On the inference of positive and negative interactions and their relation to abundance

Andrew J. Rominger

Why do rare species persist in ecosystems? Rare species seem to be at a disadvantage by pure probabilistic odds[1] and perhaps also from poorly adapted species-environment and species-species interactions[2], though negative density-dependence may help rare species persist[3,4]. The question of rarity and persistence thus remains unresolved. In a recent paper, Calatayuda et al.[5] (CEA) inferred species-species interaction networks from spatially replicated abundance data across many taxa and environments. CEA found that rare species were associated with positive interactions whereas common species were associated with negative interactions, indicating that positive interactions, such as facilitation, may help rare species persist[5]. However, the use of abundance and co-occurrence data to infer species interactions is difficult and often inaccurate[6–8]. This issue arises in no small part because the underlying null models used to infer interactions themselves are known to have type I and II error problems in real world applications[9,10]. Here, I show that the finding of rare species being more associated with positive interactions as found by CEA[5] can be explained by statistical artifacts in the inference of species interactions from abundance data. It would therefore not be supported to assign biological interpretations to these findings until more data can be brought to bear on the subject or interaction types and the persistence of rare species.

Species abundances are not evenly distributed across species[11–13], nor evenly distributed within species across space (often referred to as spatial clustering)[13–16]. These two simple observations account for the result of rare species being associated with positive interactions while common species appear associated with negative interactions. When interaction networks are inferred from spatially replicated abundance data, a null model is used to assess whether patterns of species-species co-occurrences deviate substantially enough from null expectations to suggest a non-random interaction. However, if abundances are driven by factors, both probabilistic or deterministic, other than species-species interactions, these null models may not reveal true interactions, but artifacts of other processes The unevenness of species abundances is ubiquitous and can be accounted for by purely probabilistic processes from neutral birth-death-immigration[17] to mechanistically agnostic statistical-mechanical properties of large assemblages[13], thus the simple observation of uneven abundances does by itself indicate deterministic mechanisms.

The data compiled by CEA[5] indeed conform to the ubiquity of unevenness (Supplementary Figs. 2 and 3). To evaluate whether this drives the results about interaction type and abundance, in Figure 1 I first reproduce key results from CEA's Figure 2(b-c); I then simulate purely random data that match key characteristics of observed data but contain absolutely no species interactions. These random data are simulated as follows:

1) The number of species S, number of sites M, and shape of the best fitting species abundance distribution (SAD) are sampled (with replacement) from the observed data
2) S species abundances $x_i \ldots x_S$ are sampled from the SAD
3) For each x_i, within-species counts are distributed across the M sites according to a spatial species abundance distribution (SSAD) that is either negative binomial (in the case of spatial clustering) or Poisson (in the case of spatial evenness)
4) The resulting simulated site by species matrix is fed through the same pipeline as the observed data (described in CEA[5]) to infer positive and negative interactions.

All analyses are carried out in R[18] and can be fully reproduced by installing the R package accompanying this paper, as detailed in the supplement.

In the case of a Poisson SSAD the one parameter (the mean) is fully specified by the average site-level abundance of a given species. In the case of a negative binomial SSAD, the mean parameter is again specified by the site-level average, but the size or clustering parameter k is not fully specified. To capture the rough features of the data, I sample k from a linear relationship (with noise) between the maximum likelihood estimates of k and the relative abundance of each species.

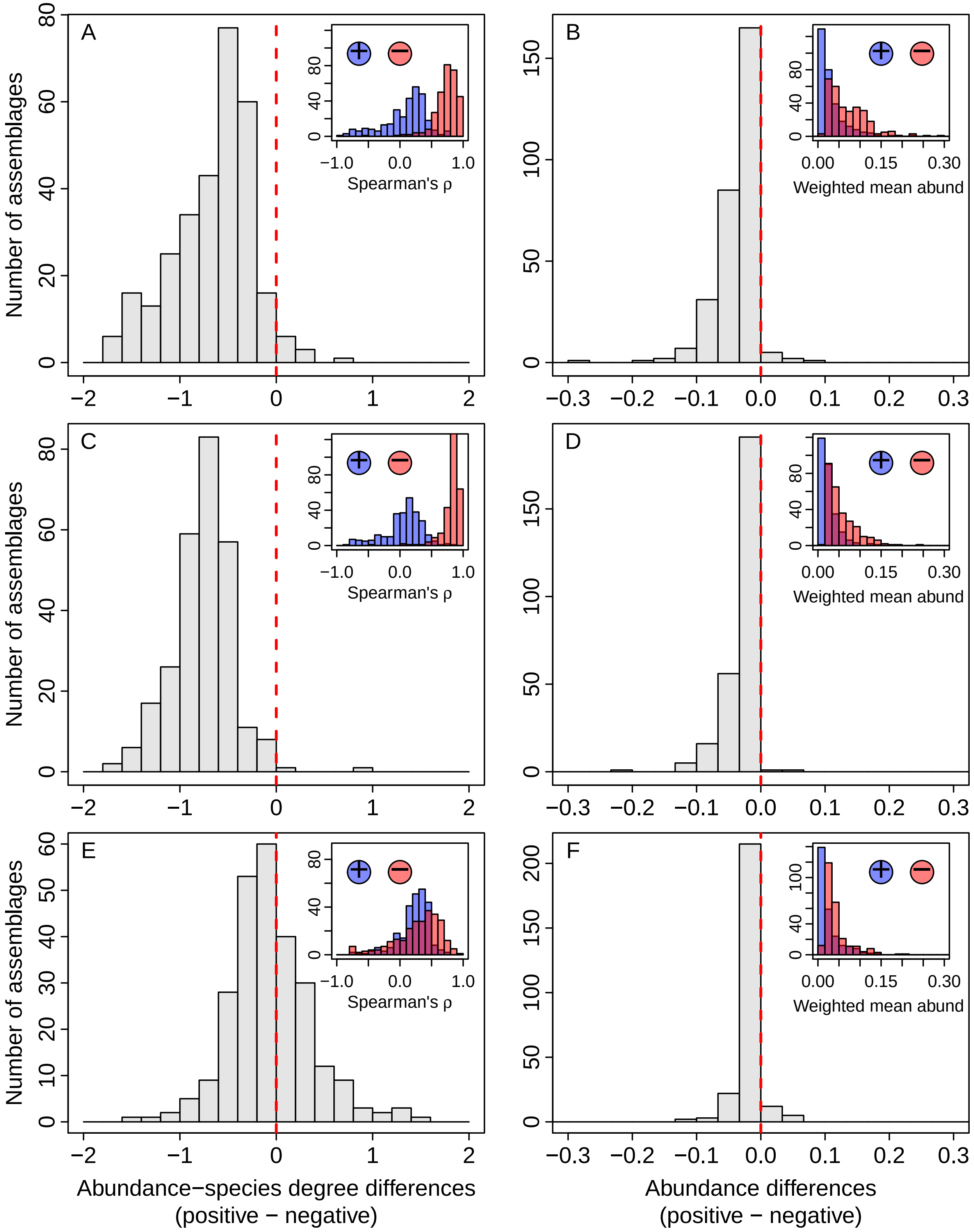

Figure 1: Distributions of correlations between network centrality and abundance (left panels) and weighted mean abundance by network type (right panels). The results of CEA Figure 2(B-C) are reproduced in this figure panels A-B; panels C-D show data simulated with a negative binomial SSAD and no species interactions; panels E-F show data simulated with a Poisson SSAD and again no species interactions.

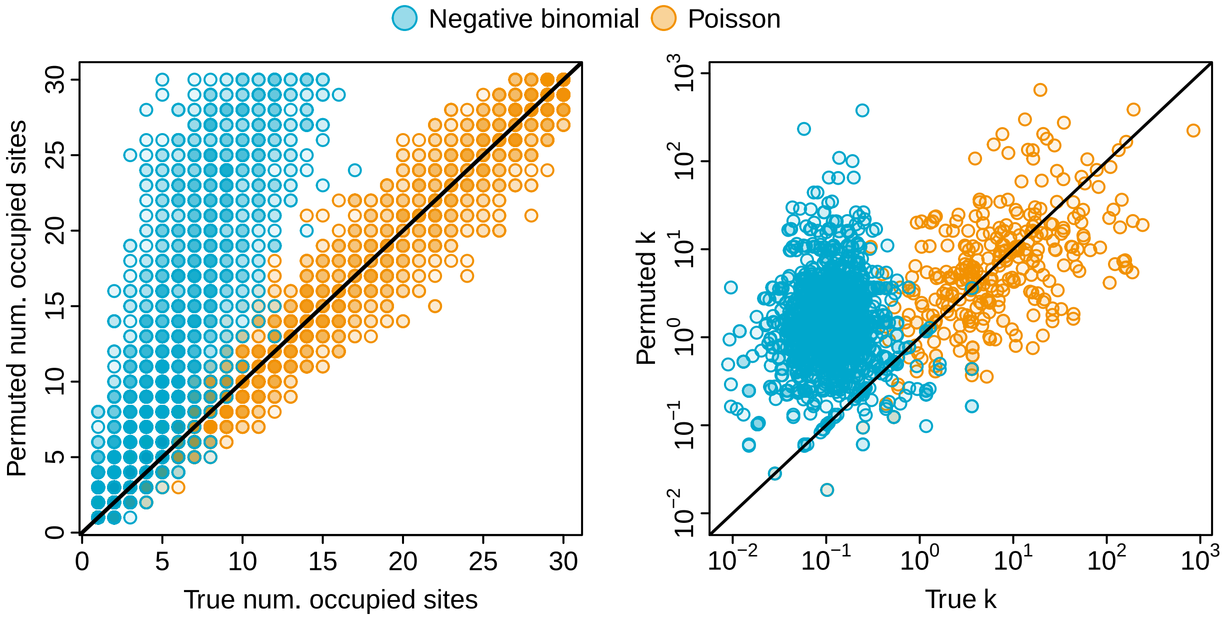

Figure 2: Comparison of true and permuted SSADs in terms of number of sites occupied (A) and inferred clustering parameter k (B). Points are semi-transparent to help display density. Lines are 1:1 lines.

Figure 1 shows that with a negative binomial SSAD, the simulated data closely match the observed findings. This correspondence largely disappears when we instead use a Poisson SSAD, highlighting the importance of spatial aggregation in driving the spurious results. However, we still observe a slight skew toward rare species being slightly more prevalent with positive interactions (Fig. 1F).

These findings do not depend on simulating SAD and SSAD shapes from the data: in Supplementary Figure 4 I show that the spurious relationship between abundance and interaction type occurs even when simulating data from just one arbitrary SAD function with the one arbitrary spatially clustered SSAD for all species. In this simulation, again, replacing the spatially clustered SSAD with a Poisson SSAD breaks the spurious association as in Figure 1.

We seek to understand why negative binomial SSADs reproduce the results while Poisson SSADs fail to. The null model algorithm used here and in CEA[5] fixes row and column marginals, but within any given species, the way its total abundance is allocated across sites by the null model has a potentially large combinatorial space to explore. I compare known SSADs to their permuted counterpart in Figure 2 and find that the null model transforms negative binomial SSADs to a more Poisson shape, while leaving Poisson SSADs probabilistically unchanged. Specifically, when starting with a negative binomial SSAD, the null model inflates the number of sites individuals are allocated to (more similarly to a Poisson SSAD) and consequently the inferred k parameter of these permuted SSADs is increase, indicating less spatial clustering.

At a mathematical level, clustered SSADs, compared to spatially even SSADs, actually increase the probability that rare species will appear aggregated with each other and common species will appear repelled, this difference between clustered and even SSADs explains the results. Consider for example two rare species, one with a single individual and the other with abundance 5, distributed across 5 sites. Their Schoener similarity is maximized when all individuals occur in the same site, such as

$$\begin{bmatrix} 1 & 5 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

If we define $Q(x_i; \mu = 1)$ as the probability of observing x_i individuals in site i given an SSAD with mean parameter μ, then the probability of the above configuration is $Q(5; \mu = 1)Q(0; \mu = 1)^4$. Under a negative binomial SSAD with $k = 0.1$ this equals 4.58×10^{-3} whereas under a Poisson SSAD this equals 5.61×10^{-5}.

Conversely, for two common species, say each with abundance 50, an example configuration that *minimizes* their Schoener similarity would be

$$\begin{bmatrix} 50 & 0 \\ 0 & 50 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

We calculate the probability of any such scenario where no abundances overlap as $4[Q(50; \mu = 10)Q(0; \mu = 10)^4]^2$. With a negative binomial SSAD with $k = 0.1$ this equals 1.41×10^{-7} whereas with a Poisson SSAD this equals 1.61×10^{-72}.

We can contrast this a configuration that would maximize the Schoener similarity between these two common species:

$$\begin{bmatrix} 10 & 10 \\ 10 & 10 \\ 10 & 10 \\ 10 & 10 \\ 10 & 10 \end{bmatrix}$$

The probability of this configuration is $Q(10; \mu = 10)^{10}$ which, for the same negative binomial equals 5.76×10^{-22}, and for the Poisson equals 9.40×10^{-10}.

Thus a clustered SSAD makes rare species appear aggregated, compared to an even SSAD that, if anything, makes common species appear more aggregated. Because the null model algorithm preserves even SSADs this is less of an issue; however, when presented with spatially clustered abundances, the null model will spuriously identify rare species as aggregated and common species as repelling each other. This statistical property accounts for the prevalence of rare species in positive interaction networks and common species in negative interaction networks.

Great caution should be used when inferring species interactions from abundance data. At an even more fundamental level than the spurious association of abundance with interaction type, in the purely random data sets simulated with a negative binomial SSAD, on average 75% of species were placed in positive interaction networks and 74% in negative interaction networks with a significance cutoff of $\alpha = 0.05$. With the Poisson SSAD these simulated numbers were 74% for positive networks and 25% for negative networks. For the observed data, on average 73% of species were placed in positive interaction networks and 60% in negative interaction networks. These are not robust estimates of species interaction.

References

1. McGill, B. J., Hadly, E. A. & Maurer, B. A. Community inertia of quaternary small mammal assemblages in north america. *Proceedings of the National Academy of Sciences* **102**, 16701–16706 (2005).

2. Hutchinson, G. E. The paradox of the plankton. *The American Naturalist* **95**, 137–145 (1961).

3. Leigh Jr, E. G. *et al.* Why do some tropical forests have so many species of trees? *Biotropica* **36**, 447–473 (2004).

4. Yenni, G., Adler, P. B. & Ernest, S. M. Strong self-limitation promotes the persistence of rare species. *Ecology* **93**, 456–461 (2012).

5. Calatayud, J. *et al.* Positive associations among rare species and their persistence in ecological assemblages. *Nat Ecol Evol* (2019).

6. Freilich, M. A., Wieters, E., Broitman, B. R., Marquet, P. A. & Navarrete, S. A. Species co-occurrence networks: Can they reveal trophic and non-trophic interactions in ecological communities? *Ecology* **99**, 690–699 (2018).

7. Carr, A., Diener, C., Baliga, N. S. & Gibbons, S. M. Use and abuse of correlation analyses in microbial ecology. *The ISME journal* **13**, 2647–2655 (2019).

8. Rajala, T., Olhede, S. C. & Murrell, D. J. When do we have the power to detect biological interactions in spatial point patterns? *Journal of Ecology* **107**, 711–721 (2019).

9. Ulrich, W. & Gotelli, N. J. Null model analysis of species associations using abundance data. *Ecology* **91**, 3384–3397 (2010).

10. Ladau, J. Validation of null model tests using neyman–Pearson hypothesis testing theory. *Theoretical Ecology* **1**, 241–248 (2008).

11. Hubbell, S. P. *The unified neutral theory of biodiversity and biogeography.* (Princeton University Press, 2001).

12. McGill, B. J. Towards a unification of unified theories of biodiversity. *Ecology letters* **13**, 627–642 (2010).

13. Harte, J. *The maximum entropy theory of ecology.* (Oxford University Press, 2011).

14. Engen, S., Lande, R. & Sæther, B.-E. A general model for analyzing taylor's spatial scaling laws. *Ecology* **89**, 2612–2622 (2008).

15. Zillio, T. & He, F. Modeling spatial aggregation of finite populations. *Ecology* **91**, 3698–3706 (2010).

16. Connolly, S. R., Hughes, T. P. & Bellwood, D. R. A unified model explains commonness and rarity on coral reefs. *Ecology letters* **20**, 477–486 (2017).

17. Kendall, D. G. Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methodological)* **11**, 230–282 (1949).

18. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, 2018).