Synthetic datasets for global ecology in a data-rich world

see CONTRIBUTORS (add your name)

work in progress

Ecologists are tasked with providing information on issues at the global scale. 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 (**???**), and . Most of these pressing topics require to be addressed (i) at the global scale and (ii) through the integration of several types of data (Thuiller et al. 2013). Because of these requirements, new sampling is not a viable solution: there is no funding structure to finance it, and there are time and scale constraints involved that make it unrealistic.

Thankfully, developments in the recent years means that ecologists can now leverage existing data, and use them to *build* new datasets (henceforth synthetic datasets) suitable for the questions at hand. 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-occurence data, as in *e.g.* GBIF, ITIS, but also taxonomic knowledge (through 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 available in a *programmatic* way. As opposed to manual collection, identification, and curation of datasets, most of these services implement web API that allow to query them, either once or on a regular basis, to retrieve records with the desired properties. This ensures that the process is repeatable, testable, transparent, and error-proof. Finally, most of the heavy-lifting for these tasks can be done through a *burgeoning ecosystem of packages and software*, that take care of handling query formatting, data retrieval, etc, and expose simple interfaces to researchers.

To us, this opens no less than a new area of research for ecologists interested in asking questions at large spatial and organizational scales – 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 scenarios. In this contribution, building on a real-life example, we (i) outline the basic approach, (ii) identify technical bottlenecks, (iii) discuss ethical considerations, and (iv) provide clear recommendations moving forward.

An illustrative case-study

Food-web data, that is the determination of trophic interactions among species, are difficult to collect. The usual approach is to assemble literature data, expert knowledge, and additional information coming from field work, either as observation of direct feeding events, or through the analysis of gut content. 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 macro-ecological mechanisms, as well as hold key to our ability to predict the spatial variation of ecological properties.

In this case study, we are interested in predicting the structure of the pine-marsh food web worldwide. One example of this system has been described in detail by (**???**). We will show how coupling these data with additional interactions reported in the literature, as well as species occurrence data, allows building a predictive model that describes the spatial variation of this system.

Interactions data

Data from the pine-marsh food web were take from **REF**, as made available in the `IWDB` database (URL). Marshes, as almost all wetlands, are critically endangered and home to a host of endemic biodiversity (Fensham et al. 2011, Minckley et al. 2013). 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 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 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). The cleaned network documented 227 interactions, between 80 genera. Using the name checking functions from the `taxize` package (Chamberlain and Szöcs 2013) revealed that all of these genus names were valid.

Because the original foodweb was sampled *locally*, there is the possibility that interactions between genera are not reported. To circumvent this, we queried the *GLOBI* database (Poelen et al. 2014) for each genus name, and retrieved all *feeding* interactions. For all *new* genera retrieved through this method, we retrieved their interactions with genera already in the network. This network has 789 genera, and a total of 9328 interactions.

Occurrence data and filtering

For each genera, we downloaded the known occurrences from GBIF and BISON. This yielded 64763 point-presence records. Because the goal is, ultimately, to do spatial modeling of the structure of the network, we removed genera for which less 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 less than 100 records were removed from the occurrence data and from the metanetwork. The final metanetwork 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*.

In itself, 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. 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. A concerted efforts by the community and funding agencies to ensure that the minimal amount of data is deposited upon publication or acquisition is needed.

Species Distribution Model

For each species in this subset of data, we retrieved the 19 `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 perform well overall on presence-only data (Elith et al. 2006). The output of these models is, for species i, the probability of an observation $\text{P}(i)$ 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.

Assembly

For each of the 4271 interactions in the metanetwork, we measured the probability of it being observed in each pixel as being the joint probability of observing both species: $\text{P}(L_{ij}) \propto \text{P}(i)\text{P}(j)$. This resulted in 4271 LDMs ("link distribution models"). Based on these informations, we generated the following illustrations. 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 the Africa, are surrounded by desertic areas in which the species studied here are not predicted to occur given the climatic variables.

Opportunities

Hypothesis testing for large-scale systems is inherently limited by the availability of suitable datasets. Perhaps as a result, macro-ecology has been guided by a search for patterns that are very broad both in scale and nature (**???**), as opposed to the testing of pre-established hypothesis. While it is obvious that collecting data at scales that are large enough to be relevant is an insurmountable effort (both because of the monetary, time, and human costs neeeded), we suggest that macroecologists should build on existing databases, and aggregate them in a way that allows direct testing of proposals stemming from theory.

Challenges

Attribution stacking and intellectual paternity:

The merging of large databases is already asking the question of proper attribution of data paternity. Namely, there are two core issues that need 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*?

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 (failible though this process may be) 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. It is now worth asking whether our total reliance on this tool (as opposed to more performing yet equally user-friendly languages as `python` or `julia`) is going to pay off in the long term. Since it makes very 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 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.

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 people, and are prone to failure. Using tools such as `taxize` (Chamberlain and Szöcs 2013) allowed to resolve a few of the uncertainties. Yet this has to 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, it can be matched across datasets.

Propagation of error:

There are some caveats to using synthetic datasets. First, the extent to which each component dataset is adequately sampled is unknown. This can create gaps in the information that is available *in fine*. Second, because it is unlikely that all component datasets were acquired using reconcilable standards and protocol, it is likely that 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, 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 for broad-scale description of community-level measures, the use of synthetic datasets is a cost and time-effective approach.

Recommendations

1. Publish data (even the small one!)
2. Publish pipeline
3. Pay attention to standard when releasing data

Acknowledgements – 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.

References

Barraquand, F. et al. 2014. Lack of quantitative training among early-career ecologists: a survey of the problem and potential solutions. - PeerJ 2: e285.

Chamberlain, S. A. and Szöcs, E. 2013. taxize: taxonomic search and retrieval in R. - F1000Research in press.

Eklof, A. et al. 2011. Relevance of evolutionary history for food web structure. - Proc. R. Soc. B Biol. Sci. 279: 1588–1596.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. - Ecography 29: 129–151.

Estes, J. A. et al. 2011. Trophic Downgrading of Planet Earth. - Science 333: 301–306.

Fensham, R. J. et al. 2011. Four desert waters: setting arid zone wetland conservation priorities through understanding patterns of endemism. - Biol. Conserv. 144: 2459–2467.

Gravel, D. et al. 2013. Inferring food web structure from predatorprey body size relationships. - Methods Ecol Evol 4: 1083–1090.

Hampton, S. E. et al. 2013. Big data and the future of ecology. - Frontiers in Ecology and the Environment 11: 156–162.

Hijmans, R. J. and Graham, C. H. 2006. The ability of climate envelope models to predict the effect of climate change on species distributions. - Glob. Change Biol. 12: 2272–2281.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. - Int. J. Climatol. 25: 1965–1978.

Minckley, T. A. et al. 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.

Poelen, J. H. et al. 2014. Global Biotic Interactions: An open infrastructure to share and analyze species-interaction datasets. - Ecological Informatics in press.

Thuiller, W. et al. 2013. A road map for integrating eco-evolutionary processes into biodiversity models. - Ecol. Lett. 16: 94–105.