Peer Community In Ecology

Open Access

RECOMMENDATION

# Upscaling the neighborhood: how to get species diversity, abundance and range distributions from local presence/absence data

Matthieu Barbier[1]

[1] Centre for Biodiversity Theory and Modelling, CNRS – Moulis, France

**Cite as:** Barbier M. Upscaling the neighborhood: how to get species diversity, abundance and range distributions from local presence/absence data. *Peer Community In Ecology*, 100009 (2019). DOI: 10.24072/pci.ecology.100009

**Published:** 10th january 2019

**Based on reviews by:**
Kevin Cazelles and one anonymous reviewer

**Correspondence:**
contact@mrcbarbier.org



**A recommendation of**

Tovo A, Formentin M, Suweis S, Stivanello S, Azaele S, and Maritan A. Inferring macro-ecological patterns from local species' occurrences. *bioRxiv* 387456, ver. 3 peer-reviewed and recommended by *PCI Evol Biol* (2019). DOI: 10.1101/387456

How do you estimate the biodiversity of a whole community, or the distribution of abundances and ranges of its species, from presence/absence data in scattered samples?

It all starts with the collector's dilemma: if you double the number of samples, you will not get double the number of species, since you will find many of the same common species, and only a few new rare ones.

This non-additivity has prompted many ecologists to study the Species-Area Relationship. A common theoretical approach has been to connect this spatial pattern to the overall distribution of how common or rare a species can be. At least since Fisher's celebrated log-series [1], ecologists have been trying to, first, infer the shape of the Species Abundance Distribution, and then, use it to predict how many species should be found in a given area or a given number of samples. This has found many applications, from microbial communities to tropical forests, from estimating the number of yet-unknown species to predicting how much biodiversity may be lost if a fraction of the habitat is removed.

In this elegant work, Tovo et al. [2] propose a method that starts only from presence/absence data over a number of samples, and provides the community's diversity, as well as its abundance and range size distributions. This method is simple, analytically explicit, and accurate:

the authors test it on the classic Pasoh and Barro Colorado Island tropical forest datasets, and on simulated data. They make a very laudable effort in both explaining its theoretical underpinnings, and proposing a straightforward step-by-step guide to applying it to data.

The core of Tovo et al's method is a simple property: the scale invariance of the Negative Binomial (NB) distribution. Subsampling from a NB gives another NB, where a single parameter has changed. Therefore, if the Species Abundance Distribution is close enough to some NB (which is flexible enough to accommodate all the data here), we can estimate how this parameter changes when going from (1) a single sample to (2) all the available samples, and from there, extrapolate to (3) the entire community.

This principle was first applied by the authors in a previous study [3] that required abundance data in the samples, rather than just presence/absence. Given that binary occurrence data is far more available in a variety of empirical settings, this extension is worthwhile (including its new predictions on range size distributions), and it deserves to be widely known and tested.

ADDITIONAL COMMENTS

1) To explain the novelty of the authors' contribution, it is useful to look at competing techniques.

Some "parametric" approaches try to infer the whole-community Species Abundance Distribution (SAD) by guessing its functional form (Gaussian, power-law, log-series...) and fitting its parameters from sampled data. The issue is that this distribution shape may not remain in the same family as we increase the sampling effort or area, so the regression problem may not be well-defined. This is where the Negative Binomial's scale invariance is useful.

Other "non-parametric" approaches have renounced guessing the whole SAD: they simply try to approximate of its tail of rare species, by looking at how many species are found in only one (or a few) samples. From this, they derive an estimate of biodiversity that is agnostic to the rest of the SAD. Tovo et al. [2] show the issue with these approaches: they extrapolate from the properties of individual samples to the whole community, but do not properly account for the bias introduced by the amount of sampling (the intermediate scale (2) in the summary above).

2) The main condition for all such approaches to work is well-mixedness: each sample should be sufficiently like a lot drawn from the same skewed lottery. As long as that condition applies, finding the best approach is a theoretical matter of probabilities and combinatorics that may, in time, be given a definite answer.

The authors also show that "well-mixed" is not as restrictive as it sounds: the method works both on real data (which is never perfectly mixed) and on simulations where species are even

more spatially clustered than the empirical data. In addition, the Negative Binomial's scale invariance entails that, if it works well enough at some spatial scale, it will also work at all higher scales (until one reaches the edges of the sufficiently-well-mixed community)

3) One may ask: why the Negative Binomial as a Species Abundance Distribution?

If one wishes for some dynamical explanation, the Negative Binomial can be derived from neutral birth and death process with immigration, as shown by the authors in [3]. But to be applied to data, it should only be able to approximate the empirical distribution well enough (at all relevant scales).

Depending on one's taste, this type of probabilistic approaches can be interpreted as:
- purely phenomenological, describing only the observational process of sampling from an existing state of affairs, not the ecological processes that gave rise to that state.
- a null model, from which everything in practice is expected to deviate to some extent.
- or a way to capture the statistical forces that tend to induce stable relationships between different patterns (as long as no ecological process opposes them strongly enough).

## References


[1] Fisher RA, Corbet AS, and Williams CB. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* (1943), 42–58. DOI: 10.2307/1411.

[2] Tovo A, Formentin M, Suweis S, Stivanello S, Azaele S, and Maritan A. Inferring macro-ecological patterns from local species' occurrences. *bioRxiv* 387456, ver. 3 peer-reviewed and recommended by *PCI Evol Biol* (2019). DOI: 10.1101/387456.

[3] Tovo A, Suweis S, Formentin M, Favretti M, Volkov I, Banavar JR, Azaele S, and Maritan A. Upscaling species richness and abundances in tropical forests. *Science Advances* 3 (2017), e1701438. DOI: 10.1126/sciadv.1701438.


## Appendix


Reviews by Kevin Cazelles and one anonymous reviewer, DOI: 10.24072/pci.ecology.100009