Forum

Oikos 000: 001–010, 2011
doi: 10.1111/j.1600-0706.2011.20301.x

Subject Editor: Paulo Guimaraes Jr. Accepted 8 September 2011

Statistical challenges in null model analysis

Nicholas J. Gotelli and Werner Ulrich

N. J. Gotelli, Dept of Biology, Univ. of Vermont, Burlington, VT 05405, USA. – W. Ulrich (ulrichw@umk.pl), Dept of Animal Ecology, Nicolaus Copernicus Univ. in Toruń, Gagarina 9, PL-87-100 Toruń, Poland.

This review identifies several important challenges in null model testing in ecology: 1) developing randomization algorithms that generate appropriate patterns for a specified null hypothesis; these randomization algorithms stake out a middle ground between formal Pearson–Neyman tests (which require a fully-specified null distribution) and specific process-based models (which require parameter values that cannot be easily and independently estimated); 2) developing metrics that specify a particular pattern in a matrix, but ideally exclude other, related patterns; 3) avoiding classification schemes based on idealized matrix patterns that may prove to be inconsistent or contradictory when tested with empirical matrices that do not have the idealized pattern; 4) testing the performance of proposed null models and metrics with artificial test matrices that contain specified levels of pattern and randomness; 5) moving beyond simple presence–absence matrices to incorporate species-level traits (such as abundance) and site-level traits (such as habitat suitability) into null model analysis; 6) creating null models that perform well with many sites, many species pairs, and varying degrees of spatial autocorrelation in species occurrence data. In spite of these challenges, the development and application of null models has continued to provide valuable insights in ecology, evolution, and biogeography for over 80 years.

> *'A null model is a pattern generating model that is based on randomization of ecological data or random sampling from a known or imagined distribution. The null model is designed with respect to some ecological or evolutionary process of interest'.* (Gotelli and Graves 1996)

From its origins in the analysis of species/genus ratios (Järvinen 1982), there is a long history of using null models to analyze patterns and test hypotheses in ecology, evolution and biogeography (Harvey et al. 1983). Although the general controversy in the 1970s over null models and competition has died down (Gotelli and Graves 1996), there are still many disputed aspects of testing and implementing null models. In this paper, we review some of the more recent challenges and controversial issues in the implementation and interpretation of null models in ecology. We focus primarily on the use of null models in biogeography, ecology, and macroecology.

1. Hypothesis testing and constraints in null model analysis

Classical Pearson–Neyman hypothesis testing (Graves 1978) addresses the dichotomy between a null hypothesis (H_0) and its alternative (H_1). If these hypotheses are mutually exclusive and collectively exhaustive, then the probability that H_0 is true, given the data ($P(H_0|\text{data})$), is $P(H_0) = 1 - P(H_1)$. The null hypothesis varies depending on the details of the test, but it is often a parsimonious expectation that the data are drawn from a single distribution, so that any patterns in the data arise only from random sampling processes. The alternative hypothesis is that patterns in the data are not the result of random variation generated by H_0. Erroneous rejection of H_0 occurs with probability α and represents a type I statistical error. Conversely, erroneous acceptance of a false null hypothesis is a type II error and occurs with probability β. The quantity $1 - \beta$ is the power of the test, the probability of correctly rejecting H_0 given that it is false (Sokal and Rohlf 1995). Following standard statistical procedure, we equate the calculated *P*-value (Fisher's evidential *P*-value ($P(\text{data}|H_0)$)) with α (but see Hubbard and Bayari 2003).

In ecological null model analysis, *'Null hypotheses entertain the possibility that nothing has happened, that a process has not occurred, or that change has not been produced by a cause of interest'* (Strong 1980), so that H_1, the alternative, effectively isolates the process of interest. This description would seem to imply that the null hypothesis is a stochastic process-based model that excludes a particular mechanism (Roughgarden 1983). For example, the neutral model (Hubbell 2001) and the equilibrium model of island biogeography (MacArthur and Wilson 1963) are two stochastic process-based models, neither of which (in their original formulation) incorporates species interactions (Fig. 1). In practice, however, ecologists have shied away from specifying a particular process model

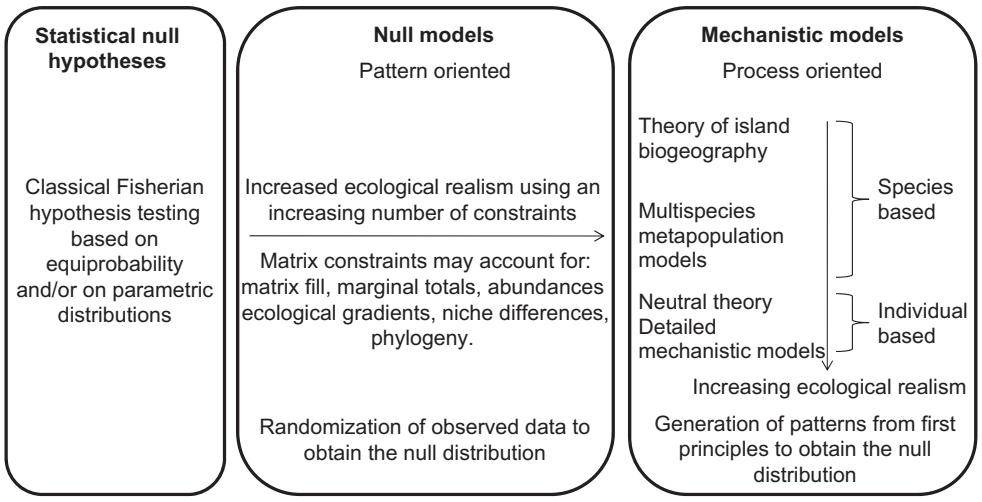

Figure 1. A conceptual model to show the relationships and differences among statistical null hypotheses, null models, and mechanistic models.

because it is usually impossible to collect all the data that would be needed to independently fit the parameters of such a model (Gotelli and McGill 2006). Even relatively simple process-based models, such as the neutral model, have proven extremely difficult to parameterize (Wootton 2005).

Instead, null model analysis specifies a statistical distribution or randomization of the observed data, designed to mimic the outcome of the random process model without specifying or estimating all of its parameters. Figure 1 depicts a gradient of models from statistical testing based on predefined theoretical distributions to mechanistic models based on parameters that specify particular ecological processes. Null models represent an intermediate point in this gradient: they create a null distribution by imposing constraints on randomization to preserve some features of the empirical data, but they do not specify all the parameters in an explicit mechanistic model.

However, insuring that these constraints match the desired properties of the null hypothesis can be difficult. For example, Atmar and Patterson (1993) hypothesized that selective, orderly extinction created patterns of nestedness in the occurrence of species on islands or habitat fragments. This mechanism also implies the alternative hypothesis, so the null hypothesis would presumably include all other mechanisms (such as island size or dispersal ability) that might influence the distribution of species on islands. However, Atmar and Patterson's (1993) null model simply randomized the placement of all species on all islands, implying that all other processes would lead to a pattern in which all species were equally common and all islands were equally suitable.

The challenge here is that this randomization algorithm is too unconstrained, and encompasses more than just the stated null hypothesis. In particular, even if there is no selective extinction, species may differ in their occurrence probabilities (some are common and some are rare), and sites may differ in their suitability (some may be highly suitable and support many species, and some may be unsuitable and support few species). Extensions of the null model approach have tried to specify different statistical universes that encompass these possibilities.

This specification and choice of null model has an important effect on the result. If all sites and species are treated as equivalent, and occurrences are randomized with no constraints, then most empirical matrices show significant nested structure (Wright et al. 1998). If the randomization is constrained to preserve observed matrix row and column totals, then only ~10% of empirical matrices exhibit nestedness (Ulrich and Gotelli 2007a). Patterns of apparent nestedness in many biogeographic data sets may be generated simply by differences in site suitability (apparent in column totals) and species occurrence probabilities (apparent in row totals), but not by an orderly sequence of extinctions.

However, imposing too many constraints on null model algorithms reduces their power and increases the chances of a type II statistical error. This tradeoff between type I and type II errors is inevitable in statistical tests, and is not unique to null model analysis. Our personal preference in null model testing is to favor the control of type I error (Gotelli and Ulrich 2011). There are two reasons for this: first, the philosophical appeal to parsimony is what precipitated the use of null models in the first place: is there any evidence that biogeographic patterns are more extreme than expected by chance (Connor and Simberloff 1979)? Most ecologists would not be satisfied to invoke a biological mechanism when a simple stochastic model that does not incorporate the mechanism can generate the same pattern. The second reason is that biogeographic data are almost always observational and often consist of little information other than the occurrence matrix itself. Statistical inference is less certain without controlled experiments, so it seems prudent to use a conservative approach.

Not all ecologists agree with this philosophy. One argument against imposing too many constraints on null models is that the procedure may become circular. If the biological processes (e.g. competition) affect the constrained elements (e.g. matrix row totals) then the effect of interest has been smuggled into the test, which reduces the sample space and

leads to excessive type II errors (Grant and Abbott 1980, Colwell and Winkler 1984). For this reason, Presley et al. (2010) recently advocated the use of the equiprobable–equiprobable model for testing for patterns of species distributions. However, the poor performance of this algorithm in the context of nestedness (Ulrich and Gotelli 2007a) and species co-occurrence (Gotelli 2000, Ladau 2008), suggests there is a real danger in overestimating the frequency of significant patterns by taking such a liberal approach.

Recently Kullback–Leibler information-based model choices have become popular as complementary approaches to classical hypothesis testing (Akaike 1973, Burnham and Anderson 2002). Information criteria assign probabilities to competing models with different numbers of free parameters and thus allow for a ranking of models from best to worst (Anderson 2008). In the context of null model analysis, we might ask whether information criteria are capable of quantifying the information content of differently constrained null models. However, a simplistic use of information criteria is problematic because we cannot equate the number of null model constraints with the number of free parameters necessary for calculating information metrics. Moreover, null models cannot simply be ranked additively by the number of constraints they contain, but should instead be chosen on the basis of their simplicity, the biological realism of their assumptions, and their performance in benchmark tests with artificial data. Information based approaches might be helpful when comparing more complex versions of stochastic neutral models (Fig. 1). Stochastic mid domain effect models (Rangel and Diniz-Filho 2005), mechanistic models of ecological drift (Hubbell 2001), and multispecies metapopulation models (Holt 1997) contain many potential parameters and could be compared with information criteria, although all of these approaches are a bit too complex to qualify as simple null models.

2. Metrics for defining pattern in null model analysis

Once a null model algorithm has been specified, the pattern in a matrix needs to be quantified with an appropriate metric. The distribution of the metric can then be estimated by simulating a large number of null matrices, and calculating the metric for each. As in other randomization tests, it is this distribution that is statistically compared to the single value of the metric calculated for the empirical data matrix (Manly 2009).

Recently, some researchers (Ladau 2008, Ladau and Schwager 2008) have argued that ecologists should use more formal criteria for testing within the Pearson–Neyman framework, including robustness (observed type I error rates should be close to the preselected ones) and bias ($\alpha \neq \beta$) of null model tests. However in much ecological research, implementation of the Pearson–Neyman framework may be problematic. In simple *t*- or *F*-tests, the metric and the associated null distributions (the respective *t*- and *F*-distributions) are naturally linked, because the distribution of the test metric is derived directly from the assumptions of the null hypothesis (e.g. normality, independence). In these cases, the null model assumptions are fully specified, and the formal properties of the associated tests can be investigated, as proposed by Ladau (2008).

However, in most ecological applications, the metric for quantifying pattern and the null distribution for generating random occurrence matrices are uncoupled. Often, investigators create metrics primarily to quantify a pattern and not because they necessarily can be used with a particular null model. As a consequence, some metrics cannot be used with some randomization algorithms. For example, Schluter's (1984) *V* ratio tests for species independence by calculating, from a presence–absence matrix, the ratio of the sum of the variances in species occurrences to the variance of the sum. However, this metric is calculated entirely from row and column totals of the observed matrix, so null model algorithms that are conditioned on these totals cannot be used with the *V*-ratio. Along the same lines, many metrics of similarity, such as the Morisita, Soerensen, and Simpson indices (Baselga 2010), as well as the recent NODF (Almeida-Neto et al. 2008), use column and/or row totals for calculation. A null model that constrains these totals reduces the sample space of these metrics and might bias the performance of the test.

How, then, should these metrics be chosen? In some cases, the choice of a metric arises naturally from the hypothesis being considered. For example, Diamond (1975) hypothesized that species interactions led to certain species pairs that would never co-occur. Counting the number of species pairs that form such 'checkerboard distributions' in a matrix is a natural metric for this question (Gotelli and McCabe 2002). But there are many ways to quantify patterns of species distribution such as nestedness, segregation, aggregation, and species home range coherence. Collapsing ecological patterns into a single metric is challenging, because ecological patterns are inherently multivariate. There may be associations among sets of species, similarities among sets of islands, and perhaps different submatrices that exhibit different kinds of patterns.

However, many metrics of species association represent an average or a sum of values calculated for individual pairs of species. A long-standing objection to null model analyses of competitive interactions is that patterns among particular sets of species will be obscured by a metric that averages over all possible pairs ('dilution effect'; Diamond and Gilpin 1982). One approach is to analyze the distribution of metrics for all of the $S(S-1)/2$ unique species pairs (Sfenthourakis et al. 2006, Gotelli and Ulrich 2010). However, this introduces new problems because there are so many statistical tests that must be conducted. Similar problems arise in genomic and proteomic analysis, in which large numbers of genes are simultaneously screened. Bayesian approaches developed for genomics (Efron 2005) can be adapted to testing for species pairs in ecological analyses (Gotelli and Ulrich 2010). Initial tests for pattern in the entire matrix followed by pair-wise analyses with adjusted *P*-values may be an effective strategy to uncover non-random species pairs and avoid a dilution effect.

A final challenge in defining an appropriate metric is ensuring that the metric uniquely quantifies the pattern of interest, and does not also measure other sorts of patterns. In many cases, metrics are implicitly equated with the patterns they are intended to describe, which can lead to a mismatch

between the formal definition of a pattern and the working definition based on the index.

The history of the nestedness concept (Almeida-Neto et al. 2007, 2008, Ulrich et al. 2009) is a good example of how this mismatch of pattern and metric can cause confusion. Patterson and Atmar (1986) originally defined nestedness as '*that the species comprising a depauperate fauna should constitute a proper subset of those in richer faunas*'. This definition of nestedness focuses on the species composition among sites, but does not consider the site composition among species. A proper nestedness metric should measure the degree to which species poor sites appear to be random subsamples of species rich sites. The discrepancy metric of Brualdi and Sanderson (1999) conforms to this definition (Fig. 2) because it quantifies only discrepancies in nestedness among sites.

However, the original temperature metric first proposed Atmar and Patterson (1993) is based on weighted discrepancies that reflect deviations of both individual sites and of individual species from a pattern of maximum nestedness (Fig. 2). Although the temperature metric performs well in most biogeographical applications (Ulrich and Gotelli 2007a), it measures something more than just deviations of individual sites from nestedness. Recently Almeida-Neto et al. (2007, 2008) introduced a metric, NODF, that is close to the original Atmar and Patterson (1993) definition of nestedness, but regards discrepancies among both sites and species. Even when used with an identical null model algorithm, these three popular metrics have the potential to classify the same matrix (Fig. 2) as nested (discrepancy index), anti-nested (NODF), or random (matrix temperature). Subtle differences in pattern definition can cause significant differences in matrix classification and therefore in the ecological interpretation. Similar contradictory behavior can be found for various diversity and evenness metrics (Tuomisto 2010, Chao et al. 2010, Loehle 2011, Almeida-Neto et al. 2011).

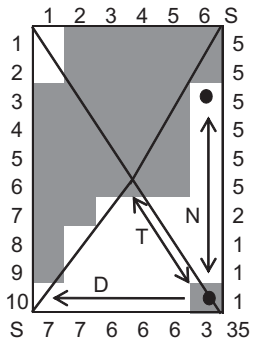

Figure 2. Three popular metrics of nestedness perform differently on a matrix of ten species and six sites. Temperature (T) is a symmetric measure of squared distances from a predefined isocline. Discrepancy (D) counts the minimal number of gaps to be filled to achieve a perfectly filled upper-right part of the matrix. NODF (N) is the averaged number of row and column gaps within the sequence of decreasing marginal totals. Both under the conservative fixed – fixed null model and under the most liberal equiprobable–equiprobable null model, these metrics give contradictory patterns. Discrepancy identifies the matrix as being significantly nested ($D = 3$, $P(H_0) < 0.01$); temperature identifies the matrix as being neither significantly nested nor anti-nested ($T = 27.3$, $P(H_0) > 0.15$); NODF points to significant anti-nestedness ($NODF = 41.1$, $P(H_0) < 0.01$).

3. Benchmark metric and algorithm performance in null model analysis

In null model and Monte Carlo analysis, randomizations are intended to provide a random sample of metric values when H_0 is true. The problem with any such null hypothesis testing is that the rejection of H_0 with probability α does not imply that H_1 is true with probability $1 - \alpha$. Such a conclusion would be allowed only if H_0 and H_1 were mutually exclusive (only one of the hypotheses is true) and collectively exhaustive (there are no other hypotheses possible). Both assumptions are often not fulfilled. Further, it is never certain whether a particular null model sufficiently captures the range of patterns specified by the null hypothesis. For instance, if we want to test whether species pairs have negative associations, our null hypothesis would be that species occurrences are independent, so that associations are random. However, very different types of association might be called random with respect to certain factors. Even with a precisely stated null hypothesis about randomness, it is still uncertain whether an associated randomization adequately approximates the pattern predicted by the null hypothesis (Navarro-Alberto and Manly 2009).

Different algorithms that are reasonable candidates for a null hypothesis may generate different distributions of test metrics. For example, there are at least three reasonable algorithms (knight's tour, swap, sums of squares reduction) that generate randomized matrices retaining marginal totals (Sanderson et al. 1998, Connor and Simberloff 1979, Miklós and Podani 2004), but not all of these algorithms generate a truly random sample of the (large) set of all matrices with the same row and column totals.

For these reasons, it is necessary to expand the traditional Neyman–Pearson testing framework to evaluate the empirical performance of any proposed null model-metric combination against a battery of artificial test matrices (Fig. 3). Although many investigators introduce new null models and metrics, and then apply them to empirical data sets, this is premature (Gotelli 2001). The performance of null models cannot be evaluated by comparing patterns with real data matrices, which contain unknown amounts of structure or randomness. By constructing a set of artificial matrices, investigators can control the amount of signal and noise in the data, and then evaluate the behavior of any candidate null model algorithm and test metric.

We begin this benchmark testing procedure (Fig. 3) by defining a set of x candidate null models. These are statistical randomization algorithms that generate matrices that are generally similar to those that might arise from a specific process-based null model. We also define a set of y candidate metrics. Each metric provides a single number that can be calculated from a matrix and quantifies a particular pattern of interest (such as nestedness or species segregation). The performance of each of the xy metric–algorithm combinations must then be tested against artificial matrices that have specified levels of randomness and structure.

We can imagine a set of 'random' matrices and a set of 'structured' matrices. For the set of artificial random matrices,

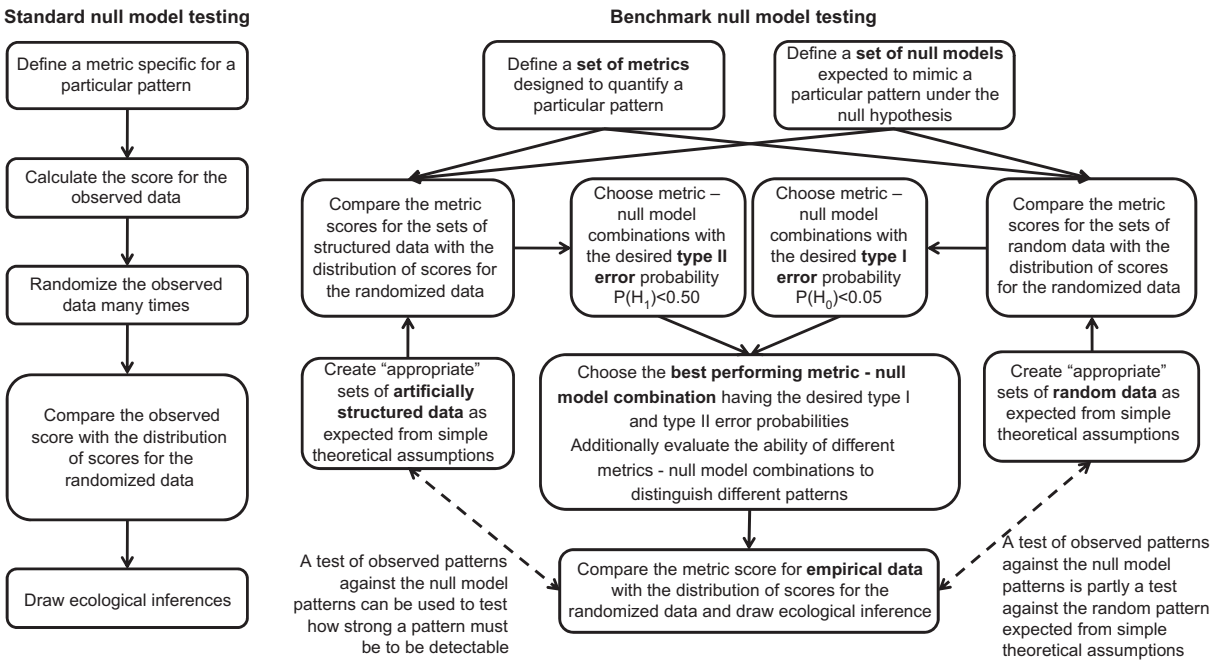

Figure 3. A conceptual model to illustrate the steps of standard null model testing and benchmark testing of proposed null models and metrics.

we want to quantify how frequently each matrix–algorithm combination rejects H_0. If this frequency is too high (by conventional standards, > 0.05), we should avoid this combination because it is prone to type I errors (operationally defined as frequent rejection of H_0 for the set of artificial random test matrices). For the set of artificial structured matrices, we also want to quantify how frequently each matrix–algorithm combination rejects H_0. If this frequency is too low, we should avoid this combination because it is prone to type II errors (operationally defined as infrequent rejection of H_0 for a set of artificially structured test matrices).

How should these test matrices be generated? The artificial random matrices should exhibit properties that are associated with the null hypothesis. If the null hypothesis is that species interactions are not important, the algorithm should place species occurrences in sites independently of other species occurrences and at random. For example, in previous tests (Ulrich and Gotelli 2007a, b) we have used a log-normal distribution of species abundances and a random uniform distribution of species per site to generate random matrices that would be expected with neutral metacommunity dynamics (Hubbell 2001). We randomly varied the parameters in these statistical distributions within specified ranges, as well as the matrix dimensions and fill, to generate a heterogeneous suite of artificial random matrices. 'Random' is here defined by the neutrality assumption of species equivalence and a lack of ecological interactions. These matrices are all random with respect to species co-occurrences, but they also incorporate heterogeneity in species incidences (row totals) and numbers of species per site (column totals).

It is more challenging to construct artificially structured matrices for benchmark testing. The two strategies used have been to begin with a highly organized matrix, and randomly swap elements in the matrix rows to introduce increasing levels of noise and species independence (Gotelli et al. 1997, Gotelli 2000). Alternatively, one can begin with a random matrix and introduce one or a few species pairs with non-random structure (Ulrich and Gotelli 2007a, b). This kind of matrix tests whether the null model algorithm can successfully recover the embedded pattern, or whether the algorithm suffers from a 'dilution effect'. It also reveals how large an effect size is necessary for statistical detection.

This benchmark testing procedure is not fool-proof, and certainly the results depend on the ways in which the artificial random and structured matrices are generated. Even for models that pass this screening, some structured matrices may appear as random (Colwell and Winkler 1984), and some random matrices may cause the null hypothesis to be rejected (Ulrich 2004). If one could specify a particular process-based null model and estimate its parameters, then other kinds of tests may be more powerful and robust, and could be analyzed with more formal statistical methods. But in the absence of such information, the operational definitions of the frequency of type I and type II errors provide reasonable benchmark criteria for evaluating and comparing the performance of different null model algorithms (Fig. 3).

4. Sample size effects in null model analysis

Most data sets for the analysis of metacommunity structure are of small or intermediate size. They contain rarely more than 100 species and/or 100 sites. The limit on the number of sites reflects the considerable time and labor that is needed for field ecologists to sample communities at multiple locations over reasonably large spatial scales. The limit on the number of species reflects the considerable taxonomic expertise needed to correctly identify field samples, and the

conscious or unconscious decisions of collectors to limit their collecting to a particular guild, taxonomic group, or trophic level, rather than attempting a comprehensive survey of an entire food web.

These limitations are easily seen in the data sets that ecologists have laboriously accumulated over the past several decades. For instance only four matrices of the well-known compilation of Atmar and Patterson (1995) had more than 100 sites. However, the recent prominence of macroecological studies based on extensive data bases of gridded maps of terrestrial species occurrence (Rahbek et al. 2007, Keil and Hawkins 2009), as well as small- and large-scale studies of taxon-rich microbial diversity (cf. Green and Bohannan 2006) have led to a substantial increase in the size of data matrices.

But null model analysis may not be well-suited to such large data sets. The general statistical problem is that with very large data sets, the null hypothesis will always be rejected unless the data were actually generated by the null model process itself. So, large data sets may often deviate significantly from null models in which row and column sums are fixed, regardless of whether species occurrences are random or not (Fayle and Manica 2010). This was not a problem in the early history of null model analysis, when ecologists worried that apparent patterns in relatively small data sets might reflect random processes.

A related problem with large matrices is in the statistical analysis of pair-wise species associations. For instance, in a matrix of 50 species there are $50 \times 49/2 = 1225$ distinct species pairs. If all pairs are independent of one another, we still expect with a 1% two-tailed test 12 'non-random' associations just by chance. Thus maximally 24 species (48%) might be involved in false positives. Having 500 species we have already 124 750 pairs and 1248 'significant' pairs just by chance. Thus it is quite probable that each of the 500 species is at least one time engaged in a false positive pair. A simple frequentist null model analysis of pair-wise association is impossible; other methods, for instance Bayesian techniques (Gotelli and Ulrich 2010), are needed.

Simple statistical analyses assume that data are randomly and independently sampled; community structure on isolated islands or habitat patches is regularly treated this way. But many macroecological data sets consist of gridded contiguous observations, and it is not clear that they should be analyzed as a set of random, independent samples. Null models that randomize occurrences within a matrix assume that occurrence probabilities are independent of sample position, so they ignore spatial autocorrelation in species occurrences. For example, null models that randomly place species occurrences in a grid of equiprobable cells will generate a uniform distribution of species richness values. But if the spatial coherence of species individual ranges is preserved in a simple 'spreading dye' null model (Jetz and Rahbek 2001), the distribution of species richness values exhibits a peak near the center of the map (the mid-domain effect; Colwell and Lees 2000), which is very different from a uniform distribution of richness across the domain. The mid-domain effect (MDE) proved to be a surprisingly controversial null model (Colwell et al. 2005). Among other things, critics objected to the random placement of ranges within a bounded range because real ranges reflect species interactions with the environment (Hawkins and Diniz-Filho 2002).

However, the MDE served as a very effective null model because it excluded geographical gradients in historical effects or contemporary climate and demonstrated that species richness gradients can arise entirely from simple geometric constraints (Colwell et al. 2004). These constraints are a realistic alternative to the implicit null hypothesis in many correlative studies where species have no dispersal constraints and can occur in any grid cell within a domain that has appropriate climatic conditions (Gotelli et al. 2009). More recent analyses have used the range cohesion effect embodied in MDE in stochastic models that also include environmental effects (Rahbek et al. 2007).

For very large matrices, and for matrices sampled at large spatial scales, the homogeneity assumption cannot be justified and traditional null models should be applied with caution. Recently Navarro-Alberto and Manly (2009) showed that any difference either in occurrence probabilities of species across sites (non-uniform column degree distributions) or species (non-uniform row degree distributions) causes some degree of spatial autocorrelation. Null models that do not correct for autocorrelation may therefore too often point to non-randomness. To our knowledge, the effect of autocorrelation on matrix structure has not been studied systematically, although Ulrich (2004) demonstrated that a neutral model with limited spatial dispersal can generate binary presence–absence matrices that are statistically segregated. Autocorrelation in species occurrences should cause a tendency towards matrix compartments with regions of higher and lower fill. For large matrices, even very small degrees of autocorrelation will be identified as being significant (Burnham and Anderson 2002).

A second type of autocorrelation is the repetition of submatrix patterns as shown in Fig. 4A. A uniform 100×10 random matrix is not identified as being structured by the common C-score using the fixed – fixed null model. However, repeated juxtaposition of this matrix generates a 100×100 matrix that appears highly structured and is identified as being significantly segregated. Repetition of structure might pose a problem in pseudoreplicated sampling designs (Hurlbert 1984) in which all contiguous sampling plots (or sites in a gridded macroecology map) are used to generate a matrix for pattern analysis.

Autocorrelation poses particular problems in pattern interpretation. Many patterns, for instance patterns of co-occurrence and nestedness, are inevitably linked to non-random cell occupancies and are therefore a special case of autocorrelation. Autocorrelation due to matrix inhomogeneity and due to non-random species associations may be indistinguishable. The development of null models for large matrices that can accommodate a moderate amount of autocorrelation is needed. Such null models should incorporate information on environmental variables that influence occurrence probabilities (Peres-Neto et al. 2001). It might be that large heterogeneous matrices need explicit process based simulations to generate appropriate null distributions (Gotelli et al. 2009).

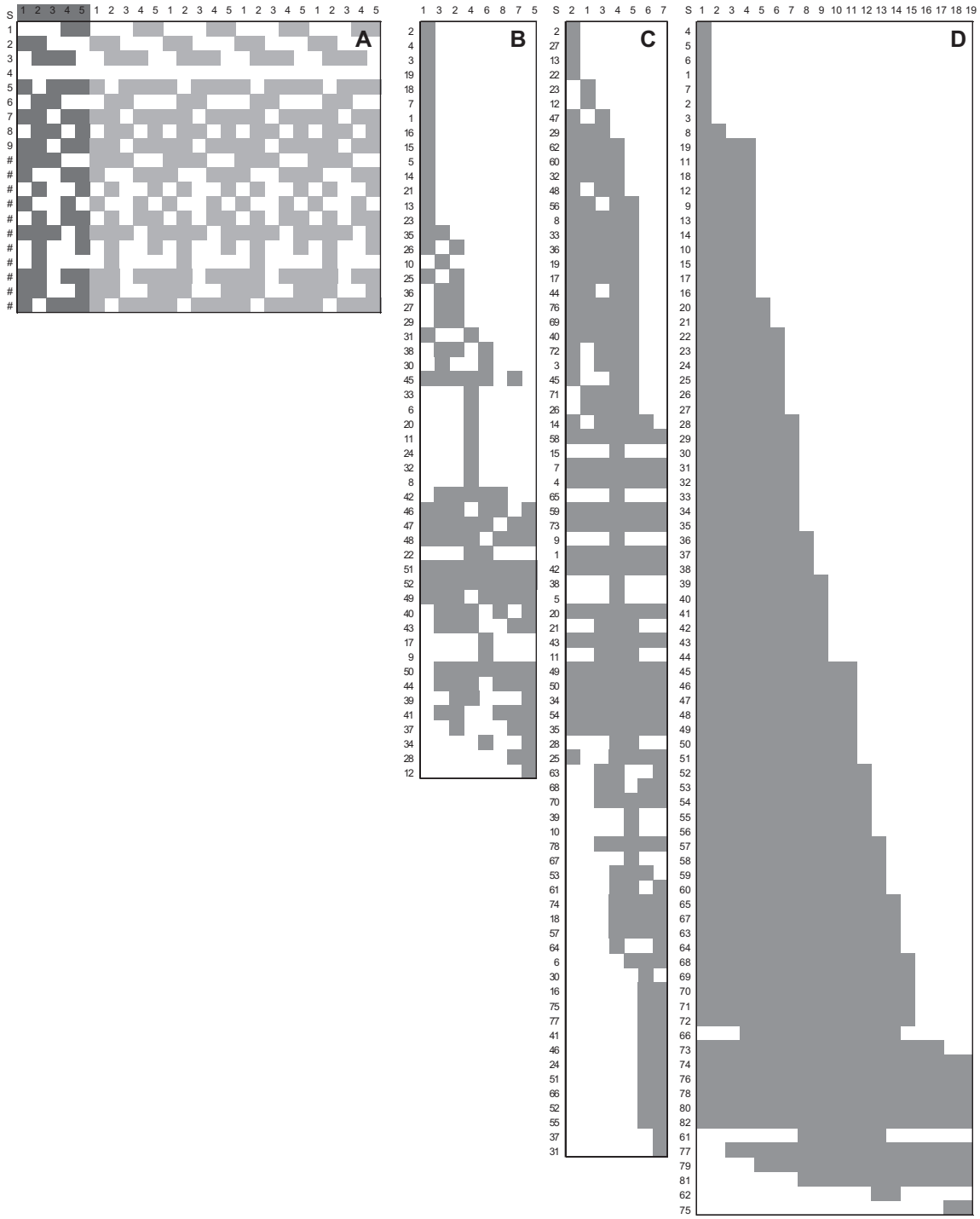

Figure 4. Four examples of matrices with seemingly contradictory patterns as detected by common metrics under the fixed–fixed null model (species occurrences are in grey; all $P(H_0) < 0.001$). (A) A presence–absence matrix of 20 species and five sites (sorted according to marginal totals) generated by a uniform random placement (the highlighted part) is not identified as being structured by the C-score (Stone and Roberts 1990) and the NODF nestedness metric (Almeida-Neto et al. 2008; fixed–fixed null model: $P(H_0) > 0.3$). However, a $5 \times$ replication and juxtaposition of this random matrix generated a 20×25 matrix that was identified by the C-score as being highly segregated ($P(H_0) < 0.001$) and by NODF as being highly anti-nested ($P(H_0) < 0.001$). (B) The herpetofauna of land bridge islands in the Sea of Cortez (Murphy 1983) is identified by temperature and discrepancy as being nested and by NODF as being antinested. (C) The Canary Island birds matrix (Bacallado 1976) appears to be segregated (C-score), nested (temperature) and as having turnover (correlation of occurrence ranks). (D) Bats along an elevation gradient (Atmar and Patterson 1995), are identified as being segregated (C-score), aggregated (nearest neighbor distance), nested (discrepancy), antinested (NODF, temperature), coherent (embedded absences), and as having turnover (correlation of occurrence ranks).

5. Matrix classification

Quantitative analysis of presence–absence matrices implies that we can position any matrix along a gradient encompassing different extremes of pattern, with random patterns occupying an intermediate position. For instance, metrics of nestedness define a gradient from nested to random to antinested (cf. Almeida-Neto et al. 2007 for the use of antinestedness). The C-score (Stone and Roberts 1990) and other metrics of species associations (cf. Leibold and Mikkelson 2002, Baselga 2010, Podani and Schmera 2011) define a gradient from species segregation to random patterns to species aggregation.

Recently Presley et al. (2010) proposed a more elaborated classification that extended the widely-cited approach of Leibold and Mikkelson (2002) to metacommunity structure. The key argument of both papers is that metacommunities can be classified into several distinct categories based on the particular patterns exhibited in presence–absence matrices. Presley et al. (2010) start with the distinction between species segregation (checkerboardiness) and coherence (species aggregation) as representing aspects of community structure. Then they subdivide coherence into 12 different types of species aggregation.

However, as noted early on by Stone and Roberts (1990), species segregation and species aggregation may be different sides of the same coin. Above a threshold matrix fill (which depends on the number of species in the matrix) any perfectly segregated (checkerboarded) presence–absences matrix can be rearranged to appear perfectly aggregated (cf. matrix **U** in Stone and Roberts 1990), simply by re-ordering the rows and the columns of the matrix. This re-ordering does not alter any of the underlying information on species occurrences in the matrix. In turn, any matrix that is strongly compartmentalized (that is having several clusters of nested species) can also be re-ordered to appear strongly segregated (checkerboarded). Any perfect checkerboarded matrix exhibits also strong species turnover (Stone and Roberts 1990), which is a change in composition across sites in a matrix that has been rearranged using ordination (reciprocal averaging) to achieve maximum turnover (Leibold and Mikkelson 2002). Indeed, turnover is always positively linked to species segregation, while it remains to be shown whether segregation necessarily implies turnover.

To assess the scatter in matrix patterns we screened the compilations of biogeographic matrices of Atmar and Patterson (1995) and Ulrich and Gotelli (2010) for matrices with seemingly contradictory patterns. Of the 435 matrices, seven were simultaneously significantly segregated (C-score metric of Stone and Roberts 1990; $P < 0.001$ as inferred from the null distributions of the fixed – fixed null model), aggregated (nearest neighbor distance test of Clark and Evans 1954 applied to a matrix ordinated by reciprocal averaging to maximize turnover), nested (discrepancy metric of Brualdi and Sanderson 1999), and exhibited turnover (measured by the coefficient of correlation of species occurrence ranks in the ordinate matrix; Ulrich and Gotelli unpubl.).

Figure 4B–D shows three typical examples of matrices with such a multiple structure. Figure 4B and 4C indicate that aggregation, segregation, turnover and coherence patterns arise in matrices with at least one compartment of higher fill and a high degree of spatial turnover. Figure 4D is a typical example of a matrix that is simultaneously nested and exhibits turnover. Our screening also showed that larger matrices seem to be especially prone to exhibiting multiple structures that are detected by null model tests. Each of the seven multiple structure matrices had more than 1000 cells, which exceeds significantly (U-test, $P < 0.001$) the average matrix size of the Atmar-Patterson and Ulrich-Gotelli data sets. These multiple structured matrices do not fit into any simple static classification scheme and may require a multivariate approach to pattern description (Podani and Schmera 2011).

Surprisingly, as many as 40 matrices were detected by discrepancy and NODF as being significantly nested and anti-nested, respectively. Only one matrix was jointly identified by the three popular nestedness metrics (NODF, discrepancy, and temperature) as being nested. Eighteen matrices were jointly detected by the nearest neighbor distance and the Morisita index (advocated by Leibold and Mikkelson 2002 and Presley et al. 2010) as being aggregated (Ulrich and Gotelli unpubl.). All of these 18 matrices are identified by the C-score as being segregated and (with one exception) by the correlation-of-occurrence-ranks metrics as having significant turnover. These analyses suggest that the interplay of species segregation, aggregation, and turnover in real metacommunities may be too complex and interwoven to generate discrete patterns that can be organized into a simple static classification scheme.

Our examples also show how difficult it may be to define patterns implicitly via metrics (see also Almeida-Neto et al. 2007, 2008). Incongruence in the assessment of structure between metrics is frequently met in ecological analyses and must be considered when proposing new patterns and classification schemes. Imprecise definitions of patterns and ad hoc introductions of metrics and algorithms that have not been subject to benchmark screening and analysis can introduce more confusion than clarity (Gotelli 2001). Relating complex patterns in presence–absence matrices to explicit ecological mechanisms remains a difficult challenge.

6. Going beyond presence–absence data

To date most biogeographic data sets only contain information on presences and absences that were obtained from floristic or faunistic surveys, and most null model tests use only the data contained within those matrices. However, additional information on both sites and species can be readily incorporated into null model analysis. For example, rather than assuming all sites of being equiprobable or fixing species richness per site, null models can place species on sites with probabilities determined by patch area (Connor et al. 2000, Jenkins 2006) or habitat suitability (Peres-Neto et al. 2001). Similarly, colonization potential of species can be estimated independent of occurrence by using estimates of body size, population size, or biomass (Gotelli et al. 2010). It is somewhat surprising how little extra biological information has so far been incorporated into null model analysis. Benchmark testing of null models that incorporate additional information can be challenging. However, a systematic comparison of model results with and without particular factors included can be very informative. Null models that incorporate independent information

occupy a worthwhile middle ground on the spectrum from standard randomizations tests (Manly 2009) to pattern-oriented modeling (Grimm et al. 2005) of detailed mechanistic processes (Gotelli et al. 2009).

Moreover, ecologists are starting to generate quantitative data matrices based on counts or estimates of abundance, biomass, or percent cover of each species. Abundance based matrices potentially contain more information on species associations than presence–absence matrices and might be better suited to infer patterns of matrix structure and the underlying processes that generate these patterns (Ulrich and Gotelli 2010, Almeida-Neto and Ulrich 2010). Abundance data might be particularly helpful to resolve questions connected to the long-lasting debate about the existence and importance of ecological assembly rules (Diamond 1975, Hubbell 2001).

Recently, Ulrich and Gotelli (2010) and Almeida-Neto and Ulrich (2010) developed and tested metrics and null model algorithms for the study of species associations and nestedness in abundance matrices. Both papers also highlighted three potential challenges in the use of these matrices. First, abundance matrices are based on counts of individuals, not species. Thus familiar metrics of nestedness, species segregation, or species turnover need precise redefinitions and new metrics that incorporate information from abundance. The subtle differences in pattern definition between presence–absence and abundance data should be taken into account when comparing patterns across matrix types. Second, abundance matrices allow for a much wider scope of potential randomization algorithms to obtain null distributions (reviewed by Ulrich and Gotelli 2010). Third, it is not yet clear how to incorporate underlying population-level processes, such as density-dependence, migration, and especially aggregation, which can potentially affect patterns in abundance matrices. More research and data collection is needed for the analysis of abundance matrices.

Acknowledgements – We thank P. Guimarães and J. A. F. Diniz-Filho for comments that improved early versions of this manuscript. WU was in part supported by grants from the Polish Science Committee (KBN 3 P04F 03422 and KBN 2 P04F 039 29). NJG acknowledges support from grants NSF DEB-0541936 and DOE DE-FG02-08ER64510.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. – In: Petrov, B. N. and Csaki, F. (eds), Proc. 2nd Int. Symp. on Information Theory. Akadémia Kiadó, Budapest, pp. 267–281.

Almeida-Neto, M. and Ulrich, W. 2010. A straightforward computational approach for measuring nestedness using quantitative matrices. – Environ. Modell. Softw. 26: 173–178.

Almeida-Neto, M. et al. 2007. On nestedness analyses: rethinking matrix temperature and anti-nestedness. – Oikos 116: 716–722.

Almeida-Neto, M. et al. 2008. A consistent metric for nestedness analysis in ecological systems: reconciling concept and quantification. – Oikos 117: 1227–1239.

Almeida-Neto, M. et al. 2011. Rethinking the relationship between nestedness and beta-diversity: a comment on Baselga (2010). – Global Ecol. Biogeogr. in press.

Anderson, D. R. 2008. Model based inference in the life sciences. – Springer.

Atmar, W. and Patterson, B. D. 1993. The measure of order and disorder in the distribution of species in fragmented habitat. – Oecologia 96: 373–382.

Atmar, W. and Patterson, B. D. 1995. The nestedness temperature calculator: a Visual Basic program, including 294 presence–absence matrices. – AICS Res., Univ. Park, NM and Field Museum, Chicago <http://aics-research.com/nestedness/tempcalc.html>.

Bacallado, J. J. 1976. Notas sobre la distribucion y evolucion de la avifauna Canaria. – In: Kunkel, G. (ed.), Biogeography and ecology in the Canary Islands. Junk, pp. 413–431.

Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. – Global Ecol. Biogeogr. 19: 134–143.

Brualdi, R. A. and Sanderson, J. G. 1999. Nested species subsets, gaps, and discrepancy. – Oecologia 119: 256–264.

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. – Springer.

Chao, A. et al. 2010. Phylogenetic diversity measures based on Hill numbers. – Phil. Trans. R. Soc. B 365: 3599–3609.

Clark, P. J. and Evans, F. C. 1954. Distance to nearest neighbour as a measure of spatial relationships in populations. – Ecology 35: 445–453.

Colwell, R. K. and Winkler, D. W. 1984. A null model for null models in biogeography. – In: Strong, D. et al. (eds), Ecological communities: conceptual issues and the evidence. Princeton Univ. Press, pp. 344–359.

Colwell, R. K. and Lees, D. C. 2000. The mid-domain effect: geometric constraints on the geography of species richness. – Trends Ecol. Evol. 15: 70–76.

Colwell, R. K. et al. 2004. The mid-domain effect and species richness patterns: what have we learned so far? – Am. Nat. 163: E1–E23.

Colwell, R. K. et al. 2005. The mid-domain effect: there's a baby in the bathwater. – Am. Nat. 166: E149–E154.

Connor, E. H. and Simberloff, D. 1979. The assembly of species communities: chance or competition? – Ecology 60: 1132–1140.

Connor, E. F. et al. 2000. Individuals–area relationships: the relationship between animal population density and area. – Ecology 81: 734–748.

Diamond, J. M. 1975. Assembly of species communities. – In: Cody, M. L. and Diamond, J. M. (eds), Ecology and evolution of communities. Harvard Univ. Press, pp. 342–444.

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

Efron, B. 2005. Bayesians, frequentists, and scientists. – J. Am. Stat. Ass. 100: 1–5.

Fayle, T. M. and Manica, A. 2010. Reducing over-reporting of deterministic co-occurrence patterns in biotic communities. – Ecol. Modell. 221: 2237–2242.

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

Gotelli, N. J. 2001. Research frontiers in null model analysis. – Global. Ecol. Biogeogr. 10: 337–343.

Gotelli, N. J. and Graves, G. R. 1996. Null models in ecology. – Smithsonian Inst. Press.

Gotelli, N. J. and McCabe, D. J. 2002. Species co-occurrence: a meta-analysis of J. M. Diamond's assembly rules model. – Ecology 83: 2091–2096.

Gotelli, N. J. and McGill, B. J. 2006. Null versus neutral models: what's the difference? – Ecography 29: 793–800.

Gotelli, N. J. and Ulrich, W. 2010. The empirical Bayes distribution as a tool to identify non-random species associations. – Oecologia 162: 463–477.

Gotelli, N. J. and Ulrich, W. 2011. Over-reporting bias in null model analysis: a response to Fayle and Manica (2010). – Ecol. Modell. 222: 1337–1339.
Gotelli, N. J. et al. 1997. Co-occurrence of Australian land birds: Diamond's assembly rules revisited. – Oikos 80: 311–324.
Gotelli, N. J. et al. 2009. Patterns and causes of species richness: a general simulation model for macroecology. – Ecol. Lett. 12: 873–886.
Gotelli, N. J. et al. 2010. Macroecological signals of species interactions in the Danish avifauna. – Proc. Natl Acad. Sci. USA 107: 530–535.
Grant, P. R. and Abbott, I. 1980. Interspecific competition, island biogeography and null hypotheses. – Evolution 34: 332–341.
Graves, S. 1978. On the Neyman–Pearson theory of testing. – Br. J. Phil. Soc. 29: 1–23.
Green, J. and Bohannan, J. M. 2006. Spatial scaling of microbial diversity. – Trends Ecol. Evol. 21: 501–507.
Grimm, V. et al. 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. – Science 310: 987–991.
Harvey, P. H. et al. 1983. Null models in ecology. – Annu. Rev. Ecol. Syst. 14: 189–211.
Hawkins, B. A. and Diniz-Filho, J. A. F. 2002. The mid-domain effect cannot explain the diversity gradient of Nearctic birds. – Global Ecol. Biogeogr. 11: 419–426.
Holt, R. D. 1997. From metapopulation dynamics to community structure: some consequences of spatial heterogeneity. – In: Hanski, I. and Gilpin, M. E. (eds), Metapopulation biology. Academic Press, pp. 149–164.
Hubbard, R. and Bayari, M. J. 2003. Confusio over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. – Am. Stat. 57: 171–182.
Hubbell, S. P. 2001. The unified theory of biogeography and biodiversity. – Princeton Univ. Press.
Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. – Ecol. Monogr. 54: 187–211.
Järvinen, O. 1982. Species-to-genus ratios in biogeography: a historical note. – J. Biogeogr. 9: 363–370.
Jenkins, D. G. 2006. In search of quorum effects in metacommunity structure: species co-occurrence analyses. – Ecology 87: 1523–1531.
Jetz, W. and Rahbek, C. 2001. Geometric constraints explain much of the species richness pattern in African birds. – Proc. Natl Acad. Sci. USA 98: 5661–5666.
Keil, P. and Hawkins, B. A. 2009. Grid versus regional species lists: are broad scale patterns of species richness robust to the violation of constant grain size. – Biodivers. Conserv. 18: 3127–3137.
Ladau, J. 2008. Validation of null model tests using Neyman–Pearson hypothesis testing theory. – Theor. Ecol. 1: 241–248.
Ladau, J. and Schwager, S. J. 2008. Robust hypothesis tests for independence in community assembly. – J. Math. Biol. 57: 537–555.
Leibold, M. A. and Mikkelson, G. M. 2002. Coherence, species turnover, and boundary clumping: elements of meta-community structure. – Oikos 97: 237–250.
Loehle, C. 2011. Complexity and the problem of ill-posed questions in ecology. – Ecol. Compl. 8: 60–67.
MacArthur, R. H. and Wilson, E. O. 1963. An equilibrium theory of insular zoogeography. – Evolution 17: 373–387.
Manly, B. J. 2009. Randomization, bootstrap and Monte Carlo methods in biology, 3rd ed. – Chapman and Hall/CRC.
Miklós, I. and Podani, J. 2004. Randomization of presence–absence matrices: comments and new algorithms. – Ecology 85: 86–92.
Murphy, R. W. 1983. The reptiles: origins and evolution. – In: Case, T. J. and Cody, M. L. (eds), Island biogeography in the Sea of Cortez. Univ. of California Press, pp. 130–158.
Navarro-Alberto, J. A. and Manly B. F. J. 2009. Null model analysis of presence–absence matrices need a definition of independence. – Popul. Ecol. 51: 505–512.
Patterson, B. D. and Atmar, W. 1986. Nested subsets and the structure of insular mammalian faunas and archipelagos. – Biol. J. Linn. Soc. 28: 65–82.
Peres-Neto, P. R. et al. 2001. Environmentally constrained null models: site suitability as occupancy criterion. – Oikos 93: 110–120.
Podani, J. and Schmera, D. 2011. A new conceptual and methodological framework for exploring and explaining pattern in presence–absence data. – Oikos in press.
Presley, S. J. et al. 2010. A comprehensive framework for the evaluation of metacommunity structure. – Oikos 119: 908–917.
Rahbek, C. et al. 2007. Predicting continental-scale patterns of bird species richness with spatially explicit models. – Proc. R. Soc. B 274: 165–174.
Rangel, T. F. L. V. B. and Diniz-Filho, J. A. F. 2005. Neutral community dynamics, the mid-domain effect and spatial patterns in species richness. – Ecol. Lett. 8: 783–790.
Roughgarden, J. 1983. Competition and theory in community ecology. – Am. Nat. 122: 583–601.
Sanderson, J. G. et al. 1998. Null matrices and the analysis of species co-occurrences. – Oecologia 116: 275–283.
Schluter, D. 1984. A variance test for detecting species associations, with some example applications. – Ecology 65: 998–1005.
Sfenthourakis, S. et al. 2006. Species co-occurrence: the case of congeneric species and a causal approach to patterns of species association. – Global Ecol. Biogeogr. 15: 39–49.
Sokal, R. R. and Rohlf, F. J. 1995. Biometry: the principles and practice of statistics in biological research, 3rd ed. – Freeman.
Stone, L. and Roberts, A. 1990. The checkerboard score and species distributions. – Oecologia 85: 74–79.
Strong, D. R. 1980. Null hypotheses in ecology. – Synthese 43: 271–285.
Tuomisto, H. 2010. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. – Ecography 33: 2–22.
Ulrich, W. 2004. Species co-occurrences and neutral models: reassessing J. M. Diamond's assembly rules. – Oikos 107: 603–609.
Ulrich, W. and Gotelli, N. J. 2007a. Null model analysis of species nestedness patterns. – Ecology 88: 1824–1831.
Ulrich, W. and Gotelli, N. J. 2007b. Disentangling community patterns of nestedness and species co-occurrence. – Oikos 116: 2053–2061.
Ulrich, W. and Gotelli, N. J. 2010. Null model analysis of species associations using abundance data. – Ecology 91: 3384–3397.
Ulrich, W. et al. 2009. A consumer's guide to nestedness analysis. – Oikos 118: 3–17.
Wootton, J. T. 2005. Field-parameterization and experimental test of the neutral theory of biodiversity. – Nature 433: 309–312.
Wright, D. H. et al. 1998. A comparative analysis of nested subset patterns of species composition. – Oecologia 113: 1–20.