Report on 'Inferring macro-ecological patterns from local species' occurrences' by Tovo et al.

Dear authors,

This is my review of your manuscript Report on 'Inferring macro-ecological patterns from local species' occurrences'. My comments are meant to be constructive, and I hope they will be helpful as you revise your manuscript. Note that I used markdown to write this review so you will find some tags in the text below. For the sake of clarity (especially for equations), I have attached a pdf version of my review.

Sincerely,

1 Overall opinion / major comments

In this manuscript Anna Tovo and colleagues propose a statistical framework that allows the inference of several biodiversity patterns based on the matrix of presence-absence matrix, the assumption that the Relative Species Abundance (RSA) follows a Negative-Binomial (NB) distribution and the absence of strong spatial correlations among the set of species considered. The manuscript is overall clear, well-structured and deals with a topic that could interest many ecologists as it proposes to derive valuable information based on the widely used presence-absence matrices.

Despite the quality of the manuscript, I believe that this manuscript requires some work to be recommended by PCI Ecology. In particular, the authors must put substantial efforts into the methods section in order to 1- better explain how the framework work, 2- better emphasize the difference between this manuscript their previous study (Tovo et al. 2017) and 3- to discuss the scope of their approach. In the following lines, I did my best to detail my concerns regarding these points.

1.1 About the theoretical framework

As mentioned above, this paper may interest many ecologists given its topic. I however think that the current version of the manuscript may appear quite impenetrable for many of them due to the lack of explanations of some mathematical concepts and notations in the methods section. Below I provide several suggestions to improve the clarity of this section.

1. I think it would be very helpful to remember what a Negative Binomial distribution is. A single sentence would be sufficient, something along this: "in a succession of Bernouilli trials in which the success

probability is ξ, the negative binomial distribution of parameter (r, ξ) models the probabilities associated to the number of trials required to obtain exactly r successes."

2. Throughout the methods section, the authors use P and $\mathcal{P}$. To me, this is quite confusing as according to the text they both are probabilities, so why using these notations?

3. Point 2 is also confusing because it seems like $P(n|1)$ is actually not a probability. Indeed $\mathcal{P}(n|r, \xi)$ is a negative binomial, so

$$\sum_{n}^{+\infty} \mathcal{P}(n|r, \xi) = 1$$

But then,

$$\sum_{n}^{+\infty} P(n|1) = c(r, \xi)$$

so, "the probability that a species has exactly n individuals – at the whole forest scale" do not sum to 1, what am I missing here? Why another "NB normalization" is needed? There is a explanation of this in (Tovo et al. 2017) but it should be clarify here.

4. *"[. . .], the conditional probability that a species has k individuals in the smaller area a = pA, given that it has total abundance n in the whole forest of area A is given by the binomial distribution"* (p.4). If I am correct, I would explain here that the assumption of spatial correlation is important to use the binomial distribution here.

5. Why do the authors use $\widehat{\xi_p}$ rather than $\widehat{\xi_p}$? Is it because it is an estimator? This should be clarified. In the same vein, $\equiv U(p|p^*, \widehat{\xi}_{p^*})$ in equation (4) may prevent readers from understanding the demonstration, this should be commented with words.

6. I think it would have been clearer to state: "S_p(k) denotes the number of species having k individuals at the scale p" which is directly applicable for $p^* $.

7. "*[. . .], by assuming that the global scale p = 1 is actually the one where we have data, [. . .].*" (p.6). I think the authors should develop what this means practically.

8. "Under the mean field hypothesis, [. . .]" (p.6). The authors should remind the readers what this means.

9. "*In words: for every scale pk, we compute the empirical average of the species observed in all subsets of k cells.*" (p.8). How the authors deal with this from a numerical standpoint, because, for instance, choosing 100 cells among 98x98, represent more than 10^{240} possibilities, what did I miss?

1.2 What's new?

There are a high degree of similarity between what is done here and what the authors did in (Tovo et al. 2017). The goal are the same, the method is here applied on two data sets included in the previous studies and many equations are identical. Moreover the reader should refer to this previous study to fully understand the

demonstration. I think it is very important that the authors better explain what has been done in their previous study and what brings this new manuscript.

My feeling is that the authors have two options. The first is, to restrict their manuscript to its novelty. What I mean here is that the authors could build on the top of (Tovo et al. 2017) without repeating the equations they have already published and make it very clear what they did at the beginning of the section and then highlight the new developments. In the current version, there are scattered sentences about the previous study but I am still wondering to what extent this study is new. For instance, p.14, the authors wrote: "*Our framework not only provides a generalization of the method recently proposed in [29], [. . .]* but I do not see the generalization in this paper. The second option is a very didactic paper to better guide readers through the framework and convince them to use it. This is the one I would recommend and I think the advice I provided above may be useful in this respect. In this didactic piece, I would suggest to add a few sentences about the numerical implementation of it, starting by mentioning where to find it.

1.3 About the scope of the methods

The authors must discuss the scope of their framework thoroughly. In the current version of the manuscript, the authors have applied their framework on simulated forests and two tropical forests. So a first question: is this method only designed for tree species in tropical forest? I guess no, otherwose they would not have present their method as a general one in the abstract. I however question its generality, especially given that I think the detectability of the species is crucial to apply this method, am I right? This would explain why tree species are well-suited to exemplify the method. I believe this is a first point to be added when discussing the scope of the method. A second is linked with the assumption introduces page 4:

> "*Assuming that the local RSA is not affected by spatial correlations due to both inter-specific and intra-specific interactions and strong environmental gradients [. . .].*"

To me, this should be further developed. For instance, how this could be tested? I think the authors should further stress that the assumptions made regard the relationships at the scale of the set of species, not the individuals species and then discuss what group of species is likely to meet the assumptions.

1.4 Specific commets

- p.1 (abstract, point 3): *"This framework, derived from first principles on the basis of biological processes[. . .]"*, I do not understand why this framework is "derived from first principles on the basis of biological processes".
- p.1 "*[. . .] as well as a new emergent pattern, which is the Relative Species Occupancy (RSO).*" The authors introduced and used the RSO throughout their manuscrit. I do not understand why the authors introduces it, is it more than a prediction of their framework that can be tested? If so, I would discuss the potential applications of the RSO in ecology.
- p.2 I would add a little more context about the RSA. In particular, I would introduce another RSA, the zero-sum multinomial that was convincingly applied by Hubbel on tropical trees in its classical Neutral Theory of Biogeography (Hubbell 2001).

- Figure 2: Is the flat line for the Data something expected?
- Table 2: All notations used must be defined. So, C should be defined in the table. Also, I would write "number" rather than using "#" and add a column references to associate an estimator with at least one paper.
- p.12: "*We provide an open source R code that perform*" a "s" is missing.

References

Hubbell, Stephen P. 2001. *The Unified Neutral Theory of Biodiversity and Biogeography*. Monographs in Population Biology 32. Princeton: Princeton University Press.

Tovo, Anna, Samir Suweis, Marco Formentin, Marco Favretti, Igor Volkov, Jayanth R. Banavar, Sandro Azaele, and Amos Maritan. 2017. "Upscaling Species Richness and Abundances in Tropical Forests." *Science Advances* 3 (10): e1701438. https://doi.org/10.1126/sciadv.1701438.