retrieval, and associated tasks, all the while exposing simple interfaces to researchers. None of these are *new* data, in the sense that these collections represent the aggregation of thousands of ecological studies; the originality lies in the ability to query, aggregate, curate, and use these data consistently using open solutions.

Hypothesis testing for large-scale systems is inherently limited by the availability of suitable datasets – most data collection results in small scale, local data, and it is not always clear how these can be used at more global scales. Perhaps as a result, developments in macroecology have primarily been driven by a search for patterns that are very broad both in scale and nature (Beck et al. 2012; Keith et al. 2012). While it is obvious that collecting exhaustive data at scales that are large enough to be relevant can be an insurmountable effort (both because of the monetary, time, and human costs needed), we suggest that macroecologists could, in parallel, build on existing databases, and aggregate them in a way that allows direct testing of proposals stemming from theory. To us, this opens no less than a new way for ecologists to ask critical research questions, spanning from the local to the global, and from the organismal to the ecosystemic, scales. Indeed, we live in a data-rich world, and a

very large amount of these data can now *easily* be collected, reducing the need for additional costly and time-consuming sampling. More importantly, this allows *rapid* evaluation of both climate change scenarios (**???**) and hypotheses. In this contribution, building on a real-life example, we (i) outline the basic approach of integrating data from a variety of sources (both in terms of provenance, and type of ecological information), (ii) identify technical bottlenecks, (iii) discuss issues related to scientific ethics and best practice, and (iv) provide clear recommendations moving forward.

1 An illustrative case-study

Food-web data, that is the determination of trophic interactions among species, are notoriously difficult to collect. The usual approach is to assemble literature data, expert knowledge, and additional information coming from field work, either as direct observation of feeding events or through gut-content analysis. Because of these technical constraints, food-web data are most often assembled in a single location. This impedes our ability to address the variation of their structure in space, which may both translate the action of macroecological mechanisms, as well as hold key to our ability to predict the spatial variation of ecological properties. As a consequence, most of the properties of food web over large (continental, global) spatial extents remain undocumented. For example, what is the relationship between latitude and connectance (the density of feeding interactions)?

One possible way to approach this question would be to collect data from different localities, and document through *e.g.* regressions the relationship between latitude and connected. The approach we will illustrate here uses broad-scale data integration to forecast the structure of a single system at the global scale. We are interested in predicting the structure of a pine-marsh food web, worldwide.

1.1 Interactions data

Food web data were take from **REF**, as made available in the `IWDB` database (URL). Marshes, like almost all wetlands, are critically endangered and serve as a home to a host of endemic biodiversity (Fensham et al. 2011; Minckley, Turner, and Weinstein 2013). Stream food webs in particular are important, both because they provide coupling between terrestrial and aquatic communities and ensure the maintenance of ecosystem services, but also because the increased pressure on wetlands makes them particularly threatened. They represent a prime example of ecosystems for which data-based prediction can be used to generate scenarios at a temporal scale relevant for conservation decisions, and faster than what sampling could allow.

The data comprising the original food web (105 nodes, including vague denominations like *Unidentified detritus* or *Terrestrial invertebrates*), were cleaned in the following way. First, all nodes were aggregated to the *genus* level. Due to high level of structure in trophic

interactions emerging from taxonomic rank alone (Eklof et al. 2011,**???**), aggregating to the genus level has the double advantage of (i) removing ambiguities on the identification of species and (ii) allowing us to integrate data when any two species from given genera interact. Second, all nodes that were not identified (`Unidentified` or `Unknown` in the original data) were removed. The cleaned network documented 227 interactions, between 80 genera.

{>>CHECK AND REDO IF NEEDED<<}

Using the name checking functions from the `taxize` package (S. A. Chamberlain and Szöcs 2013) revealed that all of these genus names were valid.

Because the original food web was sampled *locally*, there is the possibility that interactions between genera are not reported; either because species from these genera do not interact, or co-occur, in the sampling location. To circumvent this, we queried the *GLOBI* database (Poelen, Simons, and Mungall 2014) for each genus name, and retrieved all *feeding* interactions. For all *new* genera retrieved through this method, we also retrieved their interactions with genera already in the network. The inflated network (original data, and data from *GLOBI*) has 789 genera, and a total of 9328 interactions between them.

The code to reproduce this analysis is in the `1_get_data.r` suppl. file.

{>>CHECK ALL NUMERICAL VALUES<<}

1.2 Occurrence data and filtering

For each genera, we downloaded the known occurrences from GBIF and BISON. This yielded 64 763 point-presence records. Because the ultimate goal is to perform spatial modeling of the structure of the network, we removed genera for which fewer than 100 occurrences were known. This seems like a stringent filter, yet it enables to (i) maintain sufficient predictive powers for SDMs, and (ii) only work on the genera for which we have "high-quality" data. Genera with fewer than 100 records were removed from the occurrence data and from the metanetwork. The final metanetwork therefore has 4271 interactions between 188 genera. Given the curated publicly available data, it represents the current best description of feeding interactions between species of this ecosystem. A visual depiction of the network is given in *Fig. 1*.

On its own, the fact that filtering for genera with over 100 records reduced the sample size from 739 genera to 188 indicates how crucial it is that observations are reported in public databases. This is because the type of analysis we present here, although cost-effective and enabling rapid evaluation of different scenarios, is only as good as the underlying data. Since most modeling tools require a minimal sample size in order to achieve acceptable accuracy, concerted efforts by the community and funding agencies to ensure that the minimal amount of data is deposited upon publication or acquisition is needed.

The code to reproduce this analysis is in the `1_get_data.r` suppl. file.

1.3 Species Distribution Model

For each species in this subset of data, we retrieved the nineteen `bioclim` variables (Hijmans et al. 2005), with a resolution of 5 arc-minutes. This enabled us to build climatic envelope models for each species. These models tend to be more conservative than alternate modeling strategies, in that they predict smaller range sizes (Hijmans and Graham 2006), but they also perform well overall for presence-only data (Elith et al. 2006). The output of these models is, for species i, the probability of an observation $\mathrm{P}(i)$ within each pixel. We appreciate that this is a coarse analysis, but its purpose is only to highlight how the different data can be combined. A discussion of the limitations of this approach is given below.

The code to reproduce this analysis is in the `2_get_sdm.r` suppl. file.

1.4 Assembly

For each of the 4271 interactions in the metanetwork, we estimated the probability of it being observed in each pixel as the product of the probabilities of observing each species on its own: $\mathrm{P}(L_{ij}) \propto \mathrm{P}(i)\mathrm{P}(j)$. This resulted in 4271 LDMs ("link distribution models"). The code to reproduce this analysis is in the `3_get_ldm.r` suppl. file.

Based on this information, we generated the following illustrations (using `4_draw_figures.r`). First, a map of species richness (*Fig. 2A*) and number of interactions (*Fig. 2B*). Second, a map of *connectance* (*Fig. 2C*), which is the number of interactions divided by the squared species richness. Finally, a scatterplot of connectance as a function of latitude (*Fig. 2D*), which reveals a systematic macroecological trend. Interestingly, this last panel shows a strong response to this system to the fact that the tropics in Africa are surrounded by deserts in which the species studied here are not predicted to occur given the climatic variables.

2 Challenges moving forward

Attribution stacking and intellectual provenance:

The merging of large databases has already created a conflict of how to properly attribute data provenance. Here there are at least two core issues that will require community consultation in order to be resolved. First, *what is the proper mode of attribution when a very large volume of data is aggregated*? Second, *what should be the intellectual property of the synthetic dataset*? Currently, citations (whether to articles or datasets) are only counted when they are part of the main text. The simple example outlined here relies on well over a thousand references, and it makes little sense to expect that they would be provided in the main text. One intermediate solution would be to collate these references in a supplement, but it is unclear that these would be counted, and therefore contribute to the *impact* of each individual dataset. This is a problem that we argue is best solved by publishers; proper attribution

and credit is key to provide incentives to data release (Whelan, Schimel, and Baldwin 2014; Kenall, Harold, and Foote 2014). As citations are currently the currency of scientific impact, publishers have a responsibility not only to ensure that data are available (which many already do), but that they are recognized; data citation, no matter how many data are cited, is a way to achieve this goal. The synthetic dataset, on the other, can reasonably be understood as a novel product; there is technical and intellectual effort involved in producing it, and although it is a derivative work, we would encourage authors to deposit it anew.

Sharing of code and analysis pipeline:

Ideally, authors should release their analysis *pipeline* in addition to the data and explanation of the steps. The pipeline can take the form of a `makefile` (which allows to generate the results, from the raw data, without human intervention), or be all of the relevant code that allows to re-generate the figures and results. For example, we have released all of the `R` code that was used to generate the figures at **XXX**. Sharing the analysis pipeline has several advantages. First, it is a first stepv towards ensuring the quality of analyses, since reviewers can (and should reasonably be expected to) look at the source code. Second, it provides a *template* for future analyses – instead of re-developing the pipeline from scratch, authors can re-use (and acknowledge) the previous codebase and build on it. Finally, it helps identifying areas of future improvement. The development of software should primarily aim to make the work of researchers easier. Looking at commonalities in the analytical pipelines for which no ready-made solutions exists will be a great way to influence priorities in software development.

Computational literacy:

This approach hardly qualifies as *big data*; nevertheless, it relies on the management and integration of a large volume of heterogeneous information, both qualitatively larger than the current "norm". The first challenge is being able to *manage* this data; it requires data management skills that are not usually needed when the scale of the dataset is small, and, fallible though the process may be, when data can reasonably be inspected manually. The second challenge is being able to *manipulate* these data; even within the context of this simple use-case, the data do not fit in the memory of `R` (arguably the most commonly known and used software in ecology) without some adjustments. Once these issues were overcome, running the analysis involved a few hours worth of computation time. Since there is little doubt that computational approaches are going to become increasingly common in ecology (Hampton et al. 2013), and are identified by the community as both in-demand skills and as not receiving enough attention in current ecological curricula (Barraquand et al. 2014), it seems that efforts should be allocated to raise the computational literacy of ecologists, and recognize that there is value in the diversity of tools one can use to carry out more demanding studies. For example, both `Python` and `Julia` are equally as user friendly as `R` while also being more powerful and better suited for computationally- or memory-intensive analyses.

Standards and best practices:

In conducting this analysis, we noticed that a common issue was the identification of species

and genera. All of these datasets were deposited by individual scientists; whether we like it or not, individuals are prone to failure in a very different way that the "Garbage in, garbage out" idea that applies to computer programs. Using tools such as `taxize` (S. A. Chamberlain and Szöcs 2013) can allow us to resolve a few of the uncertainties, yet this must be done every time the data are queried and requires the end user to make educated guesses as to what the "true" identity of the species is. These limitations can be overcome, on two conditions. Database maintainers should implement automated curation of the data they have the stewardship of, and identify potential mistakes and correct them upstream, so that users download high-quality, high-reliability data. Data contributors should rely more extensively on biodiversity identifiers (such as TSN, GBIF, NCBI Taxonomy ID, ...), to make sure that even when there are typos in the species name, they can be matched across datasets. Constructing this dataset highlighted a fundamental issue: the rate-limiting step is rarely the availability of appropriate tools or platforms, but instead it is the adoption of common standards and the publication of data in a way that conforms to them. In addition, Maldonado et al. (2015) emphasize that point-occurence data are not always properly reported – for example, the center of a country or region can be used when no other information is known; this requires an improved dialogues betweent data collectors and data curators, to highlight which practices have the highest risk of biasing future analyses.

Propagation of error:

There are always caveats to using synthetic datasets. First, the extent to which each component dataset is adequately sampled is unknown (although there exist ways to assess the overall representativeness of the assembled dataset; Schmill et al. (2014)). This can create gaps in the information that is available when all component datasets are being merged. Second, because it is unlikely that all component datasets were acquired using reconcilable standards and protocol, it is likely that much of the quantitative information needs be discarded, and therefore the conservative position is to do qualitative analyses only. Although these have to be kept in mind, we do not think they should prevent use and evaluation of the approach we suggest. For one thing, as we illustrate, at large spatial and organizational scales, coarse- grained analyses are still able to pick up qualitative differences in community structure. Second, most emergent properties are relatively insensitive to fine- scale error; for example, Gravel et al. (2013) show that even though a simple statistical model of food-web structure mispredicts some individual interactions, it produces communities with realistic emergent properties. Which level of error is acceptable needs to be determined for each application, but we argue that the use of synthetic datasets is a particularly cost- and time-effective approach for broad-scale description of community- level measures.

3 Conclusion – why not?

- need to try
- fast, and relies on existing "classical" data
- generate testable predictions
- can identify gaps in knowledge – missing data!

- necessary step before real-time predictions Antonelli et al. (2014)

Acknowledgments – This work was funded in part through a grant from the Canadian Institute of Ecology and Evolution. TP was funded by a Starting grant from the Université de Montréal, and a NSERC Discovery Grant. DBS acknowledges a Marsden Fund Fast-Start grant (UOC-1101) and Rutherford Discovery Fellowship, both administered by the Royal Society of New Zealand. We thank Kévin Cazelles for constructive comments on the manuscript.

4 References

Antonelli, Alexandre, Fabien L. Condamine, Hannes Hettling, Karin Nilsson, R. Henrik Nilsson, Bengt Oxelman, Michael J. Sanderson, et al. 2014. "SUPERSMART: Ecology and Evolution in the Era of Big Data." e501v1. PeerJ PrePrints. https://peerj.com/preprints/501v1.

Barraquand, Frédéric, Thomas H.G. Ezard, Peter S. Jørgensen, Naupaka Zimmerman, Scott Chamberlain, Roberto Salguero-Gómez, Timothy J. Curran, and Timothée Poisot. 2014. "Lack of Quantitative Training Among Early-Career Ecologists: a Survey of the Problem and Potential Solutions." *PeerJ* 2 (March): e285. doi:10.7717/peerj.285.

Beck, Jan, Liliana Ballesteros-Mejia, Carsten M. Buchmann, Jürgen Dengler, Susanne A. Fritz, Bernd Gruber, Christian Hof, et al. 2012. "What's on the Horizon for Macroecology?" *Ecography* 35 (8) (August): 673–683. doi:10.1111/j.1600-0587.2012.07364.x.

Chamberlain, Scott A., and Eduard Szöcs. 2013. "Taxize: Taxonomic Search and Retrieval in R." *F1000Research* (October). doi:10.12688/f1000research.2-191.v2.

Eklof, Anna, Matthew R. Helmus, M. Moore, Stefano Allesina, and Anna Eklöf. 2011. "Relevance of Evolutionary History for Food Web Structure." *Proc. R. Soc. B Biol. Sci.* 279 (November 2011) (November): 1588–1596. doi:10.1098/rspb.2011.2149.

Elith, Jane, Catherine H. Graham, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, et al. 2006. "Novel Methods Improve Prediction of Species' Distributions from Occurrence Data." *Ecography* 29 (2): 129–151. doi:10.1111/j.2006.0906-7590.04596.x.

Estes, James A, John Terborgh, Justin S Brashares, Mary E Power, Joel Berger, William J Bond, Stephen R Carpenter, et al. 2011. "Trophic Downgrading of Planet Earth." *Science* 333 (6040) (July): 301–306. doi:10.1126/science.1205106.

Fensham, R. J., J. L. Silcock, Adam Kerezsy, and W. Ponder. 2011. "Four Desert Waters: Setting Arid Zone Wetland Conservation Priorities Through Understanding Patterns of Endemism." *Biol. Conserv.* 144 (10): 2459–2467. http://www.sciencedirect.com/science/article/pii/S0006320711002515.

Gilman, Sarah E, Mark C Urban, Joshua Tewksbury, George W Gilchrist, and Robert D Holt. 2010. "A Framework for Community Interactions Under Climate Change." *Trends Ecol. Evol.* 25 (6) (June): 325–331. doi:10.1016/j.tree.2010.03.002.

Gravel, Dominique, Timothée Poisot, Camille Albouy, Laure Velez, and David Mouillot. 2013. "Inferring Food Web Structure from Predatorprey Body Size Relationships." *Methods Ecol Evol* 4 (11) (November): 1083–1090. doi:10.1111/2041-210X.12103.

Hampton, Stephanie E, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. 2013. "Big Data and the Future of Ecology." *Frontiers in Ecology and the Environment* 11 (3) (March): 156–162. doi:10.1890/120103.

Hijmans, Robert J., and Catherine H. Graham. 2006. "The Ability of Climate Envelope Models to Predict the Effect of Climate Change on Species Distributions." *Glob. Change Biol.* 12 (12): 2272–2281. doi:10.1111/j.1365-2486.2006.01256.x.

Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis. 2005. "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *Int. J. Climatol.* 25 (15): 1965–1978. doi:10.1002/joc.1276.

Keith, Sally A., Tom J. Webb, Katrin Böhning-Gaese, Sean R. Connolly, Nicholas K. Dulvy, Felix Eigenbrod, Kate E. Jones, et al. 2012. "What Is Macroecology?" *Biology Letters* (August): rsbl20120672. doi:10.1098/rsbl.2012.0672.

Kenall, Amye, Simon Harold, and Christopher Foote. 2014. "An Open Future for Ecological and Evolutionary Data?" *BMC Ecology* 14 (1): 10. doi:10.1186/1472-6785-14-10. `http://www.biomedcentral.com/1472-6785/14/10`.

Maldonado, Carla, Carlos I. Molina, Alexander Zizka, Claes Persson, Charlotte M. Taylor, Joaquina Albán, Eder Chilquillo, Nina Rønsted, and Alexandre Antonelli. 2015. "Estimating Species Diversity and Distribution in the Era of Big Data: to What Extent Can We Trust Public Databases?" *Global Ecology and Biogeography* (May): n/a–n/a. doi:10.1111/geb.12326. `http://dx.doi.org/10.1111/geb.12326`.

Minckley, T. A., D. S. Turner, and S. R. Weinstein. 2013. "The Relevance of Wetland Conservation in Arid Regions: a Re-Examination of Vanishing Communities in the American Southwest." *J. Arid Environ.* 88: 213–221. `http://www.sciencedirect.com/science/article/pii/S0140196312002431`.

Poelen, Jorrit H., James D. Simons, and Chris J. Mungall. 2014. "Global Biotic Interactions: An Open Infrastructure to Share and Analyze Species-Interaction Datasets." *Ecological Informatics*. doi:10.1016/j.ecoinf.2014.08.005.

Schmill, Matthew D., Lindsey M. Gordon, Nicholas R. Magliocca, Erle C. Ellis, and Tim Oates. 2014. "GLOBE: Analytics for Assessing Global Representativeness." *2014 Fifth International Conference on Computing for Geospatial Research and Application* (Aug). doi:10.1109/com.geo.2014.21 `http://dx.doi.org/10.1109/COM.Geo.2014.21`.

Thuiller, Wilfried, Tamara Münkemüller, Sébastien Lavergne, David Mouillot, Nicolas Mouquet, Katja Schiffers, and Dominique Gravel. 2013. "A Road Map for Integrating Eco-Evolutionary Processes into Biodiversity Models." *Ecol. Lett.* 16: 94–105. doi:10.1111/ele.12104.

Whelan, Anne Marie, David S. Schimel, and J. David Baldwin. 2014. "Editorial." *Ecological Applications* 24 (1) (January): 1–2. doi:10.1890/1051-0761-24.1.1. `http://dx.doi.org/10.1890/1051-0761-24.1.1`.