Report on 'Inferring macro-ecological patterns from local species' occurrences' by Tovo et al. - Round 2

Dear authors,

This is my second review of your manuscript 'Inferring macro-ecological patterns from local species' occurrences'. Again, my comments are meant to be constructive, and I hope they will be helpful as you revise your manuscript.

Sincerely,

1 Overall opinion

The authors have taken into account most of my comments and I now have a better understanding of the study. I however think that the manuscript still requires some work, but not much.

2 Negative Binomial and Neutrality

In this section, I explain why I think the authors should better explain what a Negative Binomial (NB) on $\mathbb{R}$ is and the link of their approach with the Neutral Theory.

In their revision letter the authors wrote:

> "*We deem this characterization is somehow misleading for two reasons: i) classically $r \in N$ whereas in our framework $r \in R+$ ii) we recover a negative binomial as the equilibrium distribution of a birth an death process with immigration and we do not see any immediate correspondence between successes and fails of a sequence of binary independent trials and individuals of a species.*"

Thanks to this comment and the new version of the manuscript I now understand what I missed during the first review. The authors are referring to an extended version of the Negative Binomial (NB extended to $\mathbb{R}$). Given that the 3 references about this distribution provided in the text:

> "*Our framework exploits the form-invariance property of the Negative Binomial (NB) distribution. Such a distribution emerges as the long time behavior distribution of a birth and death stochastic dynamics, accounting for effective immigration and/or intraspecific interactions [35, 24, 29].*" (p.3)

include at least one of the authors of this study, I understand that they are quite familiar with this distribution, but for other ecologists it may not be trivial. Actually I have tried to find other usage of this form of NB and found this link https://stats.stackexchange.com/questions/310676/continuous-generalization-of-the-negative-binomial-distribution that suggests that it is not common but used by a few research groups on bioinformatics (McCarthy, Chen, and Smyth 2012). As it is not a classical distribution (the NB on $\mathbb{N}$ is classical), the authors should highlight this as well as the implications (e.g. the need of a normalizing factor in some equation). This is particularly important because some notations are used in a broader sense, for instance below equation (1), the binomial coefficients used are real number, which implies that the Gamma function is used.

Even more important is to mention the link between this approach and the theoretical work on the Neutral Theory of Biogeography (by these authors and others). This **must** be explicitly written in the manuscript. I believe that once this explained the authors should drop the justifications pertaining to the need of the absence of spatial correlation (e.g. "*Under the hypothesis of absence of spatial correlation*" - p.7). This could also be helpful to define the scope of this approach, for instance, p.7:

> The proposed method is, under the 'well mixed' hypothesis, general and not lim- ited to tropical forests.

I agree, but the authors should rather remind the reader that **this approach is limited to systems for which the neutral theory is deemed valid**.

2.1 Major comments

The new version of the manuscript better describe what is done but I still had to go through the method section to see the big picture. I think one sentence explaining that the goal of the approach is to use presence/absence data to infer fundamental parameters that will be used to derive RSA, SAC and RSO would be very helpful. The reader must understand at the end of the introduction what is built on the neutral theory and where is the inference part of the method.

2.1.1 Equation (10) - RSO

Unless I am missing something, equation (10) is wrong. I don't think (10) is a hyper-geometric (it looks like one but it is not). Two options here:

1. I am totally wrong in which case, the authors need to make this part more accessible: I know what a hypergeometric distribution is but I don't understand why it is relevant here;
2. I am right so there is something wrong here. Given the definition of $Q_{occ}(v|n, M, 1)$, I remain skeptical that a hypergeometric is the best option.

2.1.2 M and M*

Overall I do not understand the relationship between M and M* (if any).
One important question is whether or not M=A/a. I think the answer is no but p.7 the authors wrote: "*given that the forest can be tiled in M equal-sized cells of area a.*" which makes me think that it may be yes. Also I think that M=A/a is needed to derive (10) unless n is the total of individuals over A.

2.1.3 Spatial auto-correlation

I think this question is not properly handled in the current manuscript because it relies on the contrast of two point processes and for the one includin spatial auto-correlation (Thomas), we actually don't know what the clustering values means in term of auto-correlation (is it a lot or not?). I guess one way to address this comment is to progressively increase the clustering values and observe how the errors in Table are affected.

2.1.4 Detection

I was wondering whether detection probabilities can be easily integrated in the framework, it may be something to discuss.

2.2 Minor comments

- p.2 "this method have been" => "this method has been"
- p.4 "biodiversity patterns" => "biodiversity relationships"?
- Between equations (2) and (3) I would mention that the goal is to have a relationship between the birth and death ratios at two spatial scales (ξ_p and ξ_{p*}).
- equation (4): what does $\equiv U(p|p^*, \xi p^*)$ means?
- p.7 "*This assumption is not essential to our approach*" you mean the assumption of equal area, right? This is rather important to compute the RSO, am I wrong?
- p.6 I would remind the reader that it needs to use the equation (not numbered) above equation (2) in order to get (7).

References

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): 4288–97. https://doi.org/10.1093/nar/gks042.