(MR. CHAI MOLINA (Orcid ID : 0000-0001-9722-4446

Article type : Articles

Difficulties in benchmarking ecological null models: an assessment of current methods

Chai Molina[1,2,5] and Lewi Stone[3,4]

[1]Department of Ecology and Evolutionary Biology, Princeton University, Princeton NJ, USA

[2]International Institute for Applied Systems Analysis, Laxenburg, Austria

[3]Biomathematics Unit, Department of Zoology, Faculty of Life Sciences, Tel Aviv University

[4] Mathematics, School Science, RMIT University, Melbourne, Australia

[5] Corresponding Author. E-mail: chai.molina@gmail.com

Running Head: Benchmarking ecological null models

doi: 10.1002/ecy.2945

Abstract

Identifying species interactions and detecting when ecological communities are structured by them is an important problem in ecology and biogeography. Ecologists have developed specialized statistical hypothesis tests to detect patterns indicative of community-wide processes in their field data. In this respect, null model approaches have proved particularly popular. The freedom allowed in choosing the null model and statistic to construct a hypothesis test leads to a proliferation of possible hypothesis tests from which ecologists can choose to detect these processes. Here, we point out some serious shortcomings of a popular approach to choosing the best hypothesis for the ecological problem at hand that involves benchmarking different hypothesis tests by assessing their performance on artificially constructed datasets. Terminological errors concerning the use of Type-I and Type-II errors that underlie these approaches are discussed. We argue that the key benchmarking methods proposed in the literature are not a sound guide for selecting null hypothesis tests, and further, that there is no simple way to benchmark null hypothesis tests. Surprisingly, the basic problems identified here do not appear to have been addressed previously, and these methods are still being used to develop and test new null models and summary statistics, from quantifying community structure (e.g., nestedness and modularity) to analyzing ecological networks.

Keywords: Null models, Type I error, Type II error, Benchmarking, Power, Robustness, Community structure

Introduction

A long-standing debate in biogeography concerns the composition of ecological communities and the identification of species interactions that might structure them (Cody and Diamond 1975, Gotelli 1999, Weiher and Keddy 1999). As a result, ecologists have developed specialized statistical tools that test for the presence of patterns indicative of community-wide processes, such as interspecific competition, in their field data. Null model approaches have

proved particularly popular (Gotelli and Graves 1996) and over the last twenty years have been applied in thousands of published studies. Given the plethora of possible null models, Gotelli (2000) and Ulrich and Gotelli (2010) have devised a benchmarking procedure for choosing the most appropriate one for a given ecological application. Here, we argue that the benchmarking methods they propose are problematic and do not yield an appropriate yardstick for selecting a null hypothesis test.

We focus on ecological community data in the form of an abundance matrix. The entries of such a matrix, a_{ij}, represent the abundance of species-i at sample site-j as quantified by either counts of observations of individuals, or their densities. Each row in the matrix represents the abundances of a species at different sites. Columns represent the different focal communities or different sites. The entire abundance matrix represents the **metacommunity** of species at the sampled sites. Summing up the elements along row-i gives the abundance of species i across all sites, and variability in the row sums may indicate that some species colonize sites better than others. Similarly, summing the entries of the j^{th} column gives the total species abundance at site j, and variability in the column sums may indicate that some sites or focal communities are colonized more easily, or that they can support greater species richness. The distributions of the row and column sums are important defining features of a metacommunity (Connor and Simberloff 1979).

A number of null model algorithms have been developed to generate random simulated abundance matrices that are by design unstructured, in that the process by which abundances of each species at each site are allocated do not involve any species interaction or other community structure. Often a great deal of thought has been given to defining what is meant by random in this context, and devising tests to ensure that these metacommunities are truly random (Artzy-Randrup and Stone 2005; Stone and Roberts 1990,1992). Some null model algorithms generate matrices subject to realistic constraints (specific to the null hypothesis under consideration), and are thus able to simulate key features or constraints that may occur in real data, without incorporating any ecological mechanisms related to species interaction that we suspect might result in community structure. Using a null model, it is then possible to statistically test whether a given dataset is unstructured as regards species interactions, while taking into account non-random features that

might occur for other ecological reasons. Rejection of the null hypothesis suggests the presence of a non-random structuring process beyond those incorporated in the null model.

Null model algorithms typically begin with an input reference matrix, or "observed" matrix, from which the various constraints are calculated. The algorithm is then able to generate matrices that are random samples from the ensemble of all possible matrices satisfying these constraints. For example, a widely applied null model generates an ensemble of random abundance matrices whose row sums are all fixed to the values of a given (observed) abundance matrix. This ensures that the pattern of species abundances in the observed metacommunity (e.g., resulting from variation in colonization abilities) are preserved in all the simulated metacommunities generated by the null model. Thus, all random metacommunities are completely unstructured, at least over and above the deliberately imposed observed row sum constraints. Other null models impose the constraint that only column sums of the random metacommunities are kept fixed to observed values. Still other null models fix both row and column constraints or allow only a small variability in them.

Ulrich and Gotelli (2010), henceforth referred to as U&G, propose a total of 14 different null models for creating unstructured abundance matrices, which are essentially randomization algorithms. In addition to these 14 null models, U&G propose six statistics to measure community structure. From these, it is possible to create 14×6=84 hypothesis tests to detect structure in abundance data, by coupling a null model and a statistic. A key goal of U&G is to develop a benchmarking test that evaluates the performance of these different hypothesis tests, making it possible to choose the best option. Here we critically discuss their benchmarking procedure and find it unsuitable for ranking the different hypothesis tests. We argue that instead of U&G's scoring method, researchers should choose null models primarily based on biological considerations, while possibly also taking into account power and robustness to assumption violations relevant to their specific system (Heeren and D'Agostino 1987; Lehmann and Romano 2005; Ladau 2008; Ladau and Schwager 2008).

Since the benchmarking methods developed in U&G are mostly extensions of those developed by Gotelli (2000) for presence-absence data, and are suggested as a general procedure to benchmark null hypothesis tests for other aspects of community structure, e.g., nestedness and

modularity (Gotelli and Ulrich 2012, Ulrich and Gotelli 2007, 2013), much of our criticism applies also to these other studies, to which we collectively refer as **UGG**. Despite their shortcomings, the main ideas presented in UGG are being taken up not only in many areas of basic ecological research (e.g., Chaves and Anez 2004, Feeley 2003, Kembel and Hubbell 2006, Lavender et al. 2016, Lyons et al. 2016, McNickle et al. 2018, Mouillot et al. 2005, Peres-Neto et al. 2001), but also policy-oriented studies (Kobza et al 2004, Semmens et al. 2010, Schmidlin et al. 2012, Tulloch et al. 2018) and even studies of the microbiome (Li et al. 2018). Moreover, these same criticisms apply also to new benchmarking methods for testing structure in ecological networks (Vaughan et al. 2018). Surprisingly, the basic problems discussed here do not appear to have been addressed previously, yet their relevance could be of crucial importance for all these related studies.

Review of basic concepts

As an aid for the ensuing discussion, it is helpful to briefly review fundamental concepts regarding hypothesis tests (for more details, see Sokal and Rohlf 1995). A hypothesis test is a statistical tool to choose between competing hypotheses, the null (and usually simpler) hypothesis H_0 and an alternative H_a (often left unspecified). The process involves evaluating a statistic, T, that is, a function of the observed data, and determining whether its value is exceptional, under the assumption that the null hypothesis H_0 is true. One then asks, under the null hypothesis H_0, what is the probability of observing a sample for which the value of the statistic is at least as extreme as that observed? This probability is the p-value of the data. The experimenter then rejects the null hypothesis if the p-value is smaller than a preselected cut-off value, α referred to as the significance level of the test, and fails to reject it otherwise. Thus, a hypothesis test consists of a null hypothesis H_0, a test statistic T, and a significance level α. Importantly, the null hypothesis must imply a specific distribution for the statistic T which may be calculated analytically or evaluated numerically.

A Type I error, or false positive, occurs when the null hypothesis H_0 is rejected despite being true. The rate at which H_0 is rejected when the test is applied to null data (i.e., data generated by the null hypothesis) is by definition $100 \times \alpha$ percent, and referred to as the Type I

error rate (or false positive rate). The significance level is typically set at $\alpha = 0.05$ or 0.1, corresponding to Type I error rates of 5% and 10%, respectively.

Similarly, Type II errors, or false negatives, occur when H_0 is not rejected, even though the alternative H_a is true. The probability of a false negative is denoted by β, and the probability of correctly rejecting H_0, $1 - \beta$, is termed the power of the test, and reflects its ability to detect an alternative hypothesis. It is difficult to put much faith in a test having low power, since when such a test fails to reject H_0, it might well be because the test was simply not very sensitive.

Given a null hypothesis, we usually first choose the significance level α and then (if the alternative hypothesis allows) we measure β for various test-statistics. This allows a comparison between various hypothesis tests (composed of a null model, a test statistic and a fixed α) differing only in their test-statistic. By design, null hypothesis tests constructed in this manner have the same Type I error rate (determined by α), regardless of which statistic is used. We then choose the statistic that yields the lowest Type II error rate, i.e., the highest power

Benchmarking null hypothesis tests

Henceforth, we focus on the main question motivating UGG: How should one choose the best null hypothesis test for detecting species interactions that might affect the structure of a metacommunity (e.g., causing species aggregation or segregation)? Two fundamental questions arise:

1. All else being equal, what would this metacommunity look like without species interspecific interactions?
2. What feature is the most appropriate for identifying processes that create community structure?

Answering the first question corresponds to choosing a null model representing the absence of species interactions. In situations in which species differences (e.g., demographic parameters or trophic levels) are unimportant, and species can be considered functionally equivalent, this can be addressed using neutral models of community assembly (Bell 2005, Gotelli & McGill 2006). The second question corresponds to the choice of a test statistic for detecting

such interspecific interactions. When a null model has been chosen, it is simple to compare tests differing only in their test statistics (when power can be estimated). UGG's main contribution is a proposed method to answer both questions at once, i.e., to compare tests differing in both their null hypotheses and statistics. We refer to such a comparison procedure as benchmarking.

The novelty of benchmarking is in deciding on the best null model for helping identify species interactions. Selecting such a null model would naïvely require testing several null models against data about what the metacommunity of interest would look like without species interactions. But, such data are usually absent (because if we knew that species interactions haven't shaped our data, we would not be seeking a hypothesis test for identifying species interactions). Given this difficulty in null model selection, a successful benchmarking procedure would be extremely useful.

U&G present fourteen null model algorithms and 6 statistics, giving rise to 84 different null hypothesis tests. But how should a researcher choose which is the most appropriate test? The approach recommended by U&G is as follows.

U&G use another (fifteenth) algorithm to create a set of reference "unstructured test matrices", a collection of manufactured abundance matrices representing synthetic metacommunities with no interactions between species. The test matrices have row sums with a predefined log-normal distribution (because the log-normal distribution of abundance data is one of ecology's best-documented scaling laws; Preston, 1962a&b, McGill et al. 2006). However, they are constructed carefully to ensure there is no hidden mechanism that creates species segregation or aggregation. (More precisely, U&G use two pools of unstructured test matrices, which they denote M_R and M_S, respectively, that are constructed using slightly different algorithms; however, this does not affect our argument below. Gotelli (2000) uses four different algorithms to generate test matrices.) U&G also construct a set of structured test matrices by manipulating the unstructured test matrices to make species artificially aggregated or segregated.

U&G then apply the 84 null hypothesis tests to the sets of unstructured and structured test matrices, and suggest a two-step benchmarking procedure to assess their performance. First, they select the four null hypothesis tests that reject the null hypothesis for the fewest number of unstructured test matrices. They then score this subset of tests using the proportion of the

structured test matrices for which the null hypothesis was rejected. We henceforth focus on the first step of UGG's benchmarking procedure, that is, of the 84 null hypothesis tests, selecting the one that rejects the least number of unstructured matrices is considered to be the best test.

Problems with UGG's Benchmarking Method

In this section, we highlight problems with U&G's benchmarking methodology. Before highlighting the main conceptual problem and its repercussions, we address a terminological issue that obscures the overall underlying logic.

Problem 1: confusion between type I errors and power (terminological)

UGG define the "Type I error rate" of a null model as the proportion of unstructured test matrices that it rejects. This definition is nonstandard and imprecise, because the test matrices used to calculate this rejection rate are created using an algorithm different from the null model randomization algorithms U&G analyze. These unstructured test matrices are therefore not "null" with respect to any of the 84 null hypothesis tests U&G propose, despite being constructed so as to reflect no species interactions and having log-normal abundance distributions. Indeed, U&G find that for each and every null model they examine, there is a statistic with which the rejection rate for the unstructured test matrices is much greater than 5% (see table 1 in U&G). This indicates that each of the null models tested differs from the algorithm used to generate the unstructured test matrices in some way that relates to species co-occurrence (because the statistics used were selected in order to detect patterns in co-occurrence). This could not occur if the unstructured test matrices were truly null (with respect to any one of the null models used), because by definition, when $\alpha = 0.05$, exactly 5% of all null matrices are rejected.

Because the unstructured test matrices are generated by an algorithm different from the null model, the quantity that UGG measure is the frequency at which the null hypothesis *correctly* rejects the unstructured matrices. Thus, UGG measure the *power* of the various hypothesis tests ($1 - \beta$), yet refer to it as the type I error rate. (In making the power calculation, the implied alternative hypothesis is that the data were generated from the algorithm that created the unstructured test matrices.) Referring to the rejection of the unstructured test matrices as "type I errors" is a confusing and nonstandard use of a common technical term.

We emphasize that U&G are well-aware of their nonstandard use of terminology. They write that "[a]lthough the formal definition of a Type I error is incorrect rejection of a true null hypothesis, we use [a different] operational definition here of rejection of H_0 on a set of appropriate [unstructured] test matrices created by random sampling from a log-normal distribution" (U&G, p.3386). Nonetheless, the distinction between their "operational definition" and the standard one is downplayed or obscured throughout, leading to the impression that true Type I and II error rates are really being measured.

Treating the unstructured test matrices as if they were null can be thought of as simply keeping the idea of what "null" means imprecise, i.e., that "null" means conforming to some idea of being "unstructured". This is problematic because communities can be unstructured in many different ways, as is evidenced by the many possible null models U&G benchmark, all of which create abundance matrices that could in principle be considered unstructured. Using the term "null" in this way introduces a vagueness that is similar to that entailed by calling something "random": random numbers can be generated using many different distributions, but samples from a normal and a uniform distribution look very different from one another. The particulars of the random process can matter quite a bit, and similarly, so can the particulars of what it means to be "unstructured".

Problem 2: Using a particular null model to benchmark others (conceptual)

The following important problem we raise is conceptual. Contrary to standard procedure, U&G do not assume a particular null model and compare different statistics. Instead, they compare entire hypothesis tests, that is, statistics and null models taken together as a unit. But in selecting a hypothesis test, UGG also select a null model. It is therefore natural to ask, is the null hypothesis selected by U&G's benchmarking procedure our best guess for a model of the true processes shaping this metacommunity, excluding the (possible) effects of species interactions? We argue that the answer is no. Because this issue has gone unnoticed for almost two decades, we present two different arguments to support this claim.

Argument 1: The forces shaping different metacommunities may well differ, so we expect that testing for species interactions in different metacommunities will often require using different null models. To decide which of, say, two such "guesses" better describes a metacommunity, one

must study the metacommunity in question, not which of these two null models better-describes a dataset generated by a third model of metacommunities without species interactions. For example, the question of whether or not to use a null model in which column sums are constrained translates to a question about the ecological system being studied: is there an ecological reason why some of the sites being studied support more species than others (e.g., because they are more readily colonized)? Note also that while empirically studying what a metacommunity might look like in the absence of interspecific interactions (which is a counterfactual proposition) is often unfeasible, this does not detract from our logical argument.

In essence, UGG's benchmarking process is analogous to comparing apples and oranges by using plums as a reference. This can lead us to choose the model that most resembles that used to generate the test matrices (see our second argument below), and to choose a statistic with low sensitivity to the differences between the null models being benchmarked and the one that is used to do the benchmarking (see *Problem 4*).

Argument 2: U&G's benchmarking procedure is biased to choose the matrix randomization algorithm that "most resembles" the algorithm used to generate the unstructured test matrices. Indeed, U&G write: "these analyses are […] optimized for their performance on the set of matrices that we created by random sampling from a lognormal distribution of species abundances." U&G consider the unstructured test matrix generation algorithm to be a good model of real unstructured ecological metacommunities; as such, the null hypothesis tests are scored based on their performance on these unstructured test matrices.

However, U&G's test-matrix generation algorithm is composed of two distinct parts: (a) a procedure for generating log-normal and uniform species and site abundance distributions (respectively); and, (b) a "randomization algorithm" used to allocate fractions of these log-normally distributed abundances to particular entries in the matrix. And although the log-normal distribution of species abundances is well-supported in many ecological applications, U&G's choice of algorithm for distributing these species abundances across sites (i.e., columns) is not empirically supported.

Note that the species abundances of U&G's test matrices being log-normally distributed does not imply that the test-matrix randomization algorithm is better-justified than other null

models (i.e., randomization algorithms) U&G propose. To see this, observe that different randomization algorithms that preserve row sums can be applied to a particular empirically-derived abundance matrix with log-normally distributed species abundances. Matrices generated in this way will still have a log-normal abundance distribution, and the process used to generate them is not necessarily more or less "natural" than the one U&G use to generate the unstructured test matrices.

The following argument illustrates the circularity of benchmarking null models based on their performance on the unstructured test matrices generated from a procedure that is not empirically supported. Because the test-matrix randomization algorithm is not "better" or more natural than the null models (that is, randomization algorithms) being benchmarked, it could be considered a 15th possible null model, which we refer to as R_{TEST}. It is then possible to construct hypothesis tests based on this new randomization algorithm (e.g., by combining it with each of the 6 statistics of U&G, similar to the 14 null model algorithms U&G benchmarked).

We are then faced with two possibilities:

1. If we believe the test-matrix randomization algorithm R_{TEST} is preferable to the 14 null models benchmarked by U&G, then there is no reason to consider the other null models; we need only choose between null hypothesis tests constructed by pairing R_{TEST} with one of the six statistics proposed by U&G. In this case, we can then fix the significance level α and choose the statistic that yields the best power.

2. If we are uncertain whether R_{TEST} is better than the 14 null models suggested by U&G, then R_{TEST} should also be benchmarked and compared against the other 14 null models. But if we do so, the unstructured test matrices will be rejected at a rate of exactly α for any statistic chosen (because R_{TEST} was used to create them). Thus, U&G's benchmarking procedure would be biased in favour of selecting R_{TEST}.

In other words, there is no need to run the test on fourteen randomization algorithms if we already know which is best. If we do not a-priori know the best randomization algorithm to use, then we want a selection criterion that is not biased in favour of one or another randomization algorithm, a quality that UGG's benchmarking methodology lacks. (Replacing R_{TEST} with another

randomization algorithm to generate unstructured test matrices will similarly result in a benchmarking procedure that favours this new randomization algorithm.)

Lastly, we note that this problem is essentially the same criticism that U&G have towards other studies. They write (page 3385 therein) "A [different] approach is to specify a mechanistic colonization model that does not include species interactions, such as the neutral model (Bell 2005), and then use that model to create random matrices that can be used to evaluate null model procedures (Ulrich 2004). The disadvantage of this method is that the test is narrowly optimized for one particular mechanistic model, and there is no logical reason that this model should have priority." We agree, but also extend this criticism to U&G's model for generating test matrices.

Problem 3: A null hypothesis test cannot be "prone to Type I errors" (terminological)

UGG state that some tests are more prone to Type I errors and therefore not very effective. However, for a given significance level α, a test cannot be more or less prone to Type I errors, because the Type I error rate is by definition exactly $100 \times \alpha$ percent (but see the section "robustness: overview" below). The number of rejected matrices in a set of N matrices generated using the null model is binomially distributed with parameters N and α. As such, the true Type I error rate as a score for different tests is uninformative because any difference between the theoretical and observed rejection rates results from the stochastic nature of generating matrices using the null model, not the choice of statistic or null model.

Problem 4: Benchmarking encourages low power (conceptual)

Since the power of the test to correctly reject the null hypothesis when confronted with the unstructured test matrices is mislabelled as the Type I error (see *Problem 1*), UGG would like it to be around the α-level they have set, and are thus ensuring a very low power by setting $1 - \beta = 0.05$ (0.1 in Gotelli 2000). A power of 5% means that in practice, the hypothesis test will not be able to distinguish between its null model and the alternative model generating the unstructured test matrices; for 95% of the unstructured test matrices the test will incorrectly fail to reject the null hypothesis. Experimenters invariably strive to design a test with *high* power, so that if the null hypothesis H_0 is not true, they will most likely know about it. This cannot happen when the power is set at $1 - \beta = 0.05$. Moreover, as we show in a sequence of examples in Appendix

S1, seeking a statistic that is "blind" to the differences between a null model and a particular alternative model (qualitatively similar to the null model) can lead to choosing a statistic that has low power to distinguish other alternative models that are qualitatively very different from the null.

Beyond benchmarking

Having established that the benchmarking methods proposed by UGG are not a sound guide to selecting a hypothesis test, how *should* one choose from a set of hypothesis tests? Unfortunately, there is no simple answer. No single null hypothesis test is appropriate in all ecological contexts; rather, the null hypothesis and statistic should be selected on a case-by-case basis, based on the specific characteristics of the system being studied. Here, we outline some suggestions for factors that should be taken into account in this process. Most of these have been identified long ago (see for example Weiher and Keddy 1999), but bear repeating given the widespread use of UGG's benchmarking procedure for justifying null model selection, rather than ecological considerations.

As mentioned, different ecological contexts will require different null models. In particular, whether or not to incorporate row and column constraints has been a source of debate for decades (e.g., Stone and Roberts 1990; Weiher and Keddy 1999; Gotelli 2000). For example, row sums should in many cases be constrained because they reflect differences in vagility or colonizing ability. But if species are closely related, competition might influence the row sums themselves, and a species might be more common than another simply because it arrived or evolved first. As noted by Fox (1999), this issue is linked to the Narcissus effect, whereby "[s]ampling from a post-competition pool underestimates the role of competition, since its effect is already reflected in the pool" (Colwell and Winkler 1984).

Similar care should also be given to choosing a statistic that is sensitive to the ways in which we expect species interactions to manifest in the ecological metacommunity being studied. One important difficulty is that inferences based on statistics that are designed to measure the same structural property are sometimes contradictory, even when using hypothesis tests with identical null models (Stone and Roberts, 1992, Gotelli and Ulrich 2012, Strona and Fattorini 2014). This suggests that intuitions about both how to measure different metacommunity

properties (e.g., segregation, aggregation, nestedness, turnover and modularity) and the relationships between them can be misleading. Using process-based models of metacommunity formation which allow the properties in question (in contrast to the matrix randomization algorithms proposed by UGG) can help evaluate whether and how different statistics reflect these structural properties.

When the null model has been chosen, and an alternative model of metacommunity structure in the presence of species interactions is available, measuring a test's power to discern between the null and alternative hypotheses is informative. If several statistics seem plausible for detecting species interactions, the power of the resulting tests is a natural measure for ranking them. It also is particularly desirable that a test be **unbiased,** i.e., that it correctly rejects the null hypothesis more often than it does incorrectly (Lehmann and Romano 2005).

Any null model we select, however biologically plausible, will be a caricature of reality. Consequently, it is also important to see whether or not a test might still be **robust** to certain deviations from its hypotheses. Robustness testing, reviewed below, is a powerful framework, but is unfortunately underused in ecology. We also briefly discuss approaches outside the null hypothesis testing paradigm in Appendix S2.

Robustness: overview

Testing for robustness is a modern statistical procedure (Lehmann and Romano 2005, Ladau 2008, Ladau and Schwager 2008) that requires considering scenarios whereby the null model's underlying assumptions might not be satisfied, but the null hypothesis being tested still is. For example, a common procedure for testing that the mean of a dataset is 0, is to use a Z-test. However, Z-tests rely on the additional assumption that the data are normally distributed. The normality assumption, however, is independent of the null hypothesis we wish to test — that the data have mean 0 — and the result of the Z-test does not indicate whether or not this additional assumption is satisfied. In general, the experimental datasets satisfy such additional assumptions (e.g., normality) only approximately, or not at all. In such cases, the probability of rejection of the null hypothesis could be larger or smaller than the significance level of the test, even though the dataset satisfies the null hypothesis being tested (e.g., mean 0). Only in this context is it

meaningful to say that the test is prone to Type I errors (Huber 1996, Heeren and D'Agostino 1987).

Null hypothesis tests are said to be robust if the observed (nominal) Type I error rates are maintained close to the pre-selected significance level α when some assumptions of the null model are violated (e.g., Heeren and D'Agostino 1987, Sullivan and D'Agostino 1992). In this context of testing robustness, even though the full assumptions of the null model are not met, we continue to describe the rejection of data *satisfying the null hypothesis* as Type I errors. For example, the two-sample t-test (for equality of means; see Sokal and Rohlf 1995) is based on the assumption that the observations are derived from normal distributions of equal variance and there are sufficiently many samples. However, it has been shown that this test is robust to non-normality, small sample size, and in some situations, unequal variances of the sample distributions (see Sullivan and D'Agostino 1992). Thus, verifying the robustness of a hypothesis test to violations of some of its assumptions helps us know that our inferences from the test may still be valid, even when we cannot guarantee that some of the test's assumptions hold for our experimental data.

Note that UGG's benchmarking process is not a test of robustness. A test of robustness aims to measure the effects of gradual and controlled changes in the individual assumptions of a model (Sullivan and D'Agostino 1992; Lehmann and Romano 2005). Instead, UGG check whether the rejection rates of the tests are substantially changed when the entire null model is altered, comparing all the null models under consideration to an essentially unrelated model that is based on vastly different assumptions. This is why, when confronting the hypothesis tests with the unstructured test matrices, UGG find the rejection rates for some of the tests to be high (sometimes even 100%). Moreover, to use robustness for comparing different models, one must check the robustness of these algorithm to identical assumption violations. Thus, comparing the proportions of test matrices rejected is unfair because the assumption violations that transform each of the tested models into the test matrix generation algorithm are different.

Suggestions for robustness-testing in ecology

When exploring the effects of violations of the basic assumptions of hypothesis tests, it is imperative to clearly state what these assumptions are, which of these are being violated, and how. No less importantly, the biological reasons for the interest in the violation of assumptions should

be stated. Because many scenarios of assumption violations are not biologically relevant, or not relevant to the experiment in question, there is no need to make general statements as to the robustness or bias of tests for *any* assumption violation (which is no simple task). While Ladau's framework for systematically studying the effects of various violations of assumptions on a myriad of tests is very useful, his disappointment in not finding an all-purpose universally robust (and unbiased) test is unwarranted. The results of his study may be used by researchers to choose which test is relevant for their particular study design. We emphasize that this requires researchers to be intimately familiar with both the biological systems studied and the statistical methods involved. Unfortunately, many ecologists will likely find Ladau (2008) inaccessible due to the heavy technical jargon and terse description of the mathematical methods.

As an example, suppose that after careful consideration of her study system, a researcher has opted to use a model with fixed zeros (i.e., species absences), fixed row sums, and column probabilities proportional to sums (a minor variation on the model "ITR" in U&G), and an appropriate statistic (selected, for instance, based on a power analysis). One might expect the conditions of fixed row sums and fixed zeroes to fluctuate somewhat in biological data, even though, in principle, these are the constraints relevant to this study system. At the very least, random errors in measurements of these constraints would arise due to sampling. Thus, it would be useful for the researcher explore how robust the test is to gradually increasing fluctuations in the row sums, or in the locations of zeros in the abundance matrix. She could then estimate these fluctuations empirically from the experimental data (sampling errors could also be evaluated) and assess the relevance and reliability of the inferences made using this hypothesis test.

UGG and Ladau test a plethora of null models against an alternative model that is often structurally very different from the null models being benchmarked. Instead, we suggest future research into more standard tests of robustness. This would involve evaluating the robustness of specific null hypothesis tests to "gentler" changes in their underlying assumptions (ideally the magnitude of the deviation from the null hypothesis can be turned up or down using an appropriate parameter). This is a promising area of research that UGG and Ladau (2008) seem to be moving towards.

Two final caveats are that the robustness of a test is not an excuse for sloppy modelling or experimental setups, and that robustness should not be the sole criterion by which we choose our tests. Robust tests using unrealistic models or uninformative statistics are not superior to non-robust tests with realistic, biologically relevant models and statistics. Additionally, when choosing between models for explaining a particular phenomenon, the ability to explain other patterns may and should also be used as a gauge of a model's viability.

Conclusion

UGG address the question of how to detect structure in an ecological metacommunity by suggesting a method for benchmarking null hypothesis tests by comparing the results of null hypothesis tests (differing in their null models and statistics) when confronted with a dataset constructed using a null model different from all those tested. Though the goal of UGG is worthwhile, their suggested solution suffers from various statistical and methodological problems, the most important of which is that comparing tests with different null models on a set of artificially-constructed test matrices does not inform us which null model better describes a particular real ecological metacommunity. The reasons this problem has heretofore gone unnoticed are likely confusion and vagueness relating to the concepts of randomness, null models and Type I errors.

Choosing a hypothesis test is a real concern for ecologists in particular and scientists in general, and there is much confusion about how to do so throughout the literature. A simple prescription does not exist, but neither do we expect one to appear, due to the broad terms in which the problem is posed. U&G and especially Gotelli (2000) are correct to stress that there is no "all-purpose" test to use, and that sound judgment should be used when constructing a hypothesis test: "Ecologists need to move beyond the idea that there is a single "one-size-fits-all" null model that is appropriate. Rather, the null model and index should be chosen based on the kind of data […] collected and the question being asked" (Gotelli 2000).

In the previous section, we outlined general guidelines for selecting and evaluating hypothesis tests to detect community structure. The selection of a null hypothesis test must not be

based on trying to choose a null model (i.e., randomization algorithm) and statistic such that the statistic's distribution on data generated by the null model is similar to the same statistic's distribution on some other reference model. Instead, it should be grounded in intimate knowledge of the study system and thoughtful scrutiny of the biological, ecological and statistical considerations involved. If a sensible model for the alternative hypothesis is available, the power of a test to detect the alternative can inform the selection of a test statistic (statistics resulting in higher power are better). Tests that are biased (i.e., reject the null hypothesis incorrectly more often than they do correctly) should be avoided. Since "all models are wrong" (Box 1976), it is also worthwhile to evaluate a hypothesis test's robustness (i.e., its performance under biologically plausible violations of some of its basic assumptions). Note, however, that few guidelines are available for testing robustness of ecological null models (see Ladau 2008; Ladau and Schwager 2008). This challenging area requires future research to mature, and to make these tests more accessible to ecologists.

Lastly, the confusion we highlight surrounding null model selection in general, and type I errors in particular, suggests to us a real and pressing need to better train ecology graduates in statistics. In general, the current state-of-the-art statistical software is extremely powerful and readily accessible, but without a deep understanding of the theory involved, it is all too easily misused.

Acknowledgments: We acknowledge the support of the Australian Research Council (grant DDP150102472) and the Israel Science Foundation (LS and CM). We are also grateful to N. Gotelli, C. Hennig, S. Meiri and W. Ulrich, and for discussions and exchange of ideas.

Bibliography

Artzy-Randrup, Y. and Stone, L., 2005. Generating uniformly distributed random networks. Physical Review E, 72(5), p.056708.

Bell, G., 2005. The co-distribution of species in relation to the neutral theory of community ecology. Ecology, 86(7), pp.1757-1770.

Box G. E. 1976. Science and statistics. Journal of the American Statistical Association. 71(356):791–799.

Chaves, L. F. and N. Anez. 2004. Species co-occurrence and feeding behavior in sand fly transmission of American cutaneous leishmaniasis in western Venezuela. Acta Trop. 92: 219–224.

Cody, M. L., and Diamond, J. M. (eds.) 1975. Ecology and Evolution of Communities. Cambridge, MA: Belknap Press, Harvard University Press. pp. 342–444.

Colwell, R. K. and D. W. Winkler. 1984:344– 359. A null model for null models in biogeography. In D. R. Strong, D. Simberloff, L. G. Abele, and A. B. Thistle (eds.). Ecological communities: conceptual issues and the evidence Princeton University Press, Princeton, New Jersey, USA.

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

Feeley, K. J. 2003. Analysis of the avian communities of Lake Guri, Venezuela, using multiple assembly rule models. Oecologia 137: 104-113

Fox, B. J. 1999:23-57. The genesis and development of guild assembly rules. In: Weiher E., and Keddy (eds.). Ecological assembly rules. Cambridge University Press.

Gotelli, N. 1999. "ECOLOGY: How Do Communities Come Together?". Science 286 (5445): 1684–1685. doi:10.1126/science.286.5445.1684a.

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

Gotelli, N. J., and Graves, G. R. 1996. Null models in ecology. Smithsonian Institution Press, Washington, DC.

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. 2012. Statistical challenges in null model analysis. Oikos, 121(2), 171-180.

Hennig C. 2009 (personal communication).

Heeren, T. and D'Agostino, R. 1987. Robustness of the two independent samples t-test when applied to ordinal scaled data. Statistics in medicine 6.1: 79-90.

Huber, P. J. 1996. Robust Statistical Procedures. SIAM.

Kembel, S. W., and Hubbell, S. P. 2006. The phylogenetic structure of a neotropical forest tree community. Ecology. 87(Suppl): S86-S99. doi:10.1890/0012-9658(2006)87[86:TPSOAN]2.0.CO;2

Kobza, R. M., Trexler, J. C., Loftus, W. F., and Perry, S. A. 2004. Community structure of fishes inhabiting aquatic refuges in a threatened Karst wetland and its implications for ecosystem management. Biological Conservation, 116(2), 153-165.

Ladau, J. 2008. Validation of null model tests using Neyman-Pearson hypothesis testing theory. Theoretical Ecology, 1: 241-248.

Ladau, J. and Schwager S. J. 2008. Robust Hypothesis Tests for Independence in Community Assembly. Journal of Mathematical Biology, 57: 537-555.

Lavender, T. M., Schamp, B. S., and Lamb, E. G. 2016. The influence of matrix size on statistical properties of co-occurrence and limiting similarity null models. PLoS one, 11(3), e0151146.

Lehmann, E. L., and Romano, J. P. 2005. Testing statistical hypotheses. Springer New York (3rd edition).

Li, H., Li, T., Li, X., Wang, G., Lin, Q., and Qu, J. 2018. Gut microbiota in Tibetan herdsmen reflects the degree of urbanization. Frontiers in microbiology, 9.

Lyons, S. K., Amatangelo, K. L., Behrensmeyer, A. K., Bercovici, A., Blois, J. L., Davis, M., ... and Gotelli, N. J. 2016. Holocene shifts in the assembly of plant and animal communities implicate human impacts. Nature, 529(7584), 80.

McGill, B. J., Maurer, B. A., and Weiser, M. D. 2006. Empirical evaluation of neutral theory. Ecology 87.6: 1411-1423.

McNickle, G. G., Lamb, E. G., Lavender, M., Cahill Jr, J. F., Schamp, B. S., Siciliano, S. D., Condit, R., Hubbell, S. P. and Baltzer, J. L. 2018. Checkerboard score–area relationships reveal spatial scales of plant community structure. Oikos, 127(3), 415-426.

Mouillot, D., George-Nascimento, M. and Poulin, R. 2005. Richness, structure and functioning in metazoan parasite communities. Oikos, 109: 447–460. doi: 10.1111/j.0030-1299.2005.13590.x

Peres-Neto, P. R., Olden, J. D. and Jackson D. A. 2001. Environmentally constrained null models: site suitability as occupancy criterion. Oikos, 93: 110–120. doi: 10.1034/j.1600-0706.2001.930112.x

Preston, F. W. 1962a. The canonical distribution of commonness and rarity: Part I. Ecology 43.2:185-215.

Preston, F. W. 1962b. The canonical distribution of commonness and rarity: part II and II. Ecology 43.3: 410-432.

Roberts, A. and Stone, L. 1990. Island-sharing by archipelago species. Oecologia 83: 560-567.

Schmidlin, S., Schmera, D., & Baur, B. 2012. Alien molluscs affect the composition and diversity of native macroinvertebrates in a sandy flat of Lake Neuchâtel, Switzerland. Hydrobiologia, 679(1), 233-249.

Semmens, B. X., Auster P. J., and Paddack, M. J. 2010. Using Ecological Null Models to Assess the Potential for Marine Protected Area Networks to Protect Biodiversity. PLoS ONE 5(1): e8895. doi:10.1371/journal.pone.0008895

Simberloff D., Stone L., Dayan T. 1999:58-74. Ruling out a community assembly rule: the method of favored states. In: Weiher E., and Keddy (eds.). Ecological assembly rules. Cambridge University Press.

Sokal, R. R., and Rohlf, F. J. 1995. Biometry: the principles and practice of statistics in biological research. 3rd edition. W. H. Freeman and Co.: New York.

Stone, L., and Roberts, A. 1990. The checkerboard score and species distributions. Oecologia 85: 74–79.

Stone, L., and Roberts, A. 1992. Competitive exclusion, or species aggregation? Oecologia 91: 419-424

Strona, G., & Fattorini, S. 2014. On the methods to assess significance in nestedness analyses. Theory in Biosciences, 133(3-4), 179-186.

Sullivan, L. M., and D'Agostino, R. 1992. Robustness of the t-test applied to data distorted from normality by floor effects. Journal of Dental Research Vol. 71(12), 1938-1943

Tulloch, A. I., Chadès, I., and Lindenmayer, D. B. 2018. Species co-occurrence analysis predicts management outcomes for multiple threats. Nature ecology & evolution, 2(3), 465.

Ulrich, W., 2004. Species co-occurrences and neutral models: reassessing JM Diamond's assembly rules. Oikos, 107(3), pp.603-609.

Ulrich, W., and Gotelli, N. J. 2007. Null model analysis of species nestedness patterns. Ecology, 88(7), 1824-1831.

Ulrich, W., and Gotelli, N. J. 2010. Null model analysis of species associations using abundance data. Ecology, 91(11), 3384-3397.

Ulrich, W., and Gotelli, N. J. 2013. Pattern detection in null model analysis. Oikos, 122(1), 2-18.

Vaughan, I. P., Gotelli, N. J., Memmott, J., Pearson, C. E., Woodward, G. and Symondson, W. O. C. econullnetr: An R package using null models to analyse the structure of ecological networks and identify resource selection. Methods in Ecology and Evolution. 2018; 9: 728–733. DOI: 10.1111/2041-210X.12907

Weiher E., and Keddy, P. 1999. Ecological assembly rules. Cambridge University Press