Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2013) **22**, 252–260

A Journal of Macroecology

Global Ecology and Biogeography

MACROECOLOGICAL METHODS

A probabilistic model for analysing species co-occurrence

Joseph A. Veech*

Department of Biology, Texas State University, San Marcos, TX 78666, USA

ABSTRACT

Aim To develop a new probabilistic model that can be used to test for statistically significant pair-wise patterns of species co-occurrence. The model gives the probability that two species would co-occur at a frequency less than (or greater than) the observed frequency if the two species were distributed independently of one another among a set of sites. The model can be used to classify species associations as negative, positive or random.

Innovation Historically, the analysis of species co-occurrence has involved the use of data randomization. An observed species presence–absence matrix is compared with randomized matrices to determine if the observed matrix has structure, either an excess or deficit of species positively or negatively associated with each other. The computer algorithms used to randomize matrices can sometimes produce Type I and Type II errors (when the randomization algorithm produces a biased set of all possible matrices) due to the randomization process itself. The probabilistic model does not rely on any data randomization, hence it has a very low Type I error rate and is powerful having a low Type II error rate.

Main conclusions When applied to 10 different data sets the probabilistic model revealed significant positive and negative species associations in most of the data sets. Compared with previous analyses the model tended to find fewer significant associations; this may indicate a generally low rate of Type I error in the model. The model is easy to implement and requires no special software. The model could potentially transform the way that ecologists test for species co-occurrence in a wide range of ecological studies.

*Correspondence: Joseph A. Veech, Department of Biology, Texas State University, San Marcos, TX 78666, USA.
E-mail: joseph.veech@txstate.edu

Keywords
Biogeography, combinatorics, community assembly, distribution, nestedness, null model, probability.

INTRODUCTION

The analysis of species co-occurrence patterns has a long history in ecology, dating back to at least the 1970s. Analysis of co-occurrence patterns played a central role in earlier debates about the importance of competition in structuring ecological communities and the existence of assembly rules (Diamond, 1975; Connor & Simberloff, 1979, 1983; Diamond & Gilpin, 1982; Gilpin & Diamond, 1982; Gotelli & Graves, 1996; Weiher & Keddy, 1999). To some degree, the controversy continues (Ellwood *et al.*, 2009; Kennedy, 2009; Ulrich *et al.*, 2009; Chase & Myers, 2011; Collins *et al.*, 2011; Sanderson *et al.*, 2011), and the analysis of co-occurrence patterns remains a huge area of ecological and statistical endeavour (Gotelli, 2000; Ladau, 2008; Hui, 2009; Ulrich *et al.*, 2009). A crucial question has always been to what extent are co-occurrence patterns random versus structured due to some organizing process? Moreover, this question must be addressed in the context that no species occurs randomly in nature, they all respond to environmental variation. Given this challenge, ecologists have long used null models and data randomization to attempt to answer the main question. But herein lies more controversy, because the metrics for measuring co-occurrence and for testing its statistical significance are not universally agreed upon. The metrics and various algorithms for randomizing species presence–absence data have different strengths and weaknesses regarding their statistical

DOI: 10.1111/j.1466-8238.2012.00789.x

properties, computational complexities and biological realism (Gotelli, 2000; Sanderson, 2000; Miklós & Podani, 2004; Navarro-Alberto & Manly, 2009; Ulrich *et al.*, 2009).

Most notably, the metrics and algorithms differ in Type I and II error rates (Gotelli, 2000) where Type I error occurs when a randomly associated pair of species is incorrectly identified as being either positively or negatively associated and a Type II error occurs when a truly positively/negatively associated pair is incorrectly identified as being randomly associated. In part, Type I and II errors may be caused by the randomization process itself, if the algorithm does not randomize the species presence–absence matrix in an unbiased way. Here, 'unbiased' means that no particular matrix is more likely to occur (in the null set) than are others given the specified conditions of the randomization such as conserved row and column sums. Most randomization algorithms create random matrices such that species incidence rates and richness values of sample sites are conserved. That is, the randomized matrices have the same row and column sums as does the real matrix. Alternatively, sometimes just the species incidence rates are conserved. These two classes of randomization were given the labels of fixed–fixed (F–F) and fixed–equiprobable (F–E) by Gotelli & Ellison (2002a). Bias in randomization can arise when the algorithm reproduces some of the exact same random matrices multiple times or never produces some possible matrices during a given set of iterations (typically 1000–10,000). That is, the iterations do not cover the entire null space (Fig. 1). No algorithm produces every possible matrix during the randomization, except perhaps for the smallest of matrices. This issue of biased randomizations has recently attracted the attention of ecologists and statisticians (Zaman & Simberloff, 2002; Miklós & Podani, 2004; Lehsten & Harmand, 2006; Navarro-Alberto & Manly, 2009; Sanderson *et al.*, 2009; Fayle & Manica, 2010; Gotelli & Ulrich, 2011). In a greater context, the use of data randomization (as a null model and statistical test of inference) is not *distribution-free*. The whole point of the randomization is to produce a null distribution of a test statistic (a metric measuring pair-wise co-occurrence or nestedness within an entire matrix) so that the statistical significance of the observed value of the test statistic can be assessed. But this is also a *potential* source of Type I and II errors, not to mention a point of some disagreement among ecologists.

In this paper, I develop and present a distribution-free and metric-free approach to analysing species co-occurrence patterns. The model is founded upon using basic probability theory to derive exact probabilities that two species should co-occur either more or less frequently than they actually do. The model is strictly analytical; it requires no randomization. Therefore, the probabilistic model can be considered an improvement upon matrix randomization procedures; it does not require a non-biased null distribution that presumably samples all possible matrices. Moreover, the model is an example of the more parsimonious approach of applying basic math and probability theory to answer a research question in lieu of relying upon statistical inference from distributions of a test statistic.

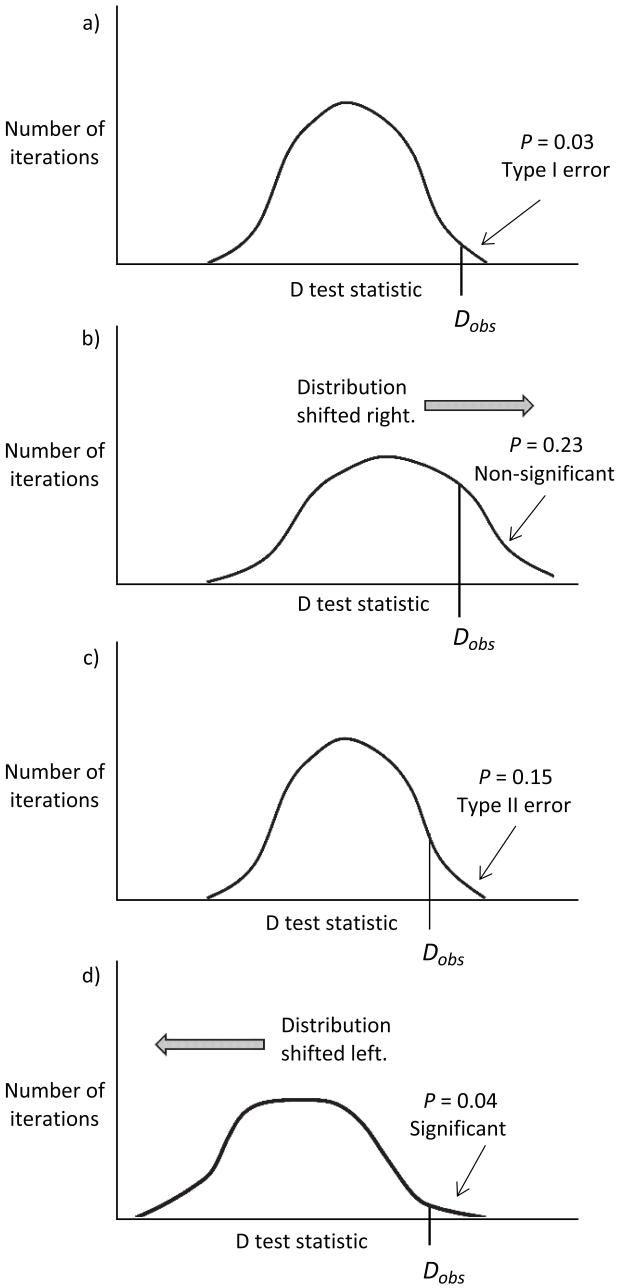

Figure 1 The distribution of a hypothetical test statistic, D, changes when the null space is completely represented. Panel (a) shows a Type I error ($P < 0.05$, the proportion of the null distribution to the right of the observed value of the test statistic); null values to the extreme right are not produced by the randomization algorithm. However, if a more complete and representative null distribution (panel b) is produced, either by a better randomization algorithm or a probabilistic model, D_{obs} is no longer significant and hence a Type I error is avoided. In panel (c) there is a Type II error ($P > 0.05$) because null values to the extreme left are not produced. When they are produced (panel d), D_{obs} is significant. Null distributions produced by some randomization algorithms may not always represent the true shape and location of the entire null space.

THE PROBABILISTIC MODEL OF SPECIES CO-OCCURRENCE

The probabilistic model of species co-occurrence allows one to analytically (i.e. without randomization or simulation) obtain the probability (P) that two selected species co-occur at a frequency either less than (P_{lt}) or greater than (P_{gt}) the observed frequency of co-occurrence. These probabilities can be obtained analytically under the condition where a species probability of occurrence at each site is equal to its observed frequency among all the sites. The following notation is used in developing the model: Q_{obs} is the observed number of sites having both species, n is the total number of sites where either species could occur, N_1 is the number of sites occupied by species 1 and N_2 the number of sites occupied by species 2. (Here, 'site' is used in a general way to denote any sample, survey area, plot, community, island or habitat fragment.) The model is based upon calculating p_j, the probability that species 1 and 2 co-occur at exactly j sites, for $j = 0$ to N. For a given N or set of sites and a given j, there is a limited number of ways that species 1 and 2 can be distributed among the sites so as to maintain or obey the observed N_1 and N_2. Therefore, the probabilistic model is essentially an exercise in combinatorics, or determining the number of ways to select r items out of a total set of n items, symbolized by $\binom{n}{r}$ or $C(n, r)$. If $\max\{0, N_1 + N_2 - N\} \leq j \leq \min\{N_1, N_2\}$ then

$$p_j = \frac{C(N, j) \times C(N - j, N_2 - j) \times C(N - N_2, N_1 - j)}{C(N, N_2) \times C(N, N_1)}. \quad (1)$$

In this equation, $C(N, j)$ is the number of ways that j co-occurrence sites could be arranged among the N sites. $C(N - j, N_2 - j)$ is the number of ways that sites having only species 2 could be arranged among those not already having both species. $C(N - N_2, N_1 - j)$ is the number of ways that sites having only species 1 could be arranged among those not having species 2. These three quantities are multiplied together to get the total number of ways that species 1 and 2 could be distributed among N sites for a given N_1, N_2 and j (Fig. 2), i.e. the numerator of the equation. The denominator simply represents the total number of ways that species 1 and 2 can be arranged among N sites without regard for j. So, the numerator is always a subset of the denominator and hence p_j is always < 1. If $j > N_1$ then $p_j = 0$. Also, if $j < N_1 + N_2 - N$ then $p_j = 0$. P_{lt} is obtained by determining p_j for all $j < Q_{obs}$ and then summing the p_j, $P_{lt} = \Sigma p_j$ for $j = 0$ to $Q_{obs} - 1$. Similarly, $P_{gt} = \Sigma p_j$ for $j = Q_{obs} + 1$ to N. When $j = Q_{obs}$, then $p_j = P_{et}$ the probability that by chance the observed co-occurrence is exactly equal to j. Therefore, $P_{lt} + P_{gt} + P_{et} = 1$ (Table 1).

The probabilistic model is also a statistical method of testing for significant patterns of co-occurrence because the quantities ($P_{lt} + P_{et}$) and ($P_{gt} + P_{et}$) can be used as P-values testing whether species 1 and 2 co-occur significantly less often or significantly more often, respectively, than expected by chance. For example, if $P_{lt} + P_{et} = 0.03$ then species 1 and 2 have a significant negative association, at a significance level of 0.05. If the two species truly are distributed independently of one another then there is only a 3% chance that the two species would co-occur at Q_{obs} or fewer sites.

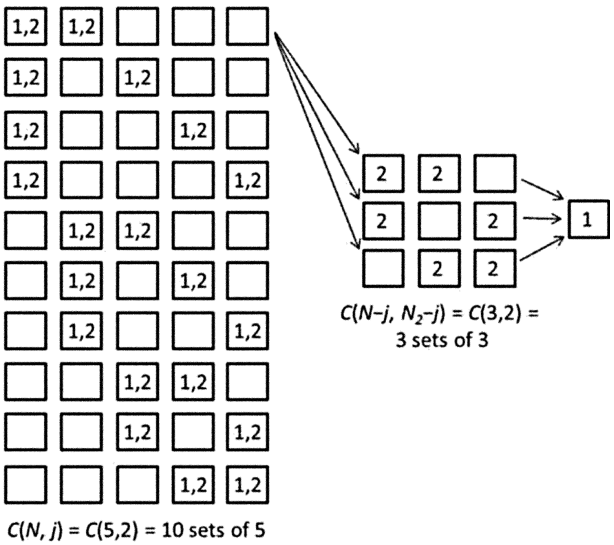

Figure 2 The number of unique combinations of species 1 and 2 distributed among a set of five sites ($n = 5$, sites are represented by boxes) where $N_1 = 3$, $N_2 = 4$ and $j = 2$. The first set of 10 combinations shows all possible arrangements of the two sites having both species, $C(N, j) = C(5,2) = 10$. Note that in each of the 10 combinations, there are three sites (empty boxes) that have not been assigned ($N - j = 3$). For each of the 10 combinations, there are three ways of arranging species 2 among those three empty sites or boxes, $C(N - j, N_2 - j) = C(3,2) = 3$. For each of those combinations, there is only one empty site ($N - j - N_2 - j$ also written as $N - N_2 = 1$) and thus only one way of placing species 1, $C(N - N_2, N_1 - j) = C(1,1) = 1$. Multiplying these quantities together, there is a total of 30 combinations possible. In this example, all sites have either species 1 or species 2. However, this need not be the case, N could be any number. In this example, if $n = 6$, there are 180 possible combinations.

Note that the probabilistic model is not equivalent to a binomial exact test. In that test, the probability of an event occurring (i.e. a 'success') in a 'success/failure' trial is used to obtain the probability of j successes in N trials. If a 'success' is defined as the co-occurrence of species 1 and 2 then the probability of success is equal to $P(1,2)$ or the probability that species 1 and 2 co-occur at a given site, $P(1,2) = N_1/N \times N_2/N$ (Bowers & Brown, 1982; Veech, 2006). In the binomial exact test, $p_j = C(N, j) \times P(1,2)^j \times [1 - P(1,2)^{N-j}]$ which is not equivalent to equation 1. The difference between the binomial exact test and the probabilistic model is probably due to species co-occurrence data violating assumptions of the former, such as independence of trials (sampling sites) and the possibility that species co-occurrence at a site (with outcomes 'success' or 'failure') does not truly represent a Bernoulli trial. The binomial exact test does not apply to analysing pairwise co-occurrence patterns.

A substantial benefit of the probabilistic model (as a statistical method) is that it completely eliminates one major source of Type I and II errors in the testing of species co-occurrence

Table 1 Calculation of co-occurrence probabilities (P_{lt}, P_{gt}, and P_{et}) for two species in a hypothetical example where there are 40 sampling sites, species 1 is found at 10 sites, and species 2 is found at 25 sites ($n = 40$, $N_1 = 10$, $N_2 = 25$).

Number of co-occurrence sites (j)	P_j	ΣP_j	$P_j \times j$
0	0.000004	0.000004	0
1	0.0001	0.0002	0.0001
2	0.0023	0.0024	0.005
3	0.0175	0.0199	0.052
4	0.0747	0.0946	0.299
5	0.1882	0.2828	0.941
6	0.2852	0.5680	1.711
7	0.2580	0.8260	1.806
8	0.1340	0.9600	1.072
9	0.0362	0.9961	0.325
10	0.0039	1.0	0.039

The probability of co-occurrence at exactly j sites (P_j) is calculated for $j = 0$ to 10 sites, $P_j = 0$ for all $j > 10$. If the two species co-occur at three sites ($Q_{obs} = 3$), then $P_{lt} = \Sigma P_j$ for $j = 0$ to 2, $P_{gt} = \Sigma P_j$ for $j = 4$ to 10, and $P_{et} = P_3$. The two species would be negatively associated at a significance level of 0.05 because $P_{lt} + P_{et} = 0.02$. There is only a 2% chance that the two species would co-occur at three or fewer sites if their distributions were truly random of one another. In this example, expected co-occurrence $Q_{exp} = \Sigma(P_j \times j) = 6.25$ sites.

patterns (although other sources still exist). Briefly, in any statistical inference test applied to data, there are at least four sources of Type I and II errors: (1) violations of assumptions about the distribution of the test statistic, (2) sampling error (due the to limited number of samples not being truly representative of the statistical population being sampled), (3) measurement error (error in the way the variables are measured or more generally error in the way the data are collected); and (4) stochastic and/or systematic natural variation in the variables or process being studied such that the variation is not accounted for in any subsequent statistical testing (i.e. the variation is caused by factors that are not being measured or assessed). In the probabilistic model, error from the first source is completely eliminated because there is no simulated or randomized distribution of a test statistic. However, this source of error exists in all other statistical tests of co-occurrence patterns. It is the error assumed when a *P*-value is derived from a null distribution that may not completely or randomly include (without any bias) all possible versions of the presence–absence matrix.

Even with this benefit, the probabilistic model still has the other sources of Type I and II errors (that need to be estimated) as do the randomization-based tests. In addition, as is common to all tests of pairwise co-occurrence patterns, the probabilistic model classifies species pairs to categories (positive, negative or random association) based upon an arbitrary significance level (e.g. $\alpha = 0.05$). This classification can sometimes represent the commission of Type I or II errors. Fortunately, the amount of *possible* Type I error is known; it equals the alpha level. For example, if 100 species pairs are tested and 12 are classified as positive or negative associations based on $\alpha = 0.05$ then five of the pairs *could* represent Type I errors (i.e. misclassifying a random association as either positive or negative) that arise due to error sources 2, 3 or 4. The amount of Type II error (i.e. misclassifying a real positive or negative association as random) cannot be known without knowing the power of the test when applied to the given data, as is true of all statistical inference tests.

Table 2 Assessment of the Type I error rate for the probabilistic model of species co-occurrence.

N	Species pairs	Mean ES	Max ES	Type I error rate
10	23,235	0.57	2.5	0.013
20	27,640	0.78	4.1	0.032
50	28,739	1.15	5.9	0.049
100	28,920	1.45	10.3	0.055

The table presents properties (species pairs, mean and maximum effect sizes) of the simulated random data sets and the Type I error rate for $\alpha = 0.05$. Effect size (ES) is defined as the absolute difference between observed and expected co-occurrence. Type I error was assessed as the proportion of the simulated randomly associated species pairs that the model revealed as incorrectly being either positively or negatively associated. n = number of sampling sites. For a given data set, the number of species pairs examined may not equal the total number of pairs [$C(241,2) = 28,920$] because pairs with expected co-occurrence < 1.0 were not examined.

In order to quantify Type I and II error rates for the probabilistic model, I simulated hypothetical data sets (species presence–absence matrices) that represented random associations between species, and other data sets that represented positive and negative associations. The random data sets allowed for the estimation of Type I error rate as the proportion of species pairs that were found to have a significant positive or negative association. The structured data sets were used to estimate Type II error rate as the proportion of species pairs that had a non-significant (random) association according to the probabilistic model. Each data set contained 241 species for a total of 28,920 pairs, and either 10, 20, 50 or 100 sampling sites. Thus, there were eight data sets, four random and four structured. The structured data sets were built from the random ones by inserting blocks of zeros and ones. The species within each data set represented a wide range of incidence values, from relatively rare species to very common, with most species having an intermediate frequency of occurrence, *P(i)*, among the sampling sites. Mean *P(i)* over all species in a data set was between 0.45 and 0.55, with *P(i)* values ranging from 0.1 to 0.9 within each data set. Thus, the eight data sets were similar except for either having structure or not and differing in the number of sampling sites.

As expected, the Type I error rate was very close to the nominal alpha level (0.05) for all data sets (Table 2). Even with this minimal level of Type I error, an adjustment of the alpha level might be desired when testing dozens or more species so as to further reduce the 'study-wide' Type I error rate. The Type II error rate (β) was also relatively low, as indicated by the effect sizes needed in order to get $\beta = 0$. Effect size was measured as

Table 3 Assessment of the Type II error rate (β) for the probabilistic model of species co-occurrence.

	Simulated data			Effect size for given Type II error rate		
N	Species pairs	Mean ES	Max ES	$\beta_{0.05} = 0$	$\beta_{0.0001} = 0$	$\beta_{0.05} = 1$
10	25 157	0.84	2.5	> 1.6	None	< 1.6
20	27 498	1.32	5.0	> 2.2	> 4.2	< 1.5
50	28 757	3.32	10.4	> 3.3	> 6.8	< 2.0
100	28 920	6.95	21.5	> 4.3	> 9.5	< 2.4

The table presents properties of the simulated structured data sets and the effect size (ES) required to manifest a given Type II error rate for $\alpha = 0.05$ and 0.0001. Effect size is defined as the absolute difference between observed and expected co-occurrence. As with all statistical tests, the power of the model is $1 - \beta$. n = number of sampling sites. For a given data set, the number of species pairs examined may not equal the total number of pairs [$C(241,2) = 28{,}920$], because pairs with expected co-occurrence < 1.0 were not examined.

the absolute difference between observed and expected co-occurrence (measured in numbers of sites) for a pair of species. Depending on the total number of sampling sites, effect size could be as low as 1.6 to 4.3 sites at $\alpha = 0.05$ and 4.2 to 9.5 sites at $\alpha = 0.0001$ while achieving $\beta = 0$ and essentially a statistical test with power ≈ 1 (Table 3). To put this in perspective, even with a conservative $\alpha = 0.0001$ and 100 sampling sites, the difference in observed and expected co-occurrence would need to equal only 10% (9.5 sites) of the total number of sites in order for the probabilistic model to reveal a significant positive or negative association. However, with fewer sites, the model is not as powerful. For data sets of 10 or fewer sites, the model does not have much power regardless of alpha level. Also, in general, the difference in observed and expected co-occurrence must be at least two sites in order for the model to have any power ($\beta \neq 1$), regardless of the number of sampling sites (Table 3). Overall, the power of the probabilistic model compares very well with Gotelli & Ulrich's (2010) Bayesian approach that has power ranging from 0.41 to 1.

The probabilistic model shares an important feature with most matrix randomization procedures (e.g. F–E and F–F algorithms), all species can potentially occur at all sites. That is, any site included in the set of *N* (probabilistic model) or included as a row in a species presence–absence matrix is 'eligible' for receiving the species during the randomization process. Randomization algorithms that conserve the observed species incidence values (column sums in site × species matrices) produce null matrices in which each species probability of occurrence is $P(i) = N_i/N$, just as in the probabilistic model. More precisely, the probabilistic model is the analytical analogue of F–E randomization algorithms. That is, for a given species, the number of occurrences among sites is the same in observed and randomized data (fixed species incidences), but number of species at a given site is not fixed. In F–E algorithms and the probabilistic model, sites may have the same, more or fewer species than what holds for the observed data. Therefore, F–E algorithms and the probabilistic model 'assign' species to sites without assuming that some sites may be more likely to have the species than are others. On the other hand, F–F algorithms make this assumption; these algorithms maintain fixed species richness values at sites (row sums) during the randomization process. Thus, sites differ in their probabilities of receiving a species based directly on their species richness value.

Such site-specific 'colonization probabilities' cannot be directly incorporated into the probabilistic model (i.e. equation 1 cannot be modified to include such probabilities). However, it is possible to apply the model in a way that takes into account the possibility that sites vary in their probability of containing (i.e. being colonized by) species *i*. The researcher can decide a priori to group sites into subsets where sites share the same or very similar colonization probabilities. Sites may often differ in species richness, area that is being sampled, habitat composition and other environmental variables as well as isolation or distance from a 'mainland' source of colonizing species (in the context of island biogeography dynamics). These differences can affect colonization probabilities, but they can be controlled by applying the model separately to subsets of similar sites. For instance, the grouping variable could be site richness itself (essentially duplicating a F–F algorithm), area of sampling site or habitat of a particular type. Of course, a researcher may have controlled for such factors in the study design phase (e.g. Gotelli & Ellison, 2002a) such that all sites can be used in a single application of the probabilistic model.

In many instances, it may be difficult to know a priori whether sites actually do differ in their probabilities of containing a given species. Often, the species distribution data (presence–absence matrix) are used to model these differences as in F–F algorithms that use the observed species richness value to determine how many species a site should receive during the randomization process. There may be good reason to randomize species presence–absence matrices in this way (Gotelli & Graves, 1996), but this restriction need not apply to analyses that are testing for non-random *pairwise* species associations across the entire suite of sampling sites where the two species could potentially co-occur.

EXAMPLES OF MODEL APPLIED TO REAL DATA SETS

I applied the probabilistic model to 10 sets of published species presence–absence data that were previously evaluated for sig-

Table 4 Number of positive, negative and random species associations for data sets analysed with the probabilistic model of species co-occurrence.

Data set	Species	Sites	Positive	Negative	Random	Percentage non-random
Galapagos Island finches	13	17	14	1	49	23.4
Great Basin Desert rodents	16	39	10	6	39	29.1
Mojave Desert rodents	12	18	0	1	36	2.7
Sonoran Desert rodents	23	38	5	3	30	21.1
Puerto Rican *Anolis* lizards	8	11	1	1	19	9.5
Jamaican *Anolis* lizards	6	9	0	0	14	0
New England forest ants	37	22	24	12	224	13.8
New England bog ants	24	22	2	0	55	3.5
Polish island beetles	71	17	58	10	770	8.1
Greek island isopods	56	14	5	6	913	1.1

The model was only applied to those species pairs whose expected number of co-occurrences $[P(1,2) \times N] > 1.0$. A significance level of $\alpha = 0.05$ was used to classify species associations.

nificant pair-wise species associations. These data sets were finches on the Galapagos Islands (Sanderson, 2000), rodents in three major deserts of North America (Bowers & Brown, 1982; Brown & Kurzius, 1987; Patterson & Brown, 1991; Fox & Brown, 1993), *Anolis* lizards on Puerto Rico and Jamaica (Haefner, 1988), forest and bog ants in New England (Gotelli & Ellison, 2002b), carabid beetles on islands in lakes of Poland (Ulrich & Zalewski, 2006; Gotelli & Ulrich, 2010) and terrestrial isopods on Greek islands in the Aegean Sea (Sfenthourakis *et al.*, 1999, 2006). Together these data sets vary considerably in taxonomic group, numbers of species and sampling sites (Table 4) and in the way that the original data were collected. They represent spatial scales ranging from landscape to biogeographic, i.e. sampling sites were separated by tens to hundreds of kilometres. In addition, the data sets were selected so that the assumption of a species being equiprobable among sites was likely to be obeyed.

For most data sets, the probabilistic model revealed instances of positive and negative species associations, although positive associations were typically more common than negative associations (Table 4). Relatively equal numbers of positive and negative associations were found for the three desert rodent data sets and the Greek isopod data set (Table 4). Some of the results contrast sharply with previous analyses of the same data sets. For rodents, Bowers & Brown (1982) reported 75, 67 and 88 negatively associated pairs along with 41, 37 and 45 positively associated pairs in the Great Basin, Mojave and Sonoran Deserts. The probabilistic model revealed 10 or fewer positive or negative associations in each rodent data set. For their ant data sets, Gotelli & Ellison (2002a) did not conduct any pairwise co-occurrence tests, although they reported large matrix-wide *C*-scores (i.e. an excess of negative associations) for forest ants. The probabilistic model also revealed some negative associations between forest ant species, although these were fewer than the positive associations (Table 4).

There was also some general agreement between the probabilistic model and some previous analyses. For Galapagos Island finches, Sanderson (2000) reported 13 'anomalous' pairings, described as those occurring more or less often than chance expectation. The probabilistic model revealed a similar number, 15 non-random associations (mostly positive). For bog ants, Gotelli & Ellison (2002a) report small or non-significant matrix-wide *C*-scores (i.e. indicating positive or random associations). Similarly, the model found only two non-random species pairs. For *Anolis* lizards, Haefner (1988) did not test for pair-wise co-occurrence, but states that 'the non-random models did not differ from the random models' for sites on Jamaica. Again, this conclusion is consistent with the probabilistic model; it did not reveal any instances of non-random associations among Jamaican *Anolis* (Table 4). For carabid beetles in Poland, Gotelli & Ulrich (2010) reported 11 negative associations which closely matches the 10 indicated by the probabilistic model; the method used by Gotelli & Ulrich (2010) did not allow for identifying positive associations. Lastly, Sfenthourakis *et al.* (2006) found 21 positive and 24 negative species associations for isopods on Greek islands, similar to the 1:1 positive:negative ratio of the probabilistic model. They tested 1225 pairs so their overall percentage of non-random pairs (4%) was slightly greater than that found by the probabilistic model (1.1%).

For each data set analysed in the present study, the model revealed many 'random' species associations. There are two possible categories of 'random' association. First, random associations might be the result of insufficient statistical power and represent Type II errors that arise from having a low alpha level. In other words, species 1 and 2 get classified as a random association (because $P > 0.05$) although the difference between observed and expected co-occurrence is substantial [expected co-occurrence, $Q_{exp} = \Sigma(P_j \times j)$; Table 1]. This may explain some of the random associations in the data sets that had relatively low numbers of sampling sites (Table 4). Indeed, among the 10 data sets, there is a positive relationship ($r = 0.74$) between number of sampling sites and the percentage of species pairs classified as non-random. As with the hypothetical simulated data, this suggests that the probabilistic model (as a statistical test) has diminished power with fewer sampling sites. This is also true for all other statistical tests of co-occurrence and nest-

edness. Second, the association between two species may be truly random, with observed ≅ expected co-occurrence. In all the data sets analysed, many of the random associations represented species pairs whose observed and expected co-occurrences did not differ by more than a few sampling sites. However, the probabilistic model will also classify as 'random' two widespread species that occur at all, or a vast majority of, sampling sites. Sfenthourakis *et al.* (2006) also noted this potential outcome in their randomization procedure, but suggested that it is a bias against finding positive associations rather than a definitive demonstration of a truly random association. The best way to resolve the difference between random and positive would be to sample more sites.

THE FUTURE OF SPECIES CO-OCCURRENCE ANALYSES

The probabilistic model could represent a major step forward in the analysis of species co-occurrence patterns. It is a different approach from that used in all previous analyses of pairwise co-occurrence and nestedness of presence–absence matrices. The probabilistic model is computationally simpler and more direct than previous methods that rely on data randomization and metrics (e.g. checkerboards, *C*-score, matrix temperature) that may not be intuitive. The metric used by the probabilistic model (number of sites where two species co-occur) could not be any more direct and intuitive. In addition, results from applying the model to separate data sets can be standardized to allow for meaningful comparison. Standardized effect sizes are often used to compare the output of different null models (Ulrich & Gotelli, 2007; Veech, 2012) where standardized effect size is calculated as the observed – expected value (of a test statistic or metric) divided by the standard deviation of the null distribution. Because there is no null distribution produced in the probabilistic model, the difference between observed and expected co-occurrence is 'standardized' on the number of sampling sites (N). Effect size standardized in this way ranges from −1 to 1 and controls for the number of sampling sites. This could be a useful metric to compare among different species pairs whose co-occurrence is based on analysing data sets where N varies.

In studies of species co-occurrence, the current trend is to analyse pairs of species instead of entire matrices (e.g. Sfenthourakis *et al.*, 2006; Veech, 2006; Sanderson *et al.*, 2009; Gotelli & Ulrich, 2010). In this regard, the probabilistic model could quickly become widely applied. The model also has some flexibility in allowing the user to test particular hypotheses. For instance, the set of sites that makes up N (and its numerical value) affects the probabilities determined by the probabilistic model. The user should decide a priori how to define this set based on the null hypothesis being tested. For example, a test of the effect of broad-scale evolutionary factors (geographic barriers, vicariant events) on co-occurrence might include sites within and outside the geographic range of species 1 but within range of species 2 (and vice versa). In this way, one might be interested in testing whether a given amount of sympatry produces a negative association between two species. In a different study, the goal might be to test the role of a present-day environmental factor (e.g. soil type) in either promoting or preventing co-occurrence, in which case only sites with the particular soil type (and within the geographic range of both species) would be included in the set defined by N.

As currently formulated, the probabilistic model does not account for imperfect detection of species. That is, the model does not allow for the possibility of false absences or failing to record a species when it is actually present at a site. In practice, this neglect of false absences probably does not bias the model toward finding more positive and fewer negative associations (or vice versa). False absences (and false presences) in the data simply represent a form of measurement error and a source of Type I and II errors that could be shared by all methods of analysing co-occurrence. However, a recent extension of occupancy modelling allows for the estimation of species detection probabilities as they might be influenced by a variety of factors including the presence of other species (MacKenzie *et al.*, 2004; Richmond *et al.*, 2010). This method could prove useful in analysing co-occurrence patterns in large data sets of tens to hundreds of species. However, to date, the method has been applied to test for co-occurrence in studies involving only a few species occurring among a set of sites that were repeatedly surveyed (Bailey *et al.*, 2009; Richmond *et al.*, 2010; Waddle *et al.*, 2010).

Regardless of the particular analysis that is used, future studies of species co-occurrence, assembly processes, species interactions and distributional patterns should attempt to make a priori predictions of the species pairs that should be (according to theory or the test hypothesis) positively, negatively and randomly associated (Gotelli & Ulrich, 2010). This species-specific prediction has generally been missing in most previous studies that tested for overall structure in presence–absence matrices or tested for significant pair-wise associations among all possible species pairs. An additional benefit is that focused hypothesis testing will also reduce the number of 'study-wide' Type I errors. The probabilistic model of species co-occurrence is conceptually intuitive and easy to apply; it could become very useful in analysing pair-wise co-occurrence patterns.

ACKNOWLEDGEMENTS

I thank Pedro Peres-Neto for helping me improve the presentation of the model and for suggesting the power analysis. Nick Gotelli and two anonymous referees provided critical and insightful comments on previous versions of this manuscript. I thank them and also recognize that they may have alternative views on the best approach for analysing co-occurrence patterns. Computer code and an executable code file are available from the author for conducting bulk runs of the probability model.

REFERENCES

Bailey, L.L., Reid, J.A., Forsman, E.D. & Nichols, J.D. (2009) Modeling co-occurrence of northern spotted and barred owls: accounting for detection probability differences. *Biological Conservation*, **142**, 2983–2989.

Bowers, M.A. & Brown, J.H. (1982) Body size and coexistence in desert rodents: chance or community structure? *Ecology*, **63**, 391–400.

Brown, J.H. & Kurzius, M.A. (1987) Composition of desert rodent faunas: combinations of coexisting species. *Annales Zoologici Fennici*, **24**, 227–237.

Chase, J.M. & Myers, J.A. (2011) Disentangling the importance of ecological niches from stochastic processes across scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2351–2363.

Collins, M.D., Simberloff, D. & Connor, E.F. (2011) Binary matrices and checkerboard distributions of birds in the Bismarck Archipelago. *Journal of Biogeography*, **38**, 2373–2383.

Connor, E.F. & Simberloff, D. (1979) The assembly of species communities: chance or competition? *Ecology*, **60**, 1132–1140.

Connor, E.F. & Simberloff, D. (1983) Interspecific competition and species co-occurrence patterns on islands: null models and the evaluation of evidence. *Oikos*, **41**, 455–465.

Diamond, J.M. (1975) Assembly of species communities. *Ecology and evolution of communities* (ed. by M.L. Cody and J.M. Diamond), pp. 342–444. Harvard University Press, Cambridge, MA.

Diamond, J.M. & Gilpin, M.E. (1982) Examination of the 'null' model of Connor and Simberloff for species co-occurrences on islands. *Oecologia*, **52**, 64–74.

Ellwood, M.D.F., Manica, A. & Foster, W.A. (2009) Stochastic and deterministic processes jointly structure tropical arthropod communities. *Ecology Letters*, **12**, 277–284.

Fayle, T.M. & Manica, A. (2010) Reducing over-reporting of deterministic co-occurrence patterns in biotic communities. *Ecological Modelling*, **221**, 2237–2242.

Fox, B.J. & Brown, J.H. (1993) Assembly rules for functional groups in North American desert rodent communities. *Oikos*, **67**, 358–370.

Gilpin, M.E. & Diamond, J.M. (1982) Factors contributing to non-randomness in species co-occurrences on islands. *Oecologia*, **52**, 75–84.

Gotelli, N.J. (2000) Null model analysis of species co-occurrence patterns. *Ecology*, **81**, 2606–2621.

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

Gotelli, N.J. & Ellison, A.M. (2002b) Biogeography at a regional scale: determinants of ant species density in New England bogs and forests. *Ecology*, **83**, 1604–1609.

Gotelli, N.J. & Graves, G.R. (1996) *Null models in ecology*. Smithsonian Institution Press, Washington, DC.

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

Gotelli, N.J. & Ulrich, W. (2011) Over-reporting bias in null model analysis: a response to Fayle and Manica (2010). *Ecological Modelling*, **222**, 1337–1339.

Haefner, J.W. (1988) Assembly rules for Greater Antillean *Anolis* lizards: competition and random models compared. *Oecologia*, **74**, 551–565.

Hui, C. (2009) On the scaling patterns of species spatial distribution and association. *Journal of Theoretical Biology*, **261**, 481–487.

Kennedy, C.R. (2009) The ecology of parasites of freshwater fishes: the search for patterns. *Parasitology*, **136**, 1653–1662.

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

Lehsten, V. & Harmand, P. (2006) Null models for species co-occurrence patterns: assessing bias and minimum iteration number for the sequential swap. *Ecography*, **29**, 786–792.

MacKenzie, D.I., Bailey, L.L. & Nichols, J.D. (2004) Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology*, **73**, 546–555.

Miklós, I. & Podani, J. (2004) Randomization of presence–absence matrices: comments and new algorithms. *Ecology*, **85**, 86–92.

Navarro-Alberto, J.A. & Manly, B.F.J. (2009) Null model analyses of presence–absence matrices need a definition of independence. *Population Ecology*, **51**, 505–512.

Patterson, B.D. & Brown, J.H. (1991) Regionally nested patterns of species composition in granivorous rodent assemblages. *Journal of Biogeography*, **18**, 395–402.

Richmond, O.M.W., Hines, J.E. & Beissinger, S.R. (2010) Two-species occupancy models: a new parameterization applied to co-occurrence of secretive rails. *Ecological Applications*, **20**, 2036–2046.

Sanderson, J.G. (2000) Testing ecological patterns: a well-known algorithm from computer science aids the evaluation of species distributions. *American Scientist*, **88**, 332–339.

Sanderson, J.G., Diamond, J.M. & Pimm, S.L. (2009) Pairwise co-existence of Bismarck and Solomon landbird species. *Evolutionary Ecology Research*, **11**, 771–786.

Sanderson, J.G., Diamond, J. & Pimm, S.L. (2011) Response to Collins *et al.* (2011). *Journal of Biogeography*, **23**, 2397.

Sfenthourakis, S., Giokas, S. & Mylonas, M. (1999) Testing for nestedness in the terrestrial isopods and snails of Kyklades Islands (Aegean Archipelago, Greece). *Ecography*, **22**, 384–395.

Sfenthourakis, S., Tzanatos, E. & Giokas, S. (2006) Species co-occurrence: the case of congeneric species and a causal approach to patterns of species association. *Global Ecology and Biogeography*, **15**, 39–49.

Ulrich, W. & Gotelli, N.J. (2007) Null model analysis of species nestedness patterns. *Ecology*, **88**, 1824–1831.

Ulrich, W. & Zalewski, M. (2006) Abundance and co-occurrence patterns of core and satellite species of ground beetles on small lake islands. *Oikos*, **114**, 338–348.

Ulrich, W., Almeida-Neto, M. & Gotelli, N.J. (2009) A consumer's guide to nestedness analysis. *Oikos*, **118**, 3–17.

Veech, J.A. (2006) A probability-based analysis of temporal and spatial co-occurrence in grassland birds. *Journal of Biogeography*, **33**, 2145–2153.

Veech, J.A. (2012) Significance testing in ecological null models. *Theoretical Ecology*, doi 10.1007/s12080-012-0159-z in press.

Waddle, J.H., Dorazio, R.M., Walls, S.C., Rice, K.G., Beauchamp, J., Schuman, M.J. & Mazzotti, F.J. (2010) A new parameterization for estimating co-occurrence of interacting species. *Ecological Applications*, **20**, 1467–1475.

Weiher, E. & Keddy, P. (1999) Assembly rules as general constraints on community composition. *Ecological assembly rules, perspectives, advances, retreats* (ed. by E. Weiher and P. Keddy), pp. 251–271. Cambridge University Press, Cambridge, UK.

Zaman, A. & Simberloff, D. (2002) Random binary matrices in biogeographical ecology – instituting a good neighbor policy. *Environmental and Ecological Statistics*, **9**, 405–421.

BIOSKETCH

Joseph Veech is broadly interested in the ecological processes that produce patterns of species diversity, co-occurrence and population change at a wide range of spatial scales. This interest also includes the development of novel statistical approaches and techniques.

Editor: Pedro Peres-Neto