Global ecology in a data-rich world

see CONTRIBUTORS

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 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 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.

Due to the limitations, … (qualitative rather than quantitative prediction)

Interactions data

Data from the pine-marsh food web were take from **REF**, as made available in the `IWDB` database (URL).

good rationale for using genus level Eklof et al. (2011) .

Occurrence data

Cleaning

Step 2 – Clean species names.

Step 3 – Species occurrence.

Step 4 – Species distribution model.

Step 5 – Interactions distribution model.

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

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 newly aggregated dataset*?

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 ... **REF**. The second challenge is being able to *manipulate* these data; small though they may be, the data from this case study do not fit in the memory of `R` (arguably the most commonly known software in ecology) without some adjustments. 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 **REF**, and are one areas where the community identifies lower skills that would be needed (Barraquand et al. 2014), it seems evident that developing computational litteracy should be a part of the core ecological curriculum.

1. Identifying species
2. Hoping that the noise averages out under large volume of data
3. Need for automated curation

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.

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

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

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