ECOLOGY LETTERS

Ecology Letters, (2020) **23**: 1050–1063 doi: 10.1111/ele.13525

IDEAS AND PERSPECTIVES

Co-occurrence is not evidence of ecological interactions

F. Guillaume Blanchet,[1]* Kevin Cazelles[2] and Dominique Gravel[1]

[1]*Département de biologie, Université de Sherbrooke, Sherbrooke J1K 2R1,QC, Canada*

[2]*Department of Integrative of Biology, University of Guelph, Guelph N1G 2W1,ON, Canada*

**Correspondence: E-mail: guillaume.blanchet@usherbrooke.ca*

The peer review history for this article is available at https://publons.com/publon/10.1111/ele.13525

Abstract

There is a rich amount of information in co-occurrence (presence–absence) data that could be used to understand community assembly. This proposition first envisioned by Forbes (1907) and then Diamond (1975) prompted the development of numerous modelling approaches (e.g. null model analysis, co-occurrence networks and, more recently, joint species distribution models). Both theory and experimental evidence support the idea that ecological interactions may affect co-occurrence, but it remains unclear to what extent the signal of interaction can be captured in observational data. It is now time to step back from the statistical developments and critically assess whether co-occurrence data are really a proxy for ecological interactions. In this paper, we present a series of arguments based on probability, sampling, food web and coexistence theories supporting that significant spatial associations between species (or lack thereof) is a poor proxy for ecological interactions. We discuss appropriate interpretations of co-occurrence, along with potential avenues to extract as much information as possible from such data.

Keywords

Co-occurrence analysis, co-occurrence networks, ecological interactions, presence–absence data, statistical inference.

Ecology Letters (2020) **23**: 1050–1063

INTRODUCTION

Co-occurrence analysis is the study of interactions between species distributions, and as such, it has been at the centre of community ecology for more than 100 years. Throughout this paper, we assumed an interaction occurs when the presence of a species has some influence on the occurrence of another. With the arrival of new statistical methods and the accumulation of observational data, co-occurrence analysis recently attracted a lot of attention from different fields (e.g. ecology and microbiology) and for various systems (e.g. boreal forests and gut microbiome). We believe there is a rich amount of information in co-occurrence data, but its interpretation should be done with care. There are several theoretical and statistical reasons explaining why there is only a weak relationship between co-occurrence and interactions. Here, we review the vast literature on co-occurrence and propose a set of arguments using probability, sampling food web and coexistence theories to support our claim that spatial associations (or lack thereof) between species should not be considered as a proxy for ecological interactions. Finally, we conclude this paper by presented different avenues and outstanding questions that need further investigation to better understand and predict ecological interactions (Box 1).

A RICH AND LONG DEBATE

It is a truism of ecology that species must co-occur to directly interact. It is also a truism of population biology that interactions impact demography, which in turn must affect co-occurrence. This explains why early on ecologists have proposed and discussed statistical methods to infer relationships among species based on presence–absence data (Forbes, 1907; Michael, 1920; Pielou & Pielou, 1967, 1968; Diamond, 1975). As early as 1907, Forbes proposed a systematic analysis of pairwise co-occurrences using the ratio between the number of observed and expected co-occurrences to determine the degree of association among pairs of fishes (Forbes, 1907; Alroy, 2015). Some 13 years later, in a modern 'plea in behalf of quantitative biology', Michael (1920) highlighted several drawbacks of Forbes' coefficient, notably he pointed out the importance of the spatial scale of sampling unit to draw meaningful conclusions about the underlying ecological relationships inferred from it. Hence, Forbes (1907) was likely the first ecologist to quantify ecological processes with an index based on an incidence matrix, whereas Michael (1920) was among the first biologists to point out potential drawbacks of such indices.

Forbes' coefficient was forgotten for years and similar approaches, grounded on the same rationale, have been developed independently (Alroy, 2015; Arita, 2016). In 1967 and 1968, Pielou and Pielou developed two statistical methods to discriminate mechanisms of co-existence among Diptera species on a bracket fungus by determining whether the frequencies of certain assemblages departed from random expectations (Pielou & Pielou, 1967, 1968). A few years later, Diamond (1975) introduced his assembly rules to explain the checkerboard distributions of bird communities on archipelagos. Diamond's assembly rules were quickly challenged by Connor & Simberloff (1979) who criticised the lack of random expectations thereof. This marked the beginning of a still ongoing debate about the link between co-occurrence data and species interactions (Gotelli & McCabe, 2002; Connor *et al.,* 2013; Diamond *et al.,* 2015) and, as a side contribution, generated a number of new techniques aimed at improving the extraction of ecological information from co-occurrence data (e.g. Whittam & Siegel-Causey, 1981).

The current array of methods available can be classified into three different categories. First, the matrix-level approaches aim at determining the main drivers of species' distribution for a given community based on the entire incidence matrix properties (Stone & Roberts, 1990; Gotelli & Ellison, 2002; Sfenthourakis *et al.,* 2006; Cardillo & Meijaard, 2010; Arita *et al.,* 2012; Ulrich & Gotelli, 2013). To do so, one or several indices are computed based on the observation data and compared to random expectations derived from null models. For instance Patterson & Atmar (1986) used nestedness to support the hypothesis that selective extinctions occurred in the mammal community of the southern Rocky Mountains. More recently, following Leibold & Mikkelson (2002), Presley *et al.* (2010) proposed a hierarchical approach based on coherence, species turnover and clumping to characterise the spatial structure of the community and hence determine the role played by colonisation and niche partitioning (D'Amen et al., 2018). The development of these techniques led to more sophisticated null models, and the successful inclusion of environmental variables (Gotelli & Ulrich, 2010) spurred enthusiasm for methods originating from research in species distribution modelling.

The second category originates from developments in species distribution models (SDMs) that predict the geographical distribution of species from abiotic variables (Elith *et al.,* 2006). Indeed SDMs, developed in the 1990s and the early 2000s were criticised for neglecting biotic interactions (Wisz *et al.,* 2013) whereas it has been repeatedly shown that biotic interactions improve the accuracy of predictions (Leathwick & Austin, 2001; Heikkinen *et al.,* 2007; Meier *et al.,* 2010; Leach *et al.,* 2016; Barbaro *et al.,* 2019). Consequently, the now so-called joint species distribution models (JSDMs; Pollock *et al.,* 2014) were developed and predict the distribution of a set of species that are potentially interdependent based on abiotic factors using the entire incidence matrix (Özesmi & Özesmi 1999; Latimer *et al.,* 2009; Ovaskainen *et al.,* 2010, 2016, 2017; Clark *et al.,* 2014; Kaldhusdal *et al.,* 2015; Warton *et al.,* 2015; Hui, 2016; Clark *et al.,* 2017; Staniczenko *et al.,* 2017). In most cases, these models provide individual species responses to the abiotic environment together with a covariance matrix whose elements capture the correlations in the incidence matrix that are not explained by the abiotic factors. Based on its mathematical definition, this matrix has been suggested as a robust way of finding significant association in co-occurrence data while accounting for environment filtering (D'Amen et al., 2018) and hence JSDMs are now used to infer interactions from ecological data (Morales-Castilla *et al.,* 2015; D'Amen et al., 2018; Barner *et al.,* 2018).

The methods in the last category directly infer ecological relationships based on the incidence matrix: for each pair of species, the two vectors of occurrence and an optional set of covariates (e.g. abiotic factors, species abundances) are combined to compute statistical associations (Veech, 2014; Morueta-Holme *et al.,* 2016; Mandakovic et al., 2018). Several techniques have been used to obtain those scores, including Fisher's tests (Veech, 2013; Arita, 2016), odds ratios (Lane *et al.,* 2014), correlations (Steele *et al.,* 2011; Faust & Raes, 2012) and Markov networks (Harris, 2016; Clark *et al.,* 2018; Popovic *et al.,* 2019). In essence, those approaches are close to Forbes (1907) coefficient proposed a century ago (and some are actually very similar, e.g. Veech, 2013; Arita, 2016), but recent approaches are now focusing on the entire set of the significant co-occurrence associations, i.e. the co-occurrence network (Araújo *et al.,* 2011; Tulloch *et al.,* 2016; Kay *et al.,* 2017). Among these methods, a dividing line must be drawn: while some approaches interpret variations in co-occurrence networks as evidence for changes in ecological interactions (Araújo *et al.,* 2011; Tulloch *et al.,* 2016; Kay *et al.,* 2017), others treat them as a direct proxy for interactions (Zelezniak *et al.,* 2015; Harris, 2016).

From the first to the third category of methods, there is a major conceptual shift from the interpretation of significant spatial associations in co-occurrence data as a potential sign of biotic interactions towards the reconstruction of entire ecological networks derived from large presence absence data sets (Faust & Raes, 2012; Wisz *et al.,* 2013; Berry & Widder, 2014; Zelezniak *et al.,* 2015; Mandakovic et al., 2018). Although inferring ecological interactions from the easiest data to acquire (presence–absence data) holds a great appeal, one should bear in mind that this is feasible only if ecological interactions leave a signal in the presence–absence data that is regular enough to be detected and interpreted by adequate statistical methods. While some recent studies have unveiled such a regular signal (e.g. Gotelli *et al.,* 2010; Cardillo, 2011), other have shown that the signal is blurred and diluted in complex networks (Cazelles *et al.,* 2016) or even absent (Brazeau & Schamp, 2019) and thus, the existence of a signal and properties thereof are still debated.

In the past 2 years, no less than four examinations have been proposed of recent statistical approaches used to infer species associations from presence–absence data (Barner *et al.,* 2018; Freilich *et al.,* 2018; Thurman *et al.,* 2019; Brazeau & Schamp, 2019). Those studies focused on specific sets of species that met two criteria: (1) regional scale species presence–absence data were available and (2) biotic interactions among the species considered were documented *a priori*. Using this information, the ability of existing statistical techniques to detect real interactions (covering the three categories described above) were evaluated. Interestingly, these studies reached similar conclusions: current methods are generally inaccurate, and thus, the spatial associations detected are poor proxies for biotic interactions. Even though these papers cast doubts on studies that equate species co-occurrences to species interactions' (Barner *et al.,* 2018), there are two major limits that preclude general conclusions to be drawn from them. First, as these investigations were carried out on specific systems, the reasons behind the poor performances observed might be, in specific or in general, related to the particularities of the system itself. Second, it could be argued that the results obtained merely pinpoint shortcomings in statistical approaches employed that could be addressed by future technical advances. Therefore, there is still a need for a critical examination of the assumptions under which (1) ecological interactions actually leave a signal in presence–absence data and (2) whether it is feasible to detect and interpret this signal properly; this is especially true given the enthusiasm around the promise of detecting interactions from presence–absence data, which may lead to inferences of ecological processes where there are none (Warren *et al.,* 2014). In the following lines, we

propose such an examination and develop seven arguments based on probability, sampling, food web and coexistence theories supporting that significant spatial associations between species (or lack thereof) is a poor proxy for ecological interactions.

INTERPRETATION OF CO-OCCURRENCES USING CONDITIONAL PROBABILITIES

Through the eyeglass of probability theory, the distribution of each species can be understood as a Bernouilli random variable (referred to as X); X = 1 for presence, X = 0 for absence where the probability of occurrence of species A and B are, respectively, $P(X_A)$ and $P(X_B)$ and the probability of the co-occurrence is $P(X_A, X_B)$. This can then be compared to the expectation assuming the two species were occurring independently from each other, i.e. $P(X_A) \times P(X_B)$, often obtained through randomisation (Gotelli & Graves, 1996; Presley *et al.,* 2010; Ulrich & Gotelli, 2013). In this respect, it is common for observations significantly larger or smaller than the random expectation to be interpreted as evidence of an ecological interaction. This is the rationale behind classical approaches such as the C-score (Stone & Roberts, 1990; Gotelli *et al.,* 2010); we refer to this departure from a random expectation as 'co-occurrence signal'.

An interaction is inferred when the presence of a species at a given location has an influence (regardless of its nature) on the probability of observing another species at that same location. This can be presented formally by stating that the conditional probability $P(X_A|X_B = 1)$ is significantly different from $P(X_A|X_B = 0)$ (see section I of the Supplementary Information for further details). This definition of interaction differs from the conventional definition of interactions used in community ecology, which states that an interaction is the effect of a species on the *per capita* growth rate of another one (Berlow *et al.,* 2004). In the following lines, we present arguments explaining in detail why co-occurrences do not imply interactions using the conditional and joint probability formalism.

Argument 1 – Species occurrences depend on the environment

Rationale

Let's assume that the occurrence of species A and B are both conditional on an environmental variable E. In other words, the occurrence probability of A and B varies along an environmental gradient. Assuming that both species do *not* interact, we may still observe a strong signal in their co-occurrence profile due to the similarity (or dissimilarity) in their environmental requirements. Figure 1a illustrates an example of how such a situation occurs in nature (on Mont Mégantic, Canada).

From a mathematical standpoint, this argument is based on the fundamental difference between the probability of co-occurrence of A and B over the entire environmental gradient

$$P(X_A, X_B), \tag{1}$$

and the expected co-occurrence of the two species for a given environmental condition

$$P(X_A, X_B|E). \tag{2}$$

In the context of SDMs, independence among species is assumed, and the general interpretation is that ecological interactions do not influence species distribution (Jeschke & Strayer, 2008). In this respect, independence should be mathematically defined as

$$P(X_A, X_B|E) = P(X_A|E)P(X_B|E), \tag{3}$$

where $P(X_A|E)$ and $P(X_B|E)$ explicitly state that the probability of occurrence of each species is conditional on the environment. Graphically, Fig. 1b and c depicts conceptually how typical deciduous and conifer species co-occur along an elevational gradient. However, the assumption of independence is often treated in the absence of environmental pressure, thus defined as:

$$P(X_A, X_B) = P(X_A)P(X_B). \tag{4}$$

The critical issue here is that eqn (3) does not imply eqn (4) (we explain why in the 'The problem of abiotic factors' section of the Supplementary Information, SI). If interactions are inferred from spatial associations over an environmental gradient, the variation in the probability of presence for one (or both) species along the gradient could generate false positives and more rarely false negatives. We illustrated such a situation in Fig. 1 where we show how the distribution of the environmental values E (Panels d–f) dramatically influences the observed co-occurrence (Panels g-i), even though the species are independent.

Conclusion

This argument suggests that any environmental condition influencing the distribution of two species may cause a strong co-occurrence signal that could be misinterpreted as ecological interactions. Furthermore, the sampling design could lead to different interpretations. Although this argument may suggest that using models that account for environmental filtering is appropriate (e.g. JSDMs, Ovaskainen *et al.,* 2010; Warton *et al.,* 2015; D'Amen et al., 2018), it should not be interpreted this way. Indeed, the co-occurrence signals (e.g. a significant positive or negative correlation value) estimated from these models could originate from any abiotic factors that impact species differently. Therefore, this correlation *cannot* be systematically interpreted as a signal of biotic interactions, as it could rather express potential non-measured environmental drivers (or combinations of them) that influence species distribution and co-distribution.

A potentially interesting way to approach this problem is to use latent variable models (e.g. Warton *et al.,* 2015; Ovaskainen *et al.,* 2017) because latent variables may be able to capture some unmeasured environmental variables. However, no distinctions are made about the type of information captured by latent variables making the use of such technique far from optimal. This difficulty of discriminating between interaction and environment using species distribution data has been shown by Godsoe *et al.* (2017) for simple interactions using simulations.

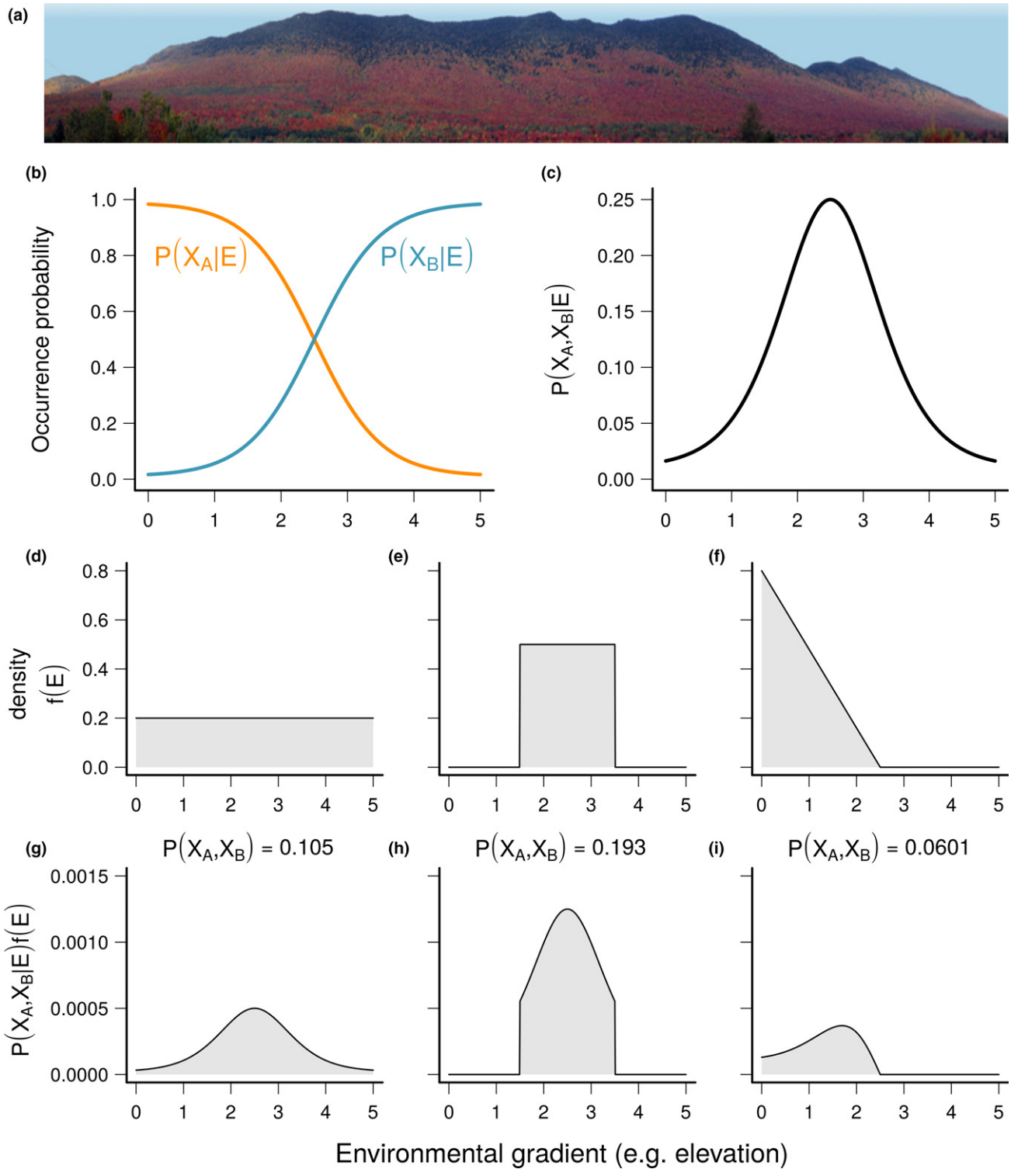

Figure 1 Species co-occurrences may depend on the abiotic environment. (a) Picture of Mont Mégantic taken in Fall showing spatial repulsion between conifers (dark green) and deciduous (red, orange and yellow) trees. The zone where tree species co-occur is caused by the elevation gradient and does not represent interaction among species *per se* (Savage & Vellend,2015). (b) Occurrence probability of species A (orange) and B (blue) along an environmental gradient (abscissa). Assuming the environmental gradient presents the full elevation of Mont Mégantic, A is a typical deciduous species, whereas B is a typical coniferous species, then panel b conceptually depicts an elevation transect of the picture in panel a. (c) Co-occurrence probability of finding both A and B along the elevation gradient. Recall that species A and B are assumed independent and as such this is the conditional probability resulting from eqn (3). (d–f) Three contrasting environmental gradients, i.e. three potential probability density functions for environmental values *E*. (g–i) Product of the scenarios in panels d–f with the conditional probability of co-occurrence presented in panel c for the two species of panel b. The marginal probability of co-occurrence for A and B, obtained through integration over the entire environmental gradient, are indicated at the top of each respective panel.

Argument 2 – The detection of the interaction between two species vanishes if either of these species interact with other species

Rationale

We focus here on the interaction among three species and assume that no other factors (biotic, environmental or others) influence their occurrence. What we show through this argument is that ecological interactions can influence the presence of a species in a specific location in unexpected ways. As species are embedded in complex networks, it becomes problematic to define a specific association without accounting for other ones. Cazelles *et al.* (2016) have already discussed this issue and showed that the higher the degree of a species (i.e. the number of interactions between this species and any other) the weaker is its statistical association with them. In other words, if an interaction between two species exists, the existence of another interaction hampers the detection of the former.

This problem is illustrated in Fig. 2 with an artificial system of three species (a herbivore (H) and two plant species (V1 and V2)). We assumed here that the two plants occur independently and that the conditional co-occurrences of the herbivore with the two plants reflect interaction strengths. Based on these assumptions, we examine how increasing the interaction strength between H and V2 while keeping the strength of the interaction between H and V1 constant affects the perceived relationship between H and V1 (see SI section 'Simulations' for computational details).

In mathematical terms, the problem highlighted in this argument is that $P(X_H)$ depends on both $P(X_{V1})$ and $P(X_{V2})$:

$$P(X_H) = P(X_H|X_{V1})P(X_{V1}) + P(X_H|X_{V2})P(X_{V2}) \quad (5)$$

but the detection of the signal in the co-occurrence data of H and one of the plants, say V1, based on the departure from random expectation: $P(X_H,X_{V1}) - P(X_H)P(X_{V1})$, does not account for the third species. As a consequence, the association profile between H and V1 changes markedly (Fig. 2a–d). This argument also highlights the necessity of having accurate knowledge of the probability of occurrence of all species considered as well as the strength of interaction between H and V1 in the absence of V2, to correctly interpret the values of the association. Even for well-known species, gathering this information can be challenging.

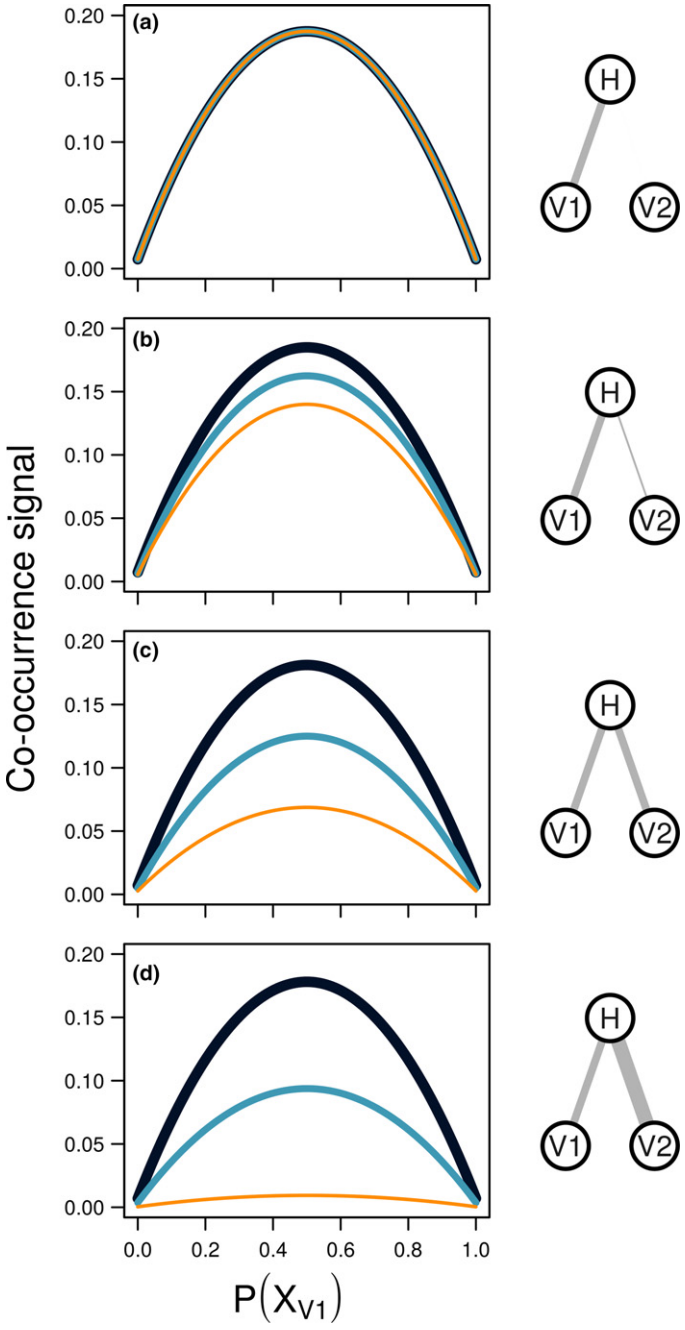

Figure 2 Co-occurrence signal in a three species system including a herbivore (H) and two plant species (V1 and V2). Note that the letter 'V' was used for 'vegetation'. The co-occurrence signal is the departure of the co-occurrence from random expectations, i.e. $P(X_H,X_{V1})-P(X_H)P(X_{V1})$. It is computed along the gradient made by the occurrence probabilities of consumer V1 ($P(X_{V1})$), whereas the occurrence probabilities of consumer V2 remains constant at $P(X_{V2}) = 0.05$ (dark blue), $P(X_{V2}) = 0.5$ (light blue) and $P(X_{V2}) = 0.95$ (orange). In (a) H and V2 are independent and thus $P(X_H|X_{V2}) = P(X_H)P(X_{V2})$, for the three other panels, this probability increases: 0.2 (b), 0.5 (c) and 0.95 (d). The interaction between H and V1 remains constant with a value of 0.75 for all panels.

Conclusion

Even though two species may interact strongly, the corresponding association values may be very low because of the interactions with other species (Cazelles *et al.*, 2016). Recently, Thurman *et al.* (2019) presented empirical results that support this theoretical finding. In their paper, they found that as more species interact, a general weakening of association strengths and trend towards positive associations can be found. As all interactions matter, it thus becomes important to find adequate approaches to characterise independent interactions while controlling for all the other interactions a species may have. A way to overcome this issue would be to keep exploring partial correlations using Bayesian (Staniczenko *et al.*, 2017) and Markov networks (Harris, 2016; Clark *et al.*, 2018). Paradoxically, to benefit from such tools and accurately detect interactions (e.g. to meaningfully use partial correlations), the entire list of interacting species as well as the full topology of the network need to be known beforehand. In a recent study, Popovic *et al.* (2019) proposed

a generalisation of the ideas proposed by Harris (2016) that can combine different data types (e.g. presence–absence, count, biomass, ordinal, etc.) in a single model using Gaussian copula. This is an interesting development because it focuses on studying relations among species using data more informative than presence–absence data. However, the ideas proposed by Popovic *et al.* (2019) and Harris (2016) suffer from the same pitfalls when used on co-occurrence data.

Argument 3 – Species associations could arise indirectly

Rationale

In ecological networks, in addition to direct interactions (Argument 2), indirect interactions may also generate non-random associations among species. For instance if a top predator feeds on a carnivore that in turn feeds on a herbivore, the top predator and the herbivore may co-occur more frequently than expected even though they do not interact directly. Using simulated (co-)occurrences data, Cazelles *et al.* (2016) have shown a decrease of the co-occurrence signal with an increase in the shortest path between two species part of the same ecological network.

In order to illustrate how indirect associations can emerge from a chain of direct interactions, we consider a chain of four trophic levels where each species feeds solely on the one directly below it in the chain (Fig. 3b). Using conditional probabilities and assuming that a predator cannot survive without its prey, we obtain.

$$P(X_H) = P(X_H|X_V)P(X_V) \quad (6)$$

$$P(X_C) = P(X_C|X_H)P(X_H) = P(X_C|X_H)P(X_H|X_V)P(X_V) \quad (7)$$

$$P(X_T) = P(X_T|X_C)P(X_C) = P(X_T|X_C)P(X_C|X_H)P(X_H|X_V)P(X_V). \quad (8)$$

With this example, we examined how increasing the strength of association between H and V affects the co-occurrence signal between H and the other species. In this case, the signal is computed as the difference between the observed co-occurrence and the expected one under the assumption that species are independent (see section 'Simulation' in SI for further details).

Figure 3a shows an increase in the co-occurrence signal for V-C and V-T as the association strength between V and H increases, meaning that the signal propagates through the network. On the other hand, this illustration also shows that the signal weakens along the chain. While the results are direct consequences of the assumption and the equations above, it also points out the difficulty in interpreting the co-occurrence signal without prior knowledge of the network (which we have in our illustration, Fig. 3b). Indeed, the sole examination of the co-occurrence signal would not allow us to determine whether the interactions T-V and C-V are direct but weak, or indirect.

Conclusion

Indirect interactions can generate non-random associations that can be interpreted similarly to the ones resulting from direct interactions. While in some cases, revealing the presence of an interaction, be it direct or indirect, is enough (e.g. to predict species distributions in the case of JSDM), this argument constitutes a major obstacle to the accurate inference of complex networks based on co-occurrence data alone. Also, because it is rarely obvious whether a particular association is direct or indirect from co-occurrence studies, such interactions could be misinterpreted. To further confound us, species may modify their interactions solely in the presence of another particular species. Studies on invasive species are rich in examples of this particularity of nature (Zavaleta *et al.*, 2001, for a review). That being said, graphical models (Popovic *et al.*, 2019) may be an interesting starting point to approach this problem because they were shown to be efficient in capturing direct association among species. A graphical model is a probabilistic model that uses a graph to express the conditional dependencies between different variables. Note that the associations measured by Popovic *et al.* (2019) are not, and have never been considered, interactions.

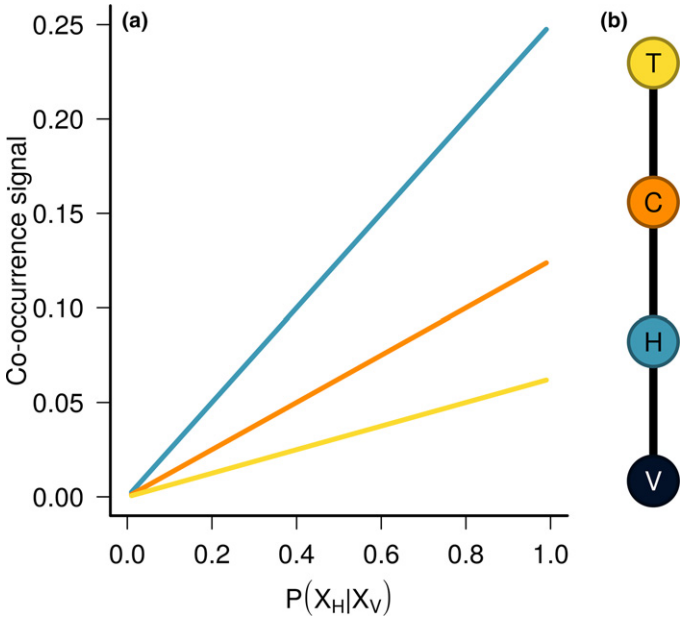

Figure 3 Significant spatial associations can emerge from indirect interactions. Three co-occurrence signals quantifying the association between species in the food chain and the resource species are computed for an increasing association between a resource an its consumer for a linear chain of four species (b) including a plant species V (the letter 'V' was used for 'vegetation'), a herbivore H, a carnivore C and a top predator T. The co-occurrence signals are calculated as follow: $P(X_{V,H})-P(X_V)\ P(X_H)$ (blue line), $P(X_{V,H})-P(X_V)\ P(X_H)$ (orange line) and $P(X_{V,T})-P(X_V)\ P(X_T)$ (yellow line) for $P(X_V) = 0.5$, $P(X_C|X_H) = 0.5$ and $P(X_T|X_C) = 0.5$.

SAMPLING IS A KEY TO MAKING CORRECT INFERENCE

In this section, we focus on the role played by different characteristics of the sampling design in the inference of species interactions from presence–absence data. To assess whether a co-occurrence is not spurious, it is important to sample enough, to sample properly and to integrate the metadata

pertaining to it (e.g. size of the sampling unit, spatial location, etc.) Assuming that the data gathered is well sampled and in large enough quantities, one can then give an interpretation of the estimated co-occurrence. In particular, the choice of the spatial scale at which to sample and the sampling effort have important impacts on the co-occurrence signal computed.

Argument 4 – Sampling scale influences measures of co-occurrence

Rationale

It has repeatedly been argued that interactions must be a major determinant of the broad geographical distribution of species (Holt & Barfield, 2009; Benning *et al.,* 2019), but also that, as a local process (acting at the individual scale), their impact may not be discernible at coarser spatial scales (Pearson & Dawson, 2003; Russell *et al.,* 2006; McGill, 2010). While the problem of sampling scale in co-occurrence studies has been raised early in the literature (Michael, 1920), biogeographers still investigate this technical but central topic (Araújo & Rozenfeld, 2014; Thuiller *et al.,* 2015; Belmaker *et al.,* 2015; Bar-Massada *et al.,* 2018). For instance using simulations, Araújo & Rozenfeld (2014) demonstrated that while negative interactions quickly vanish as the spatial extent of sampling unit increases, the imprint of positive interactions scales up. Such findings emphasise that sampling resolution needs to be carefully chosen so that the true underlying co-occurrence signal can be extracted from the data. From a mathematical standpoint, we show in the SI that this argument is general and that it is related to Argument 1 as is shown in 'The problem of abiotic factors'.

To illustrate this aspect of the sampling design, we considered two independent species A and B, that either poorly overlap (Fig. 4a) or strongly overlap (Fig. 4b). For these two scenarios, we simulated sampling and then computed co-occurrence signal along the gradient for a moving window that increases in size (see section 'Simulation' in the SI for more details). What is striking about the results obtained is that for two negatively associated species (Fig. 4a and c), a sampling area that encompasses most (but not all) of the distributional range of both species tends to overemphasise the negative association between the species. Conversely, when two species

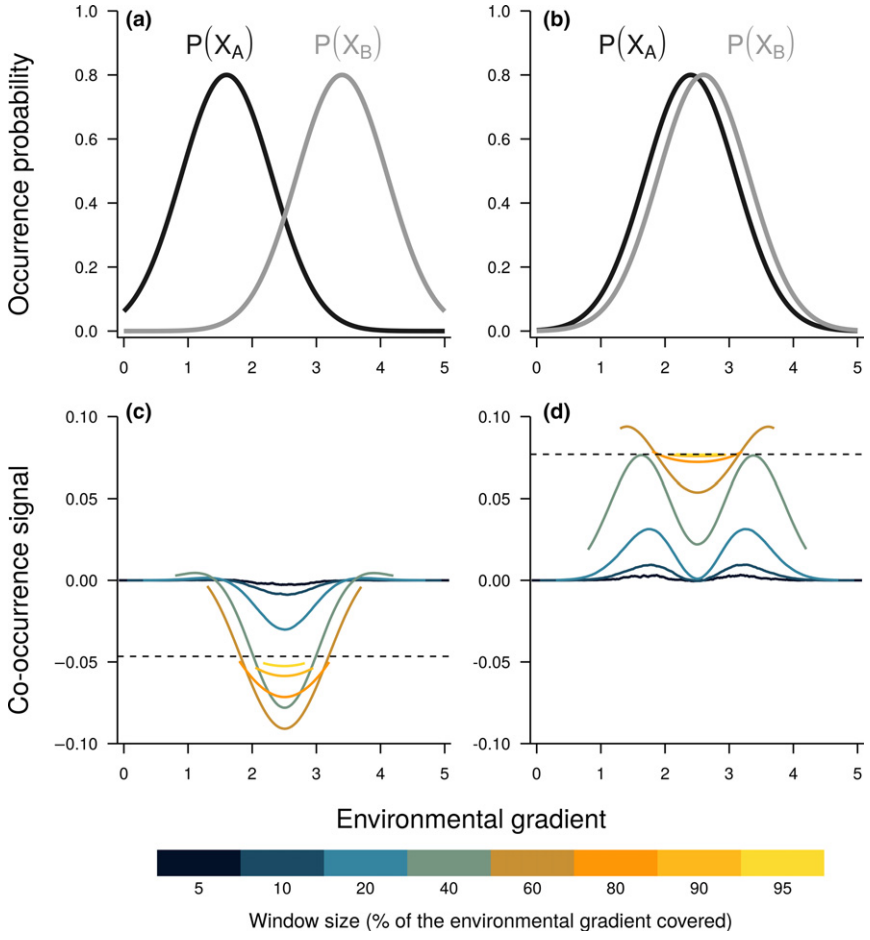

Figure 4 Co-occurrence signal and sampling scale. Top panels describe the occurrence probabilities along an environmental gradient of the independent species A and B in two contrasting scenarios. In (a) species occur in different abiotic condition whereas in (b) they share very similar environmental requirements. The corresponding bottom panels represent co-occurrence signals (measured as $P(X_A,X_B)-P(X_A)\ P(X_B)$, see 'Simulations' in SI) along the environmental gradient using moving windows of different size as the sampling area considered to assess co-occurrence structure. Dotted lines represent the co-occurrence signal computed over the entire gradient.

are positively associated (Fig. 4b and d), the co-occurrence signal varies widely, especially for a sampling area that includes between roughly 30% and 50% of the distributional range of both species. Thus, the associations detected highly depend on the portion of the environmental gradient considered.

Conclusion

In addition to the crucial importance of sampling resolution (Araújo & Rozenfeld, 2014; Thuiller *et al.,* 2015; Bar-Massada *et al.,* 2018), the portion of the environmental gradient sampled should also be carefully examined to avoid erroneous conclusions (Bar-Massada & Belmaker, 2017). To infer ecological interactions from co-occurrence data, the full distributional range of both species needs to be considered. In more colloquial terms, there are no free lunches when assessing co-occurrence through observational data.

Argument 5 – Appropriate statistical inference requires a very large sample size

Rationale

Species co-occurrence is a relatively rare phenomenon to document. To contextualise, it is already challenging to gather a large enough number of samples to estimate how a single species relates to the environment. Although there is no specific sample size prescription for species distribution models, studying model significance (e.g. through the use of confidence intervals), we generally have a good idea of whether a sample was large enough to obtain reliable results. However, to estimate a co-occurrence, many samples are required, much more than what is typically used to measure co-occurrence structure.

How many samples is enough samples?

Let's consider a simple situation with two relatively common species. If we assume that species A occurs in 40% ($P(X_A) = 0.4$) of samples and species B in 60% ($P(X_B) = 0.6$), probability theory tells us that the null expectation of co-occurrence between the two species will be $P(X_A) \times P(X_B) = 0.24$. Of course, this probability will increase (decrease) as the co-occurrence signal between the two species also increases (decreases). However, it is not readily obvious how many samples would be required to assess whether the association between species A and B is different from a null expectation or to evaluate if both species co-occur with a particular correlation level, say, 0.9, 0.5, −0.5, −0.9. This question can be approached using the multivariate Bernoulli distribution (Teugels, 1990) and binomial confidence intervals (DasGupta *et al.* (2001) compares different techniques to calculate confidence intervals on binomial data).

The results in Fig. 5 show that in the best case scenario, over 500 samples are required to reach a 95% confidence limit. Note that this example is actually conservative because when the probability of occurrence of each pair of species is either higher or lower, the number of samples increases to many thousands of samples.

What is even more worrying is that the results presented in Fig. 5 assume that the pair of species are solely influencing each other, a rare case in nature. Species often interact with a group of other species which will, in most cases, reduce the probability of co-occurrence on the considered species pair (see *Argument 2*) and in turn require that an even larger number of samples be gathered to efficiently measure the co-occurrence between the two species.

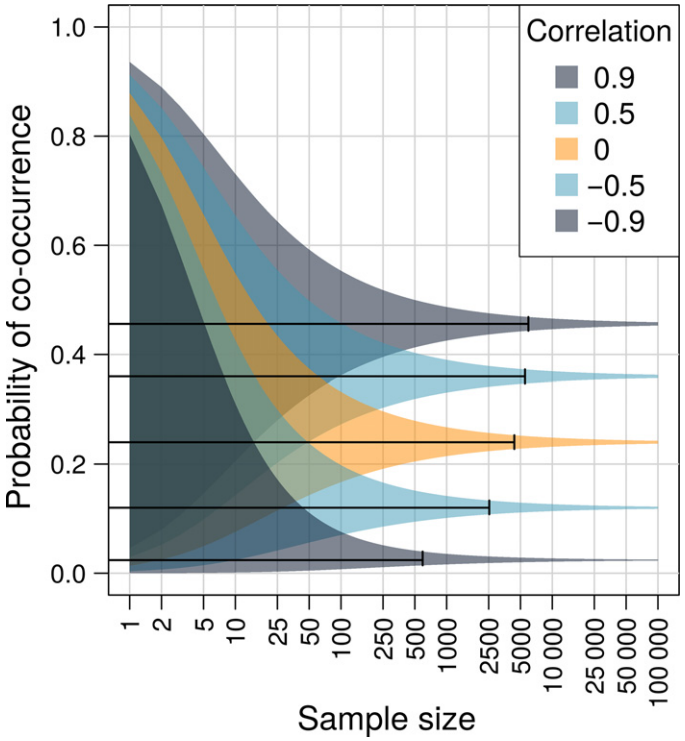

Figure 5 Co-occurrence signal and sample size. Estimated confidence intervals (coloured envelopes) given a specific sample size for a pair of species presenting different level of correlations. The probability of occurrence for the two species is 0.4 and 0.6. As such, the probabilities of co-occurrence varied depending on the correlation levels. The true probabilities of co-occurrence2 are illustrated by black horizontal lines. The short vertical black lines in each envelopes highlight the number of samples required to reach a 95% level of confidence in the estimated co-occurrence. The calculation of the confidence intervals were performed using the Wilson scores intervals, which have been shown to be accurate and robust (DasGupta *et al.,* 2001). To calculate the number of samples required to reach a 95% level of confidence, we applied a Dunn-Sidak correction (Šidák, 1967) because two species were used to compute the co-occurrence probability.

Conclusion

The number of samples required to accurately measure co-occurrence among species is impractical for most studies. As a comparison, it is common for studies in ecology focusing on co-occurrence to have a small sample size compared to what is discussed in this section. For example of the 294 data sets gathered in Atmar & Patterson (1995), only four had more than 100 samples, the largest having 202 samples. This is not unique to ecology, environmental microbiology (Rocca *et al.,* 2019) and microbiome research (e.g. Levy & Borenstein, 2013) suffer from the same problem. Granted, in the last 20 years larger data sets are becoming increasingly available (see, e.g. the data sets used by Ovaskainen *et al.,* 2017). That being

said, studies with a sample size that meet the requirements highlighted in this argument are still extremely rare.

The results of this section suggest that no statistical approach, regardless of its level of sophistication, can be used to assess spatial associations between species accurately, even for reasonably large sample size.

THE IMPRINT OF ECOLOGICAL INTERACTIONS ON CO-OCCURRENCE DATA

In this section we discuss the relationships we should expect from co-occurrence data based on current ecological theory. Whether it is from foodweb or coexistence theories, we have learned that there are different types of interactions with different strengths. In this section, we discuss how what we know of ecological interactions is expressed in co-occurrence data.

Argument 6 – Asymmetry of associations between species can blur co-occurrence signal

Rationale

Different types of interactions do not result in the same co-occurrence signal (Araújo & Rozenfeld, 2014). Most co-occurrence analyses are, however, derived from the joint species distribution (as defined above), which is a symmetric measure of spatial association. Interactions may, however, differ in magnitude and/or in their effect (positive or negative). As such, there is no reason why two species should have exactly the same effect on each other and we should therefore expect species-specific variation in the co-occurrence signal. Furthermore, some interactions such as predation, herbivory or parasitism could even lead to opposing signals, making the expectation for the joint species distribution indeterminable. This is noteworthy because these types of interactions have traditionally been the most studied ones in community ecology and are now increasingly inferred from proxies (Morales-Castilla *et al.,* 2015), including co-occurrence.

This argument is best understood with a decomposition of the joint probability of occurrence. Using the product rule (and ignoring the effect of the environment E), we find that the joint distribution of species A and B can be decomposed into the product of conditional and marginal probabilities.

$$P(X_A, X_B) = P(X_A|X_B)P(X_B) \quad (9)$$

and inversely

$$P(X_A, X_B) = P(X_B|X_A)P(X_A). \quad (10)$$

In the previous equations, the conditional occurrence probability $P(X_A|X_B)$ is the measure of the effect of species B on the occurrence of species A. Unless the marginal probabilities are exactly the same, the conditional occurrence probabilities must absolutely differ from each other to equal the joint occurrence probability. This means that the joint occurrence probability masks the variability in the strength of associations between species.

The decomposition presented above may have unexpected and far-reaching impact. For example strong negative and positive associations, such as between a predator and a prey, may cancel each other and result in a joint occurrence probability not different from the null expectation. A numerical example best illustrates this point. Let's consider a predator A with marginal occurrence $P(X_A) = 0.2$ and a prey with marginal occurrence $P(X_B) = 0.5$. We know from probability theory that their random expectation is $P(X_A) \times P(X_B) = 0.1$. Let us further assume that their realisation is $P(X_A,X_B) = 0.15$, so slightly above the expectation. Using these values and eqns 9 and 10, we can calculate the probability of finding the predator given the presence of the prey $P(X_A = 1|X_B = 1) = 0.3$ or its absence $P(X_A = 1|X_B = 0) = 0.167$. This result states that it is almost twice as probable for a predator and a prey to be found together then separated. Conversely, using the same approach, we find the conditional occurrence of the prey in the predator's absence to be $P(X_B = 1|X_A = 0) = 0.5625$, which is more than two times larger than in the presence of the predator, $P(X_B = 1|X_A = 1) = 0.25$. This simplistic example shows how variable the conditional probabilities can be and how they can have opposite effects, even if the joint occurrence is not much different from the null expectation.

Conclusion

Analysis of joint distribution of presence–absence data is not appropriate to assess interactions because not all asymmetric interactions can be identified. This particularity of co-occurrence data may lead to biased interpretation of interactions towards symmetric interactions. In this respect, conditional probabilities are more relevant to document variance in association strength as well as asymmetric associations. There are four conditional probabilities associated with a pair of co-occurring species and their comparison reveals the direction and strength of effects of one species on another. While conditional probabilities are very promising and could be extended to an entire network using Bayesian networks (Staniczenko *et al.,* 2017), they may be challenging to solve, especially when cycles are present in the network.

Argument 7 – Coexistence theory predicts that strong interactions may lead to exclusion before leaving a significant signal

Rationale

In a competition system, stable coexistence, whether it is at the local or regional scale, requires interspecific interactions to be weaker than intraspecific interactions (Chesson, 2000). The weaker competitor tends to get excluded when interaction strength increases. This narrows down the range where interactions can actually be detected using co-occurrence data: if interactions are too weak, the imprint left in co-occurrence data may be undetectable, but if interactions are too strong it may prevent coexistence from happening.

This assertion can be explored using a multi-species adaptation of the Levins (1969) metapopulation model. Such a model was presented by Hanski (1983) to illustrate the patch dynamics between a strong (species A) and a weak (species B) competitor as well as to quantify the proportion of patches occupied solely by either or both of the two species. Using this model, we can vary colonisation competition (corresponding to pre-emptive competition) or extinction competition (corresponding to competitive exclusion) (Gravel &

Massol, 2020). In doing so, we can investigate the proportion of patches where species co-occurrence vary while interaction strength increases. Intuitively, the stronger preemptive competition and competitive exclusion are, the smaller the co-occurrence will be relative to marginal occurrence (species will avoid each other). This is indeed what the model predicts. In addition, it also shows that marginal occurrences of the weak competitor rapidly decline when the interaction strength increases, resulting in very small absolute co-occurrence (Fig. 6). Given *Argument 5*, based on this result, we would need a very large sample size to document such rare phenomena. As such, it is unlikely that spatial repulsion may be detected when interaction strength is strong.

Conclusion

Strong negative interspecific interactions are incompatible with coexistence. Species may be excluded by competition before the interaction signal can be captured in co-occurrence data. In other words, a species absent regionally cannot generate any interaction signal because it will never be sampled. Obviously, the degree of spatial association between pairs of species cannot be measured following competitive exclusion – a big limitation of any co-occurrence analysis since the strong interactions most likely to impact distribution cannot be measured.

This argument could also be made for predator–prey interactions in space, but it may not apply to all types of interactions, if not opposite for positive interactions (see Gravel & Massol, 2019), it nonetheless leads to the paradox that the strong interactions we want to document with co-occurrence may be impossible to measure.

CONCLUDING REMARKS

The seven arguments we present paint a rather grim picture of the problems related to the inference of interactions from co-occurrence data. There are two broad conclusions to be drawn from them. First, the various layers of complexity inherent to ecological systems (e.g. environmental variability, diversity of biotic interactions, etc.) blur the link between interactions and co-occurrence. This is not even accounting for other more specific aspects of ecological systems. For example it is inherently challenging to measure interactions among rare or transient species, regardless of the approach used (Calatayud *et al.*,2020). Similarly, particular species may interact only in specific situations, making their assessment difficult because the context influencing these interactions may be difficult to evaluate. Also, most co-occurrence analysis considers species distributions to be at equilibrium, which could be dynamic (e.g. metapopulation) or not, a constraining assumption especially in the context of environmental changes (e.g. climate change). Second, because the relationship between interactions and co-occurrence is rarely clear, there are several technical and theoretical challenges to infer ecological interactions from co-occurrence data that still remain to be tackled.

One aspect of ecological interactions that was not discussed in this paper was the importance of temporal variations and its impact on species and their interactions. Accounting for

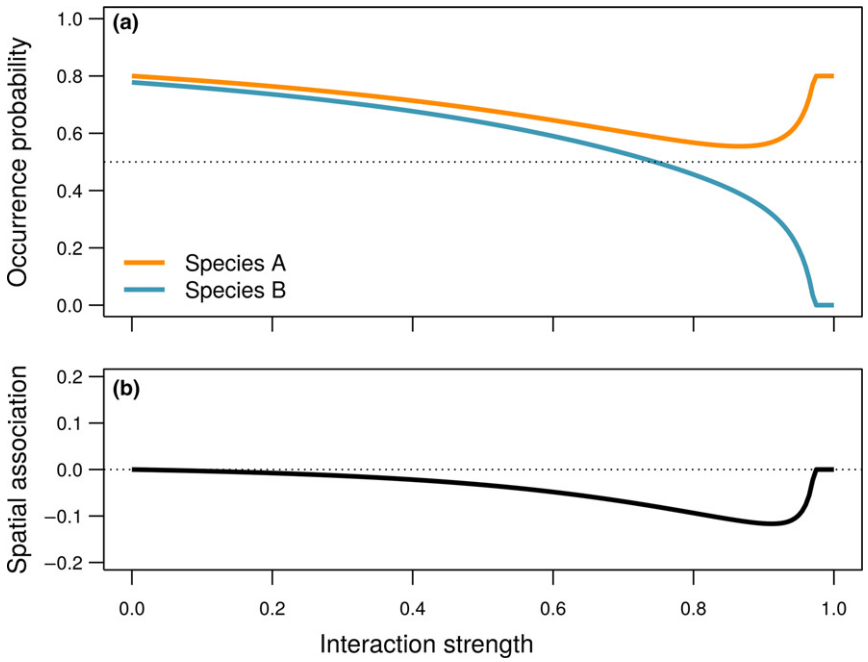

Figure 6 Co-occurrence signal and interaction strength in a metacommunity model (see Gravel & Massol (2019) for model specification). Competition for space between two species in a metacommunity impact regional occurrence (a) and co-occurrence (b). Increasing interaction strength reduces the occurrence of both species, up to the point where the weak competitor (species B, blue) is regionally extinct and the strong competitor (species A, orange) reaches its regional capacity. The strength of co-occurrence relative to the random expectation increases with interaction strength, but is hardly detectable because of a coincident reduction in the frequency of co-occurrence.

time when assessing ecological interactions is undoubtedly important and has potentially far-reaching consequences. For example hibernation, migration and phenology are all temporal drivers of change for ecological interactions. However, assessing interactions from temporal co-occurrence raises a number of additional issues that are outside the scope of this paper.

Independently and together, the arguments we developed illustrate the diversity of those challenges (Box 1). Even if statistical, sampling or theoretical solutions can be found for some, it is difficult to contemplate a solution that would solve all problems raised, while still using presence–absence data. The minimal amount of information these data carry is indeed at the core of most of the challenges we pointed out in this study. Even JSDMs, which have been seen by some as an appealing new modelling framework to efficiently study ecological interactions, are not able to tackle most of the arguments we presented above. The correlation matrix (estimated from JSDMs) describing the association among pairs of species, now often used as illustration to represent species association, is likely what triggered the interest of using JSDMs to infer species interactions. Using such a representation, it is extremely tempting to make the intellectual jump to infer ecological interactions. In this respect, we advocate that such representation should never be used when the underlying data used to construct the model is presence–absence data.

Our perspective of the problems related to co-occurrences and its use to study interactions has far-reaching implications for some historical debates in community ecology. Among others, it suggests the importance of revisiting the long-standing debates on null models initiated by Diamond (1975) and Connor & Simberloff (1979). The arguments we raised have implications for the ability to detect significant signals with null models. Scale and sampling effort problems have been debated for a while in this 'null model war'. Yet, the other arguments also need considerations. For example even the most recent developments to analyse co-occurrence data using null models (e.g. D'Amen et al., 2018) still assume species influence each other with the same level of intensity. More fundamentally, all the arguments we present show that the observation of spatial associations (or the lack thereof) may be impossible to accurately assess and interpret. We do not suggest here that the whole field pertaining to null models should be discarded. Rather, researchers should be more critical of the limits of their tools when interpreting their results.

Presence–absence data undoubtedly remain central to ecology and ecologists must certainly keep collecting them in order to broaden our knowledge on species distributions and our understanding of the factors that determine the presence of a specific community in a particular area. But it is also crucial to identify what can be inferred from such data and what cannot. Some avenues deserve to be explored bearing that in mind. A conceptually simple but technically challenging solution would be to derive interactions from abundance (count) data instead of presence–absence data. In addition to showing where a species occur, abundance data also carry more information that could be used to make more refined inferences on why a species occur at a specific location, including demographical processes, fine spatial variation, etc. All of this additional information cannot be gathered with presence–absence data. From a modelling perspective, tools exist that can be used to assess relationships among species using abundance data and infer interactions (Faisal *et al.,* 2010; Poisot *et al.,* 2015; Popovic *et al.,* 2019). From an empirical perspective, a few studies have used different ways to infer interactions from abundance (or plant cover) data both in ecology (le Roux *et al.,* 2013) and microbiology (Levy & Borenstein, 2013). Hopefully, using ecological data carrying more information than presence–absence data (such as abundance data) would provide reliable proxies for biotic interactions.

Another direction worth investigating is to study interactions through the eyeglass of conditional probabilities. Through this approach, we can get a much more direct interpretation of how a species reacts in the presence of other species by being more mathematically explicit about how species relate to each other. Occurrence probabilities could for instance be conditional on the presence of another species and on the environment, avoiding the problems associated with Argument 1 and Argument 4. Conditional probabilities are not symmetrical by default, avoiding Argument 6. It may, however, be difficult to tease apart indirect interactions (Argument 3) and the signal will still vanish when there is a high number of interactions (Argument 2). As a result, networks of conditional probabilities may still require a prohibitive sampling intensity (Argument 5). From a statistical perspective, Bayesian networks and Markov networks offer appealing avenues to investigate. There are, however, several technical challenges that will need to be solved before these could be used. Among them, the presence of cycles (species A affect B and vice-versa) is a fundamental problem, large sample size cannot be avoided and some prior knowledge of interactions is also required.

Yet another way to study species interactions is with mechanistic models where the known (or hypothesised) mechanisms of interactions are explicitly accounted for. By testing how close these mechanistic models represent data on species associations, we can then infer the underlying processes structuring species. Mechanistic models such as the general metacommunity model of Hanski (1983) for competition and its revision for all types of interactions (Gravel & Massol, 2020) could be used to further understand species interaction. However, these models are very data-hungry and therefore difficult to apply in practice. Our intuition is that working with dynamic data (e.g. for colonisation and extinctions) may avoid some of the problems we raised since in this situation, we are closer to the processes of assembly and not relying on equilibrium assumptions. Cirtwill & Stouffer (2016) proposed an avenue worth further investigation using Simberloff's classic defaunation experiment.

Experiments represent one avenue that needs to be further explored to understand how biotic interactions impact distribution. Although they are often time consuming and costly, they can increase our knowledge on interactions while remaining in the world of co-occurrences. As an example, Brazeau & Schamp (2019) have recently studied experimentally the link between competition and negative co-occurrence for flowering plants. Similarly, Kopelke *et al.* (2017)

gathered a large data set directly recording the interaction between willow tree species and sawflies, a group of parasites feeding on willow leaves. These studies are particularly interesting because they focus explicitly on interactions. A way forward would be to pursue the development of modelling approaches that can make full use of these data sets but also of theoretical ideas that can advance our understanding of ecological interaction.

Early on ecologists recognised there is a large amount of unexplored information in co-occurrence data. Powerful new statistical tools are becoming available that allow ecologists to gain new insights from co-occurrence data and efforts should continue in that way. That said, although very tempting at first, with our current knowledge, interpreting significant co-occurrence signals between species as evidence of ecological interactions should be avoided.

Box 1: Outstanding Questions

- Under which assumptions are co-occurrence signals good proxies for ecological interactions?
- How to interpret co-occurrence networks? Even though such networks are more frequently used in the literature, from a theoretical point of view, it remains unclear how they relate to ecological interaction networks.
- What are the relevant covariates required to infer ecological interactions from species distributions? In cases where distribution data are not enough to conclude, is it possible to provide additional information to infer interactions?
- How strong are interaction signals in abundance data? This can be assessed using data gathered on species known to interact. Statistically, many of the methods currently available and applied on co-occurrence data can be used (and compared) to approach this question.
- How important are detection errors when assessing interactions? There are a growing number of models that accounts for detection errors. They were never used to study the influence of detection error when measuring interactions.
- How can we account for cycles in network models? Currently, network models cannot account for cycles which are an inherent part of ecological networks.
- How can we design experiments to test if interactions have an effect (or not) on occurrence?

ACKNOWLEDGEMENTS

We thank M. Araujo, A. Rozenfeld, W. Thuiller, M. Talluto, L. Pollock, T. Poisot, O. Ovaskainen, W. Godsoe, B. Holt, P. Peres-Neto and F. Massol for insightful discussions on co-occurrence and species distribution modelling. We are also grateful to B. Blonder, L. Thurman and one anonymous reviewer for their thoughtful comments and efforts towards improving our manuscript. The picture of Mont Mégantic was taken by M. Vellend. This study was financially supported by the Canada Research Chair programme and NSERC Discovery grant to DG.

AUTHORSHIP STATEMENT

All authors made significant and equal contribution to this article.

DATA STATEMENT

No ecological/biological data were used in this study. The picture of Mont Mégantic in Fig. 1a was provided by M. Vellend.

DATA AVAILABILITY STATEMENT

The code and simulation output are available through GitHub https://github.com/TheoreticalEcosystemEcology/coocNotInteract (https://doi.org/10.5281/zenodo.3733206).

REFERENCES

Alroy, J. (2015). A new twist on a very old binary similarity coefficient. *Ecology*, 96, 575–586.

Araújo, M.B. & Rozenfeld, A. (2014). The geographic scaling of biotic interactions. *Ecography*, 37, 001–010.

Araújo, M.B., Rozenfeld, A., Rahbek, C. & Marquet, P.A. (2011). Using species co-occurrence networks to assess the impacts of climate change. *Ecography*, 34, 897–908.

Arita, H.T. (2016). Species co-occurrence analysis: pairwise versus matrix-level approaches: Correspondence. *Glob. Ecol. Biogeogr.*, 25, 1397–1400.

Arita, H.T., Christen, A., Rodríguez, P. & Soberón, J. (2012). The presence-absence matrix reloaded: the use and interpretation of range-diversity plots: Range-diversity plots. *Glob. Ecol. Biogeogr.*, 21, 282–292.

Atmar, W. & Patterson, B.D. (1995). The nestedness temperature calculator: a visual basic program, including 294 presence-absence matrices. Aics research, inc. edn. University Park, NM, and The Field Museum, Chicago.

Bar-Massada, A. & Belmaker, J. (2017). Non-stationarity in the co-occurrence patterns of species across environmental gradients. *J. Ecol.*, 105, 391–399.

Bar-Massada, A., Yang, Q., Shen, G. & Wang, X. (2018). Tree species co-occurrence patterns change across grains: insights from a subtropical forest. *Ecosphere*, 9, e02213.

Barbaro, L., Allan, E., Ampoorter, E., Castagneyrol, B., Charbonnier, Y., De Wandeler, H. *et al.* (2019). Biotic predictors complement models of bat and bird responses to climate and tree diversity in European forests. *Biol. Sci.*, 286, 20182193.

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.

Belmaker, J., Zarnetske, P., Tuanmu, M.N., Zonneveld, S., Record, S., Strecker, A. & *et al.* (2015). Empirical evidence for the scale dependence of biotic interactions: scaling of biotic interactions. *Glob. Ecol. Biogeogr.*, 24, 750–761.

Benning, J.W., Eckhart, V.M., Geber, M.A. & Moeller, D.A. (2019). Biotic interactions contribute to the geographic range limit of an annual plant: herbivory and phenology mediate fitness beyond a range margin. *Am. Nat.*, 193, 786–797.

Berlow, E.L., Neutel, A.M., Cohen, J.E., de Ruiter, P.C., Ebenman, B., Emmerson, M. *et al.* (2004). Interaction strengths in food webs: Issues and opportunities. *J. Anim. Ecol.*, 73, 585–598.

Berry, D. & Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*, 5, 219, 1-14.

Brazeau, H.A. & Schamp, B.S. (2019). Examining the link between competition and negative co-occurrence patterns. *Oikos*, 128, 1358-1366.

Calatayud, J., Andivia, E., Escudero, A., Melián, C.J., Bernardo-Madrid, R., Stoffel, M. *et al*. (2020). Positive associations among rare species and their persistence in ecological assemblages. *Nat. Ecol. Evol*., 4, 40–45.

Cardillo, M. (2011). Phylogenetic structure of mammal assemblages at large geographical scales: linking phylogenetic community ecology with macroecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366, 2545–2553.

Cardillo, M. & Meijaard, E. (2010). Phylogeny and co-occurrence of mammal species on Southeast Asian islands. *Global Ecol. Biogeog*. 19, 465-474.

Cazelles, K., Araújo, M.B., Mouquet, N. & Gravel, D. (2016). A theory for species co-occurrence in interaction networks. *Theor. Ecol.*, 9, 39–48.

Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst*., 31, 343–366.

Cirtwill, A.R. & Stouffer, D.B. (2016). Knowledge of predator-prey interactions improves predictions of immigration and extinction in island biogeography: Predator-prey interactions and island biogeography. *Glob. Ecol. Biogeogr*., 25, 900–911.

Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K. (2014). More than the sum of the parts: forest climate response from joint species distribution models. *Ecol. Appl*., 24, 990–999.

Clark, J.S., Nemergut, D., Seyednasrollah, B., Turner, P.J. & Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecol. Monogr*., 87, 34–56.

Clark, N.J., Wells, K. & Lindberg, O. (2018). Unravelling changing interspecific interactions across environmental gradients using Markov random fields. *Ecology*, 99, 1277–1283.

Connor, E.F., Collins, M.D. & Simberloff, D. (2013). The checkered history of checkerboard distributions. *Ecology*, 94, 2403–2414.

Connor, E.F. & Simberloff, D. (1979). The assembly of species communities: Chance or competition? *Ecology*, 60, 1132.

D'Amen, M., Mod, H.K., Gotelli, N.J. & Guisan, A. (2018). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. *Ecography*, 41, 1233–1244.

DasGupta, A., Cai, T.T. & Brown, L.D. (2001). Interval estimation for a binomial proportion. *Stat. Sci*., 16, 101–133.

Diamond, J., Pimm, S.L. & Sanderson, J.G. (2015). The checkered history of checkerboard distributions: comment. *Ecology*, 96, 3386–3388.

Diamond, J.M. (1975). Assembly of species communities. In: *Ecology and Evolution of Communities* (eds Cody, M.L. & Diamond, J.M.). Harvard Univ Press, Cambridge, Mass, pp. 342–444.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A. et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129–151.

Faisal, A., Dondelinger, F., Husmeier, D. & Beale, C.M. (2010). Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. *Ecological Informatics*, 5, 451–464.

Faust, K. & Raes, J. (2012). Microbial interactions: From networks to models. *Nat. Rev. Microbiol*., 10, 538–550.

Forbes, S. (1907). On the Local Distribution of Certain Illinois Fishes: An Essay in Statistical Ecology. No. v. 7, no. 8 in Bulletin of the Illinois State Laboratory of Natural History. Illinois State Laboratory of Natural History.

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.

Godsoe, W., Franklin, J. & Blanchet, F.G. (2017). Effects of biotic interactions on modeled species' distribution can be masked by environmental gradients. *Ecol. Evol*., 7, 654–664.

Gotelli, N.J. & Ellison, A.M. (2002). Assembly rules for New England ant assemblages. *Oikos*, 99, 591–599.

Gotelli, N.J. & Graves, G.R. (1996). *Null Models in Ecology*. Smithsonian Institution Press, Washington.

Gotelli, N.J., Graves, G.R. & Rahbek, C. (2010). Macroecological signals of species interactions in the Danish avifauna. *Proc. Natl Acad. Sci*., 107, 5030–5035.

Gotelli, N.J. & McCabe, D.J. (2002). Species co-occurrence: A meta-analysis of JM Diamond's assembly rules model. *Ecology*, 83, 2091–2096.

Gotelli, N.J. & Ulrich, W. (2010). The empirical Bayes approach as a tool to identify non-random species associations. *Oecologia*, 162, 463–477.

Gravel, D. & Massol, F. (2020). Toward a general theory of metacommunity ecology. In: *Theoretical Ecology: concepts and applications*. Oxford University Press, Oxford, p. 195-284.

Hanski, I. (1983). Coexistence of competitors in patchy environment. *Ecology*, 64, 493–500.

Harris, D.J. (2016). Inferring species interactions from co-occurrence data with Markov networks. *Ecology*, 97, 3308–3314.

Heikkinen, R.K., Luoto, M., Virkkala, R., Pearson, R.G. & Körber, J.H. (2007). Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Glob. Ecol. Biogeogr*., 16, 754–763.

Holt, R.D. & Barfield, M. (2009). Trophic interactions and range limits: the diverse roles of predation. *Proceedings of the Royal Society B: Biological Sciences*, 276, 1435–1442.

Hui, F.K. (2016). boral- Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R. *Methods Ecol. Evol*., 7, 744–750.

Jeschke, J.M. & Strayer, D.L. (2008). Usefulness of Bioclimatic Models for Studying Climate Change and Invasive Species. *Ann. N. Y. Acad. Sci*., 1134, 1–24.

Kaldhusdal, A., Brandl, R., Müller, J., Möst, L. & Hothorn, T. (2015). Spatio-phylogenetic multispecies distribution models. *Methods Ecol. Evol*., 6, 187–197.

Kay, G.M., Tulloch, A., Barton, P.S., Cunningham, S.A., Driscoll, D.A. & Lindenmayer, D.B. (2017). Species co-occurrence networks show reptile community reorganization under agricultural transformation. *Ecography*, 41, 113-125.

Kopelke, J.P., Nyman, T., Cazelles, K., Gravel, D., Vissault, S. & Roslin, T. (2017). Food-web structure of willow-galling sawflies and their natural enemies across Europe. *Ecology*, 98, 1730–1730.

Lane, P.W., Lindenmayer, D.B., Barton, P.S., Blanchard, W. & Westgate, M.J. (2014). Visualization of species pairwise associations: a case study of surrogacy in bird assemblages. *Ecol. Evol*., 4, 3279–3289.

Latimer, A.M., Banerjee, S., Sang, H., Mosher, E.S. & Silander, J.A. (2009). Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecol. Lett*., 12, 144–154.

Leach, K., Montgomery, W.I. & Reid, N. (2016). Modelling the influence of biotic factors on species distribution patterns. *Ecol. Model*., 337, 96–106.

Leathwick, J.R. & Austin, M.P. (2001). Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology*, 82, 2560–2573.

Leibold, M.A. & Mikkelson, G.M. (2002). Coherence, species turnover, and boundary clumping: elements of meta-community structure. *Oikos*, 97, 237–250.

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull. Entomol. Soc. Am*., 15, 237–240.

Levy, R. & Borenstein, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl Acad. Sci*., 110, 12804–12809.

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E. (2018). Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Scientific Reports*, 8. 5875, 1-12.

McGill, B.J. (2010). Matters of scale. *Science*, 328, 575–576.

Meier, E.S., Kienast, F., Pearman, P.B., Svenning, J.C., Thuiller, W., Araújo, M.B. *et al.* (2010). Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography*, 33, 1038–1048.

Michael, E.L. (1920). Marine ecology and the coefficient of association: a plea in behalf of quantitative biology. *J. Ecol.*, 8, 54.
Morales-Castilla, I., Matias, M.G., Gravel, D. & Araújo, M.B. (2015). Inferring biotic interactions from proxies. *Trends Ecol. Evol.*, 30, 347–356.
Morueta-Holme, N., Blonder, B., Sandel, B., McGill, B.J., Peet, R.K., Ott, J.E. *et al.* (2016). A network approach for inferring species associations from co-occurrence data. *Ecography*, 39, 1139–1150.
Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.*, 7, 549–555.
Ovaskainen, O., Hottola, J. & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91, 2514–2521.
Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D. *et al.* (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.*, 20, 561–576.
Patterson, B.D. & Atmar, W. (1986). Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biol. J. Lin. Soc.*, 28, 65–82.
Pearson, R.G. & Dawson, T.P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob. Ecol. Biogeogr.*, 12, 361–371.
Pielou, D. & Pielou, E. (1967). The detection of different degrees of coexistence. *J. Theor. Biol.*, 16, 427–437.
Pielou, D.P. & Pielou, E.C. (1968). Association among species of infrequent occurrence: the insect and spider fauna of Polyporus betulinus (Bulliard) Fries. *J. Theor. Biol.*, 21, 202–216.
Poisot, T., Stouffer, D.B. & Gravel, D. (2015). Beyond species: why ecological interaction networks vary through space and time. *Oikos*, 124, 243–251.
Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M. *et al.* (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol. Evol.*, 5, 397–406.
Popovic, G.C., Warton, D.I., Thomson, F.J., Hui, F.K.C. & Moles, A.T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods Ecol. Evol.*, 10, 1571–1583.
Presley, S.J., Higgins, C.L. & Willig, M.R. (2010). A comprehensive framework for the evaluation of metacommunity structure. *Oikos*, 119, 908–917.
Rocca, J.D., Simonin, M., Blaszczak, J.R., Ernakovich, J.G., Gibbons, S.M., Midani, F.S. et al. (2019). The microbiome stress project: toward a global meta-analysis of environmental stressors and their effects on microbial communities. *Front. Microbiol.*, 9. 3272, 1-14.
le Roux, P.C., Lenoir, J., Pellissier, L., Wisz, M.S. & Luoto, M. (2013). Horizontal, but not vertical, biotic interactions affect fine-scale plant distribution patterns in a low-energy system. *Ecology*, 94, 671–682.
Russell, R., Wood, S., Allison, G. & Menge, B. (2006). Scale, environment, and trophic status: the context dependency of community saturation in rocky intertidal communities. *Am. Nat.*, 167, E158–E170.
Savage, J. & Vellend, M. (2015). Elevational shifts, biotic homogenization and time lags in vegetation change during 40 years of climate warming. *Ecography*, 38, 546–555.
Sfenthourakis, S., Tzanatos, E. & Giokas, S. (2006). Species co-occurrence: The case of congeneric species and a causal approach to patterns of species association. *Glob. Ecol. Biogeogr.*, 15, 39–49.
Staniczenko, P.P., Sivasubramaniam, P., Suttle, K.B. & Pearson, R.G. (2017). Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecol. Lett.*, 20, 693–707.
Steele, J.A., Countway, P.D., Xia, L., Vigil, P.D., Beman, J.M., Kim, D.Y. *et al.* (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.*, 5, 1414–1425.
Stone, L. & Roberts, A. (1990). The checkerboard score and species distributions. *Oecologia*, 85, 74–79.
Teugels, J.L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *J. Multi. Anal.*, 32, 256–268.
Thuiller, W., Pollock, L.J., Gueguen, M. & Münkemüller, T. (2015). From species distributions to meta-communities. *Ecol. Lett.*, 18, 1321–1328.
Thurman, L.L., Barner, A.K., Garcia, T.S. & Chestnut, T. (2019). Testing the link between species interactions and species co-occurrence in a trophic network. *Ecography*, 42, 1658–1670.
Tulloch, A.I.T., Chadès, I., Dujardin, Y., Westgate, M.J., Lane, P.W. & Lindenmayer, D. (2016). Dynamic species co-occurrence networks require dynamic biodiversity surrogates. *Ecography*, 39, 1185–1196.
Ulrich, W. & Gotelli, N.J. (2013). Pattern detection in null model analysis. *Oikos*, 122, 2–18.
Veech, J.A. (2013). A probabilistic model for analysing species co-occurrence: probabilistic model. *Glob. Ecol. Biogeogr.*, 22, 252–260.
Veech, J.A. (2014). The pairwise approach to analysing species co-occurrence. *J. Biogeogr.*, 41, 1029–1035.
Warren, D.L., Cardillo, M., Rosauer, D.F. & Bolnick, D.I. (2014). Mistaking geography for biology: inferring processes from species distributions. *Trends Ecol. Evol.*, 29, 572–580.
Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., & Walker, S.C. *et al.* (2015). So many variables: joint modeling in community ecology. *Trends Ecol. Evol.*, 30, 766–779.
Whittam, T.S. & Siegel-Causey, D. (1981). Species interactions and community structure in alaskan seabird colonies. *Ecology*, 62, 1515–1524.
Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F. *et al.* (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.*, 88, 15–30.
Zavaleta, E.S., Hobbs, R.J. & Mooney, H.A. (2001). Viewing invasive species removal in a whole-ecosystem context. *Trends Ecol. Evol.*, 16, 454–459.
Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D.R., Bork, P. & Patil, K.R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl Acad. Sci.*, 112, 6449–6454.
Özesmi, S.L. & Özesmi, U. (1999). An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecol. Model.*, 116, 15–31.
Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Asso.*, 62, 626–633.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Editor, Elizabeth Jeffers
Manuscript received 20 January 2020
First decision made 24 February 2020
Manuscript accepted 7 April 2020