

## CAN WE INFER SPECIES INTERACTIONS FROM CO-OCCURRENCE PATTERNS? A REPLY TO PETERSON *ET AL.* (2020)

CHRISTOPHER R. STEPHENS<sup>1,2</sup>, CONSTANTINO GONZÁLEZ-SALAZAR<sup>1,3</sup>,  
MARÍA DEL CARMEN VILLALOBOS-SEGURA<sup>4</sup> and PABLO A. MARQUET<sup>1,5</sup>

<sup>1</sup>*C3 - Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico.* <sup>2</sup>*Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico City, Mexico; stephens@nucleares.unam.mx.* <sup>3</sup>*Departamento de Ciencias Ambientales, CBS Universidad Autónoma Metropolitana, Unidad Lerma; Estado de México, Mexico; cgsalazar7@gmail.com.* <sup>4</sup>*Laboratorio Ecología de Enfermedades y Una Salud, Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, Mexico City, Mexico.* <sup>5</sup>*Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile; Instituto de Ecología y Biodiversidad (IEB), Santiago, Chile and The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 8731, USA.*

Our response to Peterson *et al.*'s (2020) polemic can be summarised quite succinctly thus: **Co-occurrence is a necessary but not sufficient condition for a biotic interaction to occur.**

Unfortunately, Peterson *et al.*'s criticism is geared towards convincing the reader that we are pushing the thesis that “co-occurrence = biotic interaction,” i.e., that co-occurrence is a sufficient condition for a biotic interaction.

Peterson *et al.* have misrepresented our “body of work” in several important ways. Firstly, it has been reduced to a discussion of only “co-occurrence networks” and, moreover, to a particular (over) interpretation of our results. Our “body of work” is quite general—a Bayesian statistical inference framework that uses spatial data to calculate conditional probabilities,  $P(C|X)$ , from a definition of co-occurrences of  $C$  and  $X$  which represent variable(s) whose distribution in space and/or time is known from empirical data. Its chief innovation is that it allows variables of any spatio-temporal resolution to be included, compared, contrasted and combined. The conditional probabilities  $P(C|X)$  can then be used as building blocks in the construction of models that represent two distinct and important elements: ecological niches and communities.

Peterson *et al.* pose the question of whether a connection between co-occurrence and biotic interaction should be expected to exist, arguing that: “Quite generally, spatial co-incidence of species’ distributions may be a consequence of geographic constraint, history, shared climate or substrate preferences, migra-

tory patterns, or many other factors.” Of course! Our methodology allows for all these multiple factors to be included, compared and contrasted in their contribution to the niche and distribution of a species. The only requirement is that a spatio-temporal data representation of the corresponding variables be at hand. Moreover, it allows one to study confounding between one type of factor and another.

Peterson *et al.* also state that: “patterns of co-incidence and non-co-incidence are no indication of the processes causing them,” citing (Bell, 2005), who concentrates on community assembly, showing that: “bulk properties of communities, such as overall diversity, may be strongly affected by local dispersal and stochastic drift and, for this reason, are adequately represented by neutral models in many cases, despite the operation of systematic processes of local selection.” The emphasis here is that it is “bulk” properties that are unlikely to reflect processes. This is why it has been difficult to find systems where global community indices show significantly non-neutral signatures. However, that a checkerboard matrix does not produce a significant overall score does not mean there are no significant correlations between a subset of species within the matrix. Peterson *et al.* further claim that we have tested the predictive capacity of our models in only one single case. However, our methodology has been tested in multiple scenarios, not just a “single test,” including tests where intensive experimental protocols including large field-work and laboratory teams were combined to obtain data that provided a true

out-of-sample test of our methods (Berzunza et al., 2015; Stephens et al., 2016). They also state that our result “needs to be examined critically, using a properly constructed battery of tests (Gotelli 2000).” We have emphasised that our methodology already uses a “properly constructed battery of tests” as the null hypothesis in our binomial test is just SIM2 of Gotelli (2000).

Beyond questioning whether one could possibly expect to see a relation between co-occurrence data and biotic relationships in the first place, Peterson *et al.* question the use of “primary” data sources in light of their well-known biases. As Peterson *et al.* in multiple papers over a long period of time have used such data bases for niche modelling and species distributions we presume that they do not believe that such databases are useless. Moreover, in Atauchi et al. (2018), Peterson and collaborators used “massively biased” primary data, such as GBIF, to model the Peruvian bird species *Phytotoma raimondii*, using as inputs five known food source species, to determine if these biotic factors played a role in the niche of the bird species thereby being inconsistent with the Eltonian Noise Hypothesis (ENH). The authors concluded that: “We rejected the ENH for this case: *biotic interactions* improved the model.” Our point here is not to argue that primary data sources are without problems but, rather, the inconsistency of Peterson *et al.*’s posture that such data cannot be used as a proxy representation of biotic interactions in the case of our work but can in theirs. However, the inconsistency of their argument is much less important than whether or not primary data sources can be used to create predictive models whose results can be validated with known data or new data as we have done.

Peterson *et al.*’s paper mentions several criticisms of our diagnostic  $\epsilon$ : that it is asymmetric, is grid size dependent, can have divisions by zero and that we misinterpret its significance values. We emphasise that  $\epsilon$  is just a binomial test, an extremely well known and studied statistic. The properties that Peterson *et al.* question are all well-known attributes of this diagnostic. Asymmetry, for example, is a direct consequence of the fact that the proportion of occurrences of  $X$  that have a co-occurrence with  $C$  is not the same as the proportion of occurrences of  $C$  that have a co-occurrence of  $X$ . Similarly, as grid size determines the size of the statistical ensemble under consideration, and  $\epsilon$  is a measure of statistical significance, it is perfectly correct that it should de-

pend on the grid size, while the fact that one can have divisions by zero is due to the fact that we are trying to approximate a smooth distribution with a finite sample. Thus, it is sometimes appropriate to smooth the approximate distribution using, for example, Additive or Laplace smoothing. The fourth criticism is our claim that associations are significant when  $\epsilon > 1.96$ . Here, Peterson *et al.* seem to be under the misapprehension that it is the normality of the distribution of values of  $\epsilon$  itself that are being discussed. The distribution of a statistic and the probability distribution of the data from which the statistic is derived, however, are not the same thing.

The next line of criticism is that “the entire probabilistic argument behind the  $\epsilon$  index very doubtful” as “the numbers in that formula, in general, cannot be regarded as probabilities, but as proportions of observations, for a given species, or proportions of observed co-occurrences, for pairs of species, in a particular database.” Of course, the numbers in that formula refer to proportions - that is how a probability is defined in the frequentist perspective - of observations in a particular database, but not, contrary to Peterson *et al.*’s comments, to the “true” probability representing the “true” species distribution in their terms. Obviously, the probability that is calculated from a database represents an under sampling relative to the true distribution.

All Peterson *et al.*’s worked examples are based on the incorrect assumption that our work states that statistically significant co-occurrence is a *sufficient* condition for a biotic interaction. As the “labels” associated with the possibly interacting species are important to understand the nature of any interaction, it is a rather fruitless enterprise to actively seek examples where the associated labels do not naturally lead towards a hypothesis of a likely biotic interaction, as in the case of Peterson *et al.*’s example using the families Trogonidae (Aves) and Scarabeidae (Insecta), where they claim that there can be no biotic component to the interaction as scarabs are terrestrial and trogons are frugivorous. However, contrary to their statements, not all scarabs are terrestrial (Vulinec et al., 2007) and trogons are not exclusively frugivorous (Remsen et al., 1993). Thus, some possible biotic interactions are: some trogon species consume some scarab species, or some fruit consuming scarabs (Reyes and Morón, 2005) consume the same food source as some trogons. This does not mean,

of course, that co-occurrence is due exclusively to these potential biotic interactions.

The second example considers the explanation of the significant co-occurrence of two rodents *Dipodomys merriami* and *Perognathus longimembris*. Peterson *et al.* wish to infer, wrongly, that a large positive value of  $\epsilon$  signifies within our methodology “mutualism, symbiosis” while, in contrast, there is evidence of competition between these two species which comes from controlled experiments where one set of species were excluded from a given area and the relative numbers of any remnant species were recorded. The observation that the removal of some species led to increases in the abundance of some others was interpreted as evidence of competition, although this might not be the case, as shown by Davidson *et al.* (1984). This, micro-level competition however, will not necessarily lead to manifestations at a macro level, where, on the contrary, one might expect to see one species being a niche variable for the other if, for example, they share food resources. One can investigate this phenomenon by creating niche and community models using both climate and rodent species as covariates and noting that the interaction is not explainable by both species being adapted to the same climate. It is then an interesting exercise to construct the niche models and networks for both species by including in potential food sources. One finds for example, that *Larrea tridentata* (creosotebush), *Prosopis glandulosa* (honey mesquite), *Gutierrezia californica* and *Gutierrezia ramulosa* all have significant  $\epsilon$  values with both species. Thus, one can verify that one reason, among various, for a positive interaction may be due to shared food resources. The final worked example of Peterson *et al.* concerns calculating  $\epsilon$  values for six cat species occurring in Mexico but from two distinct data sources. The criticism there is that  $\epsilon$  is “database dependent,” as the results they derive from the SNIB and the IUCN extent-of occurrence datasets (IUCN, 2016) are different.” We completely agree, and so it should be. IUCN data are not primary data: it is the result of a subjective model based on expert opinion, so to call them two different data bases as if they were two different representations of the same thing, both using primary data, is extremely misleading.

In their discussion of our work in epidemiology and public health, Peterson *et al.* state that our “key assertion is that geographic co-occurrence implies biotic interactions such as vectoring and host-

ing pathogens.” No! Our key assertion, again, is that geographic co-occurrence is a necessary but not sufficient condition for such biotic interactions to be present. Thus, it is not a “rash” conclusion to state that a *Leishmania* vector must co-occur with one or more *Leishmania* hosts and vice versa and, contrary to their claim, we have never asserted that all *Lutzomyia* species are vectors, or that all are of the same competence.

In the case of flavivirus, Peterson *et al.* state: “Despite known ZIKV infections in numerous species of bats in Africa and Asia, they have not been found to be competent reservoirs, *contra* the predictions of González-Salazar *et al.* (2017).” This is a false statement. In that paper, we state clearly: “Finally, we emphasise again that the scope of the model of the present paper is to serve as a focus for future studies and show that potentially useful information can be gleaned from the method, which, at this level, is not capable of predicting detailed elements such as potential host competency.” In particular, our methodology is clearly predictive of contact between pathogen and potential mammalian hosts. The question is more, what is the result of that contact? Given the enormous importance of emerging and re-emerging diseases, anything that can focus attention on potential risk factors is extremely valuable.

In closing, we note that network reconstruction based on co-occurrence of living entities or their activities in space and time is becoming a major research direction in the life sciences, from biochemistry, to genomics, to ecology. The reason for this is because it is a fundamental topic, which will provide important clues on phenomena such as co-existence and ecosystem functioning. Several techniques have been developed, such as MaxEnt methods, Boolean networks, and Bayesian approaches. As most research directions and theories in ecology, they proceed by continuous refinement and approximation of complex natural systems. We do not claim that ours is the best approach; we limit ourselves to present it and show that it makes sensible predictions that can be tested, and we acknowledge that, as with most approaches, it can be refined. Isn’t that what science is about? ... increasing our knowledge of nature by developing theories and models and testing the hypothesis derived from their application in order to refine them?

## REFERENCES

- Atauchi, P. J., Peterson, A. T. and J. Flanagan. 2018. Species distribution models for Peruvian plantcutter improve with consideration of biotic interactions. *Journal of Avian Biology*, 49:e01617.
- Bell, G. 2005. The co-distribution of species in relation to the neutral theory of community ecology. *Ecology* 86:1757-1770
- Berzunza-Cruz M, Rodríguez-Moreno A, Gutiérrez-Granados G, González-Salazar C, Stephens C.R., Hidalgo-Mihart M., et al. 2015. *Leishmania (L.) mexicana* infected bats in Mexico: Novel Potential Reservoirs. *PLoS Neglected Tropical Diseases* 9:1-15.
- Davidson, D. W., R. S. Inouye, and J. H. Brown. 1984. Granivory in a Desert Ecosystem: Experimental Evidence for Indirect Facilitation of Ants by Rodents. *Ecology* 65,:1780-86.
- Gotelli, N.J. 2000. Null model analysis of species co-occurrence patterns. *Ecology* 81:2606-2621
- Remsen Jr, J.V., Hyde, M.A., and A. Chapman. 1993. The diets of Neotropical trogons, motmots, barbets and toucans. *Condor* 95: 178-192.
- Peterson, A. T. Soberón, J., Ramsey, J. R., and L. Osorio-Olvera. 2020. Co-occurrence networks do not support identification of biotic interactions. *Biodiversity Informatics* 17:1-10.
- Reyes, N.E. and M.A. Morón. 2005. Fauna de Coleoptera Melolonthidae y Passalidae de Tzucacab y Conkal, Yucatán, México. *Acta Zoológica Mexicana* 21:15-49.
- Vulinec, K., Mellow, D. J. and C.R.V da Fonseca. 2007. Arboreal foraging height in a common neotropical dung beetle, *Canthon subhyalinus* Harold (Coleoptera: Scarabaeidae). *Coleopterists' Bulletin* 61:75-81.