On the inference of positive and negative species associations and their relation to abundance

Abstract

The prevalence of rare species in ecosystems begs the question of how they persist. In a recent paper, Calatayuda et al. (CEA) provided a new hypothesis that rare species, in contrast to common species, share unique microhabitats and/or preferentially engage in mutualistic interactions. CEA support this hypotheses by reconstructing association networks from spatially replicated abundance data finding that rare species are over-representing in positive association networks while common species are over-representing in negative association networks. However, the use of abundance and co-occurrence data to infer true species associations is difficult and often inaccurate. Here, I show that the finding of rare species being more represented in positive association networks can be explained by statistical artifacts in the inference of species associations from abundance data. I caution against the inference of ecological association networks from abundance data alone.

Why do rare species persist in ecosystems? Rare species seem to be at a disadvantage by pure probabilistic odds (McGill *et al.*, 2005) and perhaps also from poorly adapted species-environment and species-species interactions (Hutchinson, 1961), though negative density-dependence may help buoy rare species (Leigh Jr *et al.*, 2004; Yenni *et al.*, 2012). The question of rarity and persistence thus remains unresolved. In a recent paper, Calatayud *et al.* (2019) (CEA) have contributed toward helping resolve this question. They compiled an impressive collection of datasets, across many taxa and environments, capturing spatially replicated species abundance measures. With these data they inferred species-species association networks. Such association networks are hypothesized to reflect both potential species-species interactions and/or shared environmental preferences, though there is debate about their accuracy and interpretation (Sander *et al.*, 2017; Barner *et al.*, 2018; Freilich *et al.*, 2018; Carr *et al.*, 2019; Rajala *et al.*, 2019; Blanchet *et al.*, 2020). CEA found that rare species were statistically over-represented in positive-positive species association networks, while common species were statistically over-represented in negative-negative species association networks (Calatayud *et al.*, 2019). CEA interpreted this finding as possible evidence that the persistence of rare species may be aided by positive species interactions, such as mutualism or facilitation, or by shared use of similar microhabitats. However, this result could be compromised by the unreliability of inferring species associations from abundance data alone. Here, I show that the correlation between abundance and association type (positive or negative) as reported by CEA can be explained by statistical artifacts. These artifacts arise because of spatial clustering in intra-specific abundances. It would therefore not be supported to assign biological interpretations to correlations between association types and abundances until more data can be brought to bear on the subject.

When association networks are inferred from spatially replicated abundance data, species-species co-occurrences are quantified by a metric (e.g., CEA use Schoener similarity (Schoener, 1968)) and then a null model is used to assess whether these co-occurrence metrics deviate substantially enough from null expectations to suggest a non-random association, either in the positive or negative direction. However, seemingly non-random patterns in abundance can arise from many processes, including neutrality, that are not driven by species interactions or associations. As such, deviations of abundance patterns from null models might not, by itself, indicate true associations or interactions. One critical, and widely observed, property of species abundances is that they are not evenly distributed across species nor across space within a given species (often referred to as spatial clustering) (McGill & Collins, 2003; Engen *et al.*, 2008; Zillio & He, 2010; Harte, 2011; Connolly *et al.*, 2017). Both ubiquitous patterns can be accounted for by purely probabilistic processes from neutral birth-death-immigration (Kendall, 1949; Hubbell, 2001) to mechanistically agnostic statistical-mechanical properties of large assemblages (Harte, 2011). Importantly, the negative binomial probability distribution both accurately reflects many empirical measurements of spatial variation in intra-specific abundances (Harte, 2011; Connolly *et al.*, 2017) and is independently derived by disparate null or neutral ecological theories (Kendall, 1949; Engen *et al.*, 2008; Harte, 2011).

Thus, the simple observation of uneven or clustered intra-specific abundances does by itself indicate the influence of deterministic species associations. The data compiled by CEA (Calatayud *et al.*, 2019) indeed confirm the ubiquity of uneven species abundances both at an intra-specific level across space (Supplementary Fig. 5) at an inter-specific level (Supplementary Fig. 6). For consistency, I will refer to the spatial distribution of intra-specific abundances as the spatial species abundance distribution (SSAD) following Harte (Harte, 2011) and the inter-specific distribution of abundances as the species abundance distribution SAD. To reiterate for clarity, the SSAD is a measure of spatial variability in *intra*-specific abundances across space and is measured once for each species; the SAD is a measure of variability in abundance across species (i.e. *inter*-specific abundances).

Using simulation, I show that intra-specific spatial clustering of the SSAD alone is sufficient to reproduce the apparent correlation between abundance and association type reported by CEA. Spatial clustering is not, by itself, evidence that rare species positively associate with each other while common species negatively associate. Therefore, the observations reported by CEA do not tell us about species associations, but rather that the null models used do not preserve important aspects of the SSAD.

In Figure 1 I first reproduce key results from CEA's Figure 2(B-C). Then to evaluate whether these results can be produced simply from spatial clustering alone I simulate purely random data (with absolutely no association or interaction between species) that match the unevenness of abundances found in the observed data. These random data are simulated as follows:

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

All analyses are carried out in R (R Core Team, 2018) and can be fully reproduced by installing the R package (https://github.com/ajrominger/RarePlusComMinus) 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 (Supplementary Fig. 5).

Figure 1 A–D shows that with a negative binomial SSAD, simulated data closely match observed findings: the correlation between abundance and species' network degree skews more negative in positive association networks (i.e. more rare species are more highly connected in positive association networks), and positive associations networks tend to contain more rare species than negative networks. This correspondence between real and simulated patterns largely disappears when we instead use a Poisson SSAD, highlighting the importance of spatial aggregation in driving the spurious results.

My findings do not depend on simulating SAD and SSAD shapes from the data: in Supplementary Figure 7 I show that the spurious relationship between abundance and association type occurs even when simulating data from just one arbitrary SAD function with the one arbitrary spatially clustered SSAD for all species.

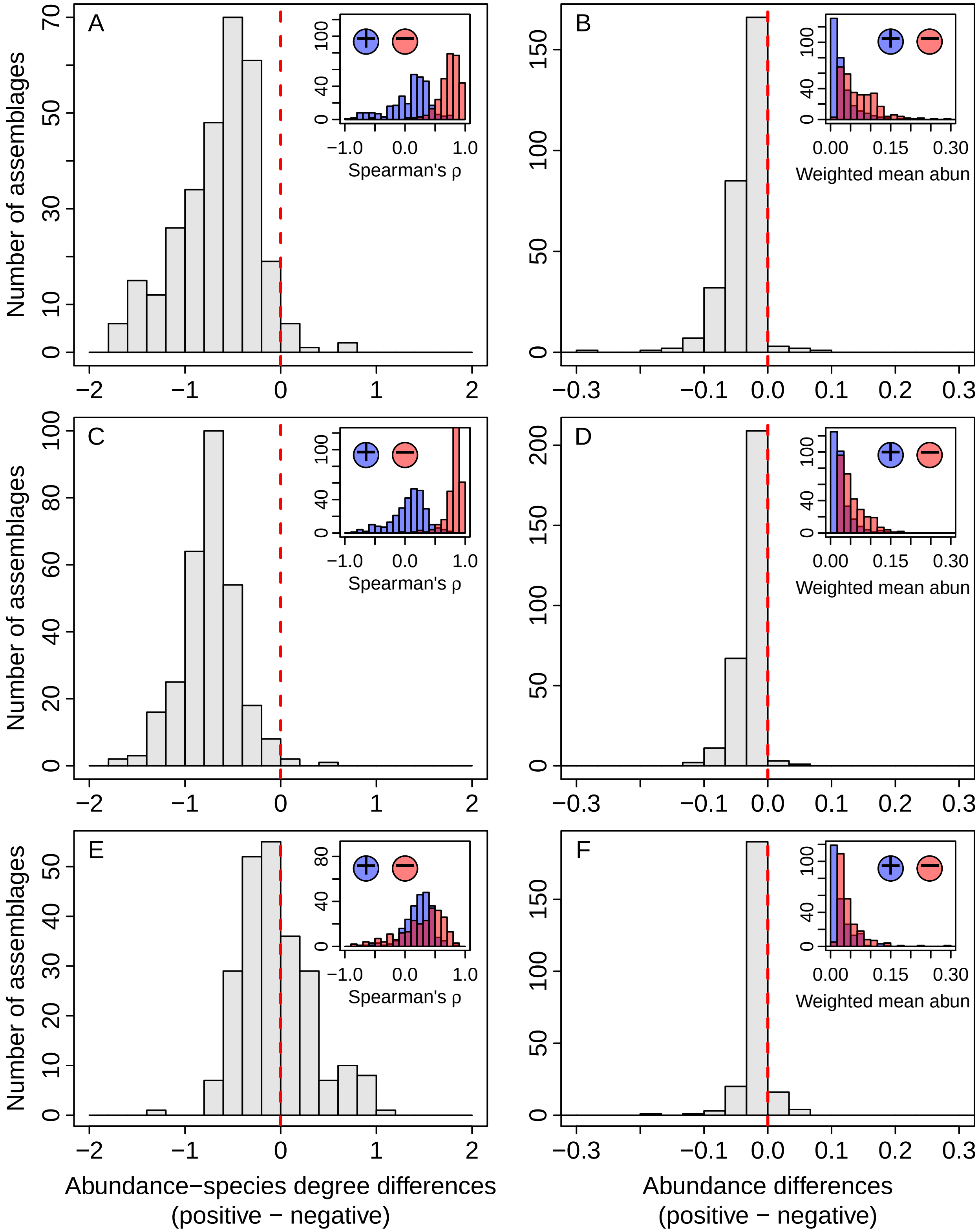

Figure 1: Distributions of correlations between network centrality (i.e. species degree) and abundance (left panels) and distributions of weighted mean abundances (right panels). The main figures show the differences between positive and negative association networks, while the inset figures show the sepparate distributions for each network. The results of CEA Figure 2(B-C) are reproduced here in panels A-B; panels C-D show data simulated with a negative binomial SSAD and no species associations; panels E-F show data simulated with a Poisson SSAD and no species associations.

In this simulation, again, replacing the spatially clustered SSAD with a Poisson SSAD breaks the spurious connection between abundance and association type as in Figure 1 (E-F).

Why do negative binomial SSADs reproduce the results while Poisson SSADs fail to? The null model algorithm used here and in CEA fixes row and column marginals, thus the empirical shape of the SAD is preserved, and the *total* abundances (across all species) at each site are also preserved. However, the way a species' total abundance is allocated across sites by the null model has a potentially large combinatorial space to explore. In Figure 2 I compare summary statistics of known SSADs to their permuted counterparts 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 increases the inferred k parameter, indicating less spatial clustering in the permuted matrices compared to their non-permuted, negative binomial starting points.

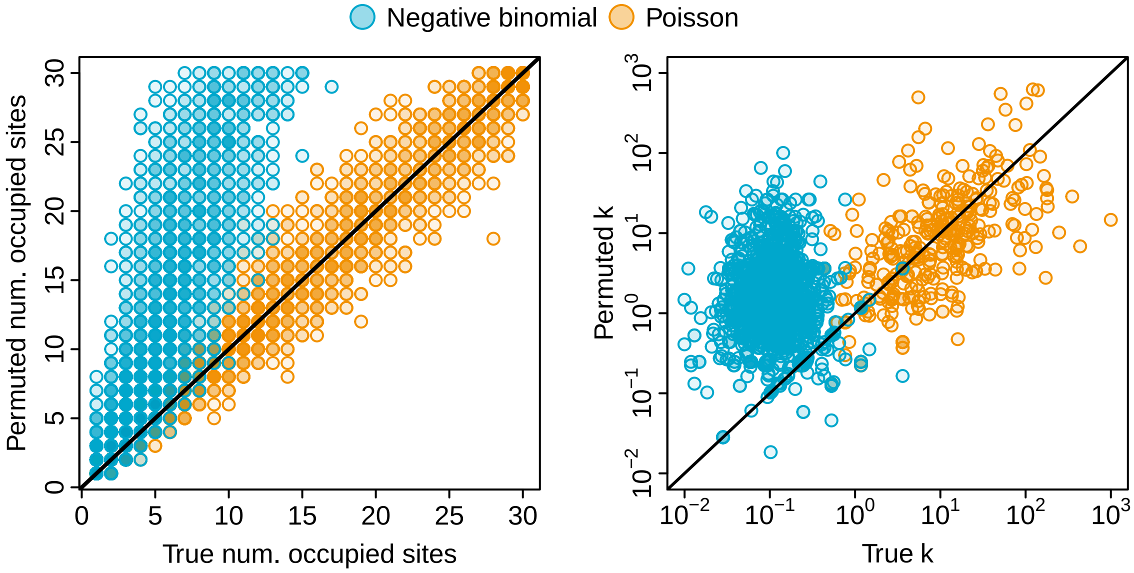

Figure 2: Comparison of SSAD statistics for true and permuted site by species matrices. Colors correspond to the true, un-permuted SSAD. Panel (A) shows how permutation affects number of occupied sites. Panel (B) shows how permutation affects maximum likelihood estimates of the clustering parameter k (B). Points are semi-transparent to help display density. Lines are 1:1 lines.

The negative binomial SSAD appears to be the key to producing presumably spurious relationships between abundance and positive or negative association networks. One might then expect that a null model which preserves the shape of the SSAD for each species would account for statistical artifacts deriving from the SSAD. CEA indeed explore such a null model algorithm (the "independent swap algorithm" (Kembel *et al.*, 2010; Ulrich & Gotelli, 2010); null model III in the CEA supplement) and find it still supports their results. I similarly apply the independent swap algorithm to data simulated with a negative binomial SSAD and no real species associations or interactions. I find that the same spurious relationships between abundance and association type, even when using the independent swap algorithm (Supp. Fig. 8). This again confirms that such association networks and further biological interpretations of them cannot be drawn from abundance data alone.

At a mathematical level, clustered SSADs as compared to spatially even SSADs, increase the probability that rare species will appear positively associated with each other and common species will appear negatively associated. 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 at the same site, such as this site by species matrix

$$X_{rare} = \begin{bmatrix} 1 & 5 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

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 $P(X_{rare}) = Q(5; \mu = 1) \left(Q(0; \mu = 1)^4 \right)$. Under a negative binomial SSAD with $k = 0.1$, $P(X_{rare}) = 4.58 \times 10^{-3}$ whereas under a Poisson SSAD $P(X_{rare}) = 5.61 \times 10^{-5}$.

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

$$Y_{min} = \begin{bmatrix} 50 & 0 \\ 0 & 50 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

We calculate the probability of any such scenario where no abundances overlap as $P(Y_{min}) = 4\left(\left(Q(50; \mu = 10) Q(0; \mu = 10)^4 \right)^2 \right)$. With a negative binomial SSAD with $k = 0.1$, $P(Y_{min}) = 1.41 \times 10^{-7}$ whereas with a Poisson SSAD $P(Y_{min}) = 1.61 \times 10^{-72}$.

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

$$Y_{max} = \begin{bmatrix} 10 & 10 \\ 10 & 10 \\ 10 & 10 \\ 10 & 10 \\ 10 & 10 \end{bmatrix}$$

The probability of this configuration is $P(Y_{max}) = Q(10; \mu = 10)^{10}$. For the same negative binomial $P(Y_{max}) = 5.76 \times 10^{-22}$, and for the Poisson $P(Y_{max}) = 9.40 \times 10^{-10}$.

Thus a spatially clustered SSAD, compared to a spatially even SSAD, gives more probability to configurations where rare species appear aggregated and common species appear over-dispersed. Because the null model algorithm permutes site by species matrices to resemble more Poisson-like SSADs this probabilistic difference between spatially clustered versus even SSADs accounts for the prevalence of rare species in positive association networks and common species in negative association networks.

Caution should be used when inferring species association from abundance data. More fundamentally than the spurious correlation of abundance with association type, my analysis shows that statistically significant species associations are inferred from data simulated without any real species associations. In data simulated with a negative binomial SSAD, on average 75% of species were placed in positive association networks and 75% in negative association networks with a significance cutoff of $\alpha = 0.05$. With the Poisson SSAD these simulated numbers were 72% for positive networks and 25% for negative networks. For the observed data, on average 73% of species were placed in positive association networks and 60% in negative association networks.

It is becoming increasingly appreciated that abundance data alone are not sufficient to distinguish between different ecological processes (McGill *et al.*, 2007; Morlon *et al.*, 2009). The question of why rare species persist is fascinating, and CEA should be commended for making a concerted effort to illuminate possible mechanisms underlying the phenomenon; however, to reach robust conclusions, other types of data, such as actual experimental measurement of shared environmental associations or species-species interaction strengths, are needed in addition to abundance data.

Data and Code Availability

All data and code needed to reproduce the results of this manuscript are available at https://github.com/ajrominger/RarePlusComMinus and a detailed description of the analytical approach is available in the supplement.

References

Barner, A.K., Coblentz, K.E., Hacker, S.D. & Menge, B.A. (2018) Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology*, **99**, 557–566.

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020) Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, **23**, 1050–1063.

Calatayud, J., Andivia, E., Escudero, A., Melián, C.J., Bernardo-Madrid, R., Stoffel, M., Aponte, C., Medina, N.G., Molina-Venegas, R., Arnan, X., Rosvall, M., Neuman, M., Noriega, J.A., Alves-Martins, F., Draper, I., Luzuriaga, A., Ballesteros-Cánovas, J.A., Morales-Molino, C., Ferrandis, P., Herrero, A., Pataro, L., Juen, L., Cea, A. & Madrigal-González, J. (2019) Positive associations among rare species and their persistence in ecological assemblages. *Nat Ecol Evol*.

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

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

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

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

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

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

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

Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., Blomberg, S.P. & Webb, C.O. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**, 1463–1464.

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

Leigh Jr, E.G., Davidar, P., Dick, C.W., Terborgh, J., Puyravaud, J.-P., Steege, H. ter & Wright, S.J. (2004) Why do some tropical forests have so many species of trees? *Biotropica*, **36**, 447–473.

McGill, B. & Collins, C. (2003) A unified theory for macroecology based on spatial patterns of abundance. *Evolutionary Ecology Research*, **5**, 469–492.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F. & others (2007) Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology letters*, **10**, 995–1015.

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

Morlon, H., White, E.P., Etienne, R.S., Green, J.L., Ostling, A., Alonso, D., Enquist, B.J., He, F., Hurlbert, A., Magurran, A.E. & others (2009) Taking species abundance distributions beyond individuals. *Ecology Letters*, **12**, 488–501.

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

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

Sander, E.L., Wootton, J.T. & Allesina, S. (2017) Ecological network inference from long-term presence-absence data. *Scientific reports*, **7**, 7154.

Schoener, T.W. (1968) The anolis lizards of bimini: Resource partitioning in a complex fauna. *Ecology*, **49**, 704–726.

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

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

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