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 question at large scales – we live in a data-rich world, and a very large amount of these data can now *easily* be collected to address questions at large scales, without 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

Rodents and their parasites are distributed throughout Eurasia, in a wide range of habitats (**???**). Both can serve as vectors for human pathogens (**???**), and this is especially a problem in areas of important ecological contact or high poverty. Predicting where rodents and their parasite will distribute is emerging as a major public health challenge in some areas of the world (**???**). Obtaining data from these systems is difficult (**???**) – rodents are hard to trap, have a large area of distribution, and parasitic investigation requires a tremendous amount of man-hours. This is especially true since parasites do not interact consistently with the same hosts through space (Poisot et al. 2013, Canard et al. 2014) On the other hand, rodents distributions are usually well-predicted by climatic variables (**???**), which allows for predictive approaches instead of, or in complement to, additional sampling.

In this case-study, we will show how several tools can be easily integrated to (i) assemble a new dataset and (ii) use it to

Step 1 – Species interactions. Hadfield et al. (2014) have established a list of parasitic interactions between XX species of rodents and YY species of fleas in Eurasia. These data have been deposited to the *mangal* database (**???**), and can be retrieved directly from `R`.

```
library(rmangal)
library(betalink)
api <- mangalapi()
interaction_metadata <- getDataset(api, 2)
interaction_data <- metaweb(toIgraph(api, interaction_metadata$networks))
```

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*?

1. Data curation
2. Large memory required
3. Need computational litteracy

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

Recommendations

1. Publish pipeline!!!
2. 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

Canard, E. F. et al. 2014. Empirical evaluation of neutral interactions in host-parasite networks. - Am. Nat. 183: 468–479.

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

Hadfield, J. D. et al. 2014. A Tale of Two Phylogenies: Comparative Analyses of Ecological Interactions. - Am. Nat.: 000–000.

74 Poisot, T. et al. 2013. Facultative and obligate parasite communities exhibit different network properties. -
75 Parasitology 140: 1340–1345.

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