vol. 184, no. 5 the american naturalist november 2014

Note

Zeta Diversity as a Concept and Metric That Unifies Incidence-Based Biodiversity Patterns

Cang Hui[1] and Melodie A. McGeoch[2,*]

1. Centre for Invasion Biology, Department of Mathematical Sciences, Stellenbosch University, Matieland 7602, South Africa; and African Institute for Mathematical Sciences, Muizenberg 7945, South Africa; 2. School of Biological Sciences, Monash University, Melbourne 3800, Australia

Submitted February 26, 2014; Accepted June 6, 2014; Electronically published September 24, 2014

Online enhancements: appendix, zip file.

abstract: Patterns in species incidence and compositional turnover are central to understanding what drives biodiversity. Here we propose zeta (ζ) diversity, the number of species shared by multiple assemblages, as a concept and metric that unifies incidence-based diversity measures, patterns, and relationships. Unlike other measures of species compositional turnover, zeta diversity partitioning quantifies the complete set of diversity components for multiple assemblages, comprehensively representing the spatial structure of multispecies distributions. To illustrate the application and ecological value of zeta diversity, we show how it scales with sample number, grain, and distance. Zeta diversity reconciles several different biodiversity patterns, including the species accumulation curve, the species-area relationship, multispecies occupancy patterns, and scaling of species endemism. Exponential and power-law forms of zeta diversity are associated with stochastic versus niche assembly processes. Zeta diversity may provide new insights on biodiversity patterns, the processes driving them, and their response to environmental change.

Keywords: macroecology, beta diversity, occupancy, distance decay, scaling, turnover.

Introduction

Spatial variation in the presence or absence of species in assemblages, or compositional diversity, underpins the study of biodiversity. One of the main hurdles to understanding relationships between various theories of biodiversity is differences in mathematical language and the lack of unified sets of equations (McGill 2010). There is currently no single measure that connects the range of assemblage patterns constructed from species presence-absence, or "incidence," data. This prevents the mathematical relationships between them from being formulated. By providing a common currency, a single measure would have significant advantages for modeling and understanding the mechanistic basis of spatial patterns in diversity. The integration of biodiversity models in this way is a central goal of ecology (Scheiner and Willig 2008).

How and why biodiversity changes between sites and habitats, and the consequences of this variation, are often examined through species richness and composition per se. Measures of spatial variation in the compositional similarity of assemblages are commonly based on β diversity. These are derived from partitioning regional γ diversity into α and β components and use either Whittaker's (1960) multiplicative ($\beta = \gamma/\alpha$) or Lande's (1996) additive ($\gamma = \alpha + \beta$) diversity partitioning. A range of assemblage patterns, such as species-area relationships, interspecific range size distributions, and patterns of rarity and endemism, are also used (Gaston and Blackburn 2000; McGill 2010).

All existing measures of compositional similarity and difference were originally derived for pairwise comparisons of individual assemblages (sites, samples, or areas), regardless of the partitioning approach used (Jost et al. 2011; McGlinn and Hurlbert 2012). When comparisons of three or more assemblages are involved, the average of the pairwise similarities is used. As a result, none of the metrics of presence-absence (incidence)-based species turnover across sites is able to calculate all diversity components. In other words, the diversity components of three or more assemblages cannot all be expressed with only α and β. For example, in a three-assemblage case, the species shared exclusively by pairs of assemblages within the comparison cannot be calculated from only α and β, and neither can the species shared by all three assemblages. As a result, pairwise metrics are not sufficient for representing assemblage similarity across multiple sites (Chao et al. 2008).

The few existing multiple-assemblage, incidence-based measures have shortcomings (Koch 1957; Diserud and Ødegaard 2007). These include (1) inference problems as a consequence of averaging nonindependent pairwise val-

[*] Corresponding author; e-mail: melodie.mcgeoch@monash.edu.

Am. Nat. 2014. Vol. 184, pp. 684–694. 0003-0147/2014/18405-55316$15.00.
DOI: 10.1086/678125

ues and (2) the fact that when the number of sites considered is large, the values become less reliable and more difficult to interpret (Jost 2007). Another approach used for accommodating multiple samples in comparisons is incremental pooling of nested samples or areas by means of a hierarchical sample design (Crist et al. 2003). This approach has contributed significantly to understanding how β diversity changes with spatial scale, but it does not also allow for complete partitioning of diversity components across multiple assemblages. Ideally, the diversity metric should show how species incidence and turnover vary continuously with the addition of independent or nested sites across space.

Here we propose zeta (ζ) diversity as a concept and metric that captures all diversity components produced by assemblage partitioning. As a result, it reconciles existing descriptors of species incidence and compositional turnover. We illustrate the scale dependence of ζ diversity by showing how it changes with sampling grain and extent, and we relate this to hierarchical β diversity partitioning and to the distance decay of similarity. Using 291 real species-by-site matrices (Atmar and Patterson 1995), we identify the most common forms of ζ diversity (power law and exponential), the ecological implications of these, and their relationship with incidence-based assemblage patterns. We also show how ζ diversity can be used to produce general formulas for a range of biodiversity patterns. These include sample-based species accumulation curves, the endemics-effort relationship, and the diverse forms of species occupancy frequency distributions. We conclude by recommending ζ diversity as a concept and metric that unifies incidence-based biodiversity patterns. The use of ζ diversity may provide new insights about the drivers of species composition and turnover, co-occurrence, community assembly processes, and the consequences of environmental change for biodiversity.

Zeta Diversity Partitioning

Let the ζ component, ζ_i, be the mean number of species shared by i sites (fig. 1). Note that ζ_1 (where $i = 1$) is simply the mean number of species across all sites. Since species shared by i sites will necessarily be among those shared by $i - 1$ sites, the number of shared species ζ_i declines monotonically with i. All incidence-based, pairwise β diversity metrics can be expressed with ζ_1 and ζ_2. However, with three or more sites, the diversity components (e.g., A–G in the inset of fig. 1, where a component [or partition] is the subset of species shared by a particular set of sites; Lande 1996) cannot all be estimated with α and pairwise β components only (i.e., ζ_1 and ζ_2). The higher-order ζ components are needed to do so.

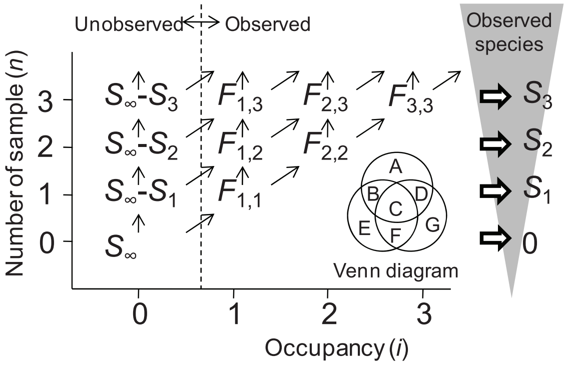

Figure 1: Recursive diagram of diversity partitioning across multiple assemblages (sites or samples). Initially, there are S_∞ unknown species in the assemblage (lower left). With an increase in number of samples, the number of unknown species gradually declines to $S_\infty - S_n$. With each site added, $F_{1,n}/n$ new species are added to the species inventory, and the cumulative number of known species increases from S_{n-1} to S_n. The addition of a site also increases by 1 the occupancy of some discovered species, while the occupancy of other discovered species remains unchanged (those that do not occur in the newly added site). The Venn diagram inset shows the diversity partitioning of three assemblages with species partitioned into seven disjoint sets (components A–G). Let AB represent the joint set of species in components A and B and $|A|$ the number of species in component A. Using the ζ component definition, $\zeta_1 = (|\mathrm{ABCD}| + |\mathrm{EFBC}| + |\mathrm{GFCD}|)/3$, $\zeta_2 = (|\mathrm{BC}| + |\mathrm{CD}| + |\mathrm{FC}|)/3$, and $\zeta_3 = (|C|)$; the number of species newly discovered in one, two, and three samples are $S_1 = \zeta_1$, $S_2 = (|\mathrm{ABCDEF}| + |\mathrm{EFCBDG}| + |\mathrm{GFCDAB}|)/3 = 2\zeta_1 - \zeta_2$, and $S_3 = |\mathrm{ABCDEFG}| = 3\zeta_1 - 3\zeta_2 + \zeta_3$, respectively; also, $F_{1,3} = |\mathrm{AEG}| = 3\zeta_1 - 6\zeta_2 + 3\zeta_3$, $F_{2,3} = |\mathrm{BDF}| = 3\zeta_2 - 3\zeta_3$, and $F_{3,3} = \zeta_3$. As ζ_3 represents a unique diversity component, it cannot be expressed as a function of ζ_1 and ζ_2 (the β-diversity component) only.

This procedure for diversity partitioning can be illustrated in a cumulative and recursive way (fig. 1). Let S_n be the total number of species across n sites and $F_{i,n}$ the number of species that occupy i sites out of the total n sites surveyed. Using the inclusion-exclusion principle (see "Incidence-Based Zeta Diversity Partitioning" in the appendix, available online), the following general formulas can be derived deductively using ζ components:

$$S_n = \sum_{k=1}^{n} (-1)^{k+1} \cdot C_n^k \cdot \zeta_k, \tag{1}$$

$$F_{i,n} = C_n^i \cdot \sum_{k=1}^{n-i+1} (-1)^{k+1} \cdot C_{n-i}^{k-1} \cdot \zeta_{i+k-1}, \tag{2}$$

where C_n^i $(= n!/[i!(n - i)!])$ is the number of combinations of choosing i from n sites and k is the standard index of summation.

Adding one extra site to a survey with $n - 1$ sites will add $F_{1,n}/n$ new species (fig. 1). A species occupying i sites is either present in the new site (i.e., now occupies $i + 1$ sites) or absent from it (i.e., its occupancy remains i sites).

This means that the number of species in the region (S_∞), with the number of surveyed sites increasing to infinity, can be estimated as follows (see "Incidence-Based Zeta Diversity Partitioning" in the appendix):

$$S_\infty = S_n + \sum_{k=n+1}^{\infty} \frac{F_{1,k}}{k}. \quad (3)$$

When sampling within a habitat or region, the number of unique species per sample $F_{1,n}/n$ declines monotonically with n. If the series $<F_{1,n}/n>$ is mathematically convergent to 0 (meaning that the limit of $\sum (F_{1,n}/n)$ is finite, as it would be for a clearly delimited sampling extent), there will be an asymptote to S_∞; otherwise, there will be no asymptote. ζ diversity thus represents the first approach for analyzing continuous changes in multispecies occupancy (presence-absence) and turnover across discrete, independent sites.

Zeta Diversity, Sample Number, Grain, and Distance

In the previous section, we expressed ζ as a function of the number of sites or samples (i). However, ζ can also be expressed as a function of other survey design parameters, such as distance, area, or grain, and for either aggregate (including hierarchical, nested) or independent sampling schemes (sensu Scheiner et al. 2011). ζ diversity relationships can therefore be used for estimating sampling completeness, for understanding how diversity is affected by the spatial properties of the samplings scheme (grain, distance, and extent), and for analyzing multiscale patterns of species diversity (Veech and Crist 2010; McGlinn and Hurlbert 2012). Here we examine the form of the relationship between ζ diversity and these ecologically relevant parameters.

Zeta Diversity Decline with Sample Number

The number of species shared by samples declines monotonically with sample number (i). However, the exact form of the relationship between ζ diversity and i (hereafter, "ζ diversity decline") is variable. We examined the fit of seven parametric models to ζ diversity decline with increasing sample number, using 291 empirical species-by-site matrices (Atmar and Patterson 1995) and the adjusted R^2 (see "Form of the Relationship between Zeta Diversity and Sample Number (Zeta Decline)" in the appendix). The power law provided the best fit in 57% (167) of cases and the exponential in a further 26% (76) of cases (fig. A1; figs. A1–A3 available online), with the selected model fitting extremely well (adjusted $R^2 > 0.95$ for all matrices; adjusted $R^2 > 0.99$ for more than 80% of the matrices; supplementary table, available online). Together, these two forms accounted for more than 80% of observed relationships. As a result, here we discuss only the implications of these two specific forms of ζ diversity decline for incidence-based biodiversity patterns. The exponential and power-law forms of ζ diversity decline are underpinned by distinct hypotheses about ecological process; that is, they represent species turnover as either largely stochastic (exponential ζ) or driven principally by niche differentiation processes (power-law ζ; Munoz et al. 2008; Scheiner et al. 2011).

First, the probability that a species shared by $i - 1$ sites is also found to be shared by i sites can be expressed as the ζ component ratio, ζ_i/ζ_{i-1}. If this probability is independent of i (e.g., $\zeta_2/\zeta_1 = \zeta_{101}/\zeta_{100}$), then the form of ζ diversity decline is exponential ($= a \cdot e^{-b \cdot i}$; fig. 2*A*). This means that with every new site, the chance of an already discovered species being found again in the new site does not depend on the species' current occupancy. Species with high or low regional occupancy will have, counterintuitively, an equal chance of being found in the new site. A null model with all the species having the same probability of occurring in a site, regardless of the heterogeneity across sites, will produce this exponential form of ζ diversity decline. In this null model, the predicted number of occupied sites is the same for all species, and variation in realized occupancy and turnover arises from stochastic species assembly. This can happen, for example, when species with relatively similar or large range sizes overlap spatially to form a local assemblage or where strong environmental flows (wind or water) result in stochastic establishment of propagules and occurrence patterns (fig. 2*A*).

Alternatively, if the ζ component ratio (ζ_i/ζ_{i-1}) is dependent on i, in most cases increasing with i (e.g., $\zeta_2/\zeta_1 < \zeta_{101}/\zeta_{100}$), this means that the chance of finding a common species in a new site is larger than finding a rare one. This case is consistent with the scale-heritage assumption, which holds when each species in the community has an occupancy status that is partially inherited across spatial scales (Hui and McGeoch 2008). In other words, the status of a species as either common or rare, based on current sampling effort, is a useful predictor of its likely occupancy status with the addition of new sites. The simplest two-coefficient model of ζ diversity that exhibits this scale-heritage property is the power law, $\zeta_i = c \cdot i^{-d}$ (fig. 2*B*). Null models for the species-by-site matrix with species differing in their probability of occupying a site (e.g., species have different site or habitat preferences) commonly produce this power-law form of the ζ diversity decline with sample number. Communities with nonrandom co-occurrence patterns, such as those with clear niche or range differentiation, and competitively structured com-

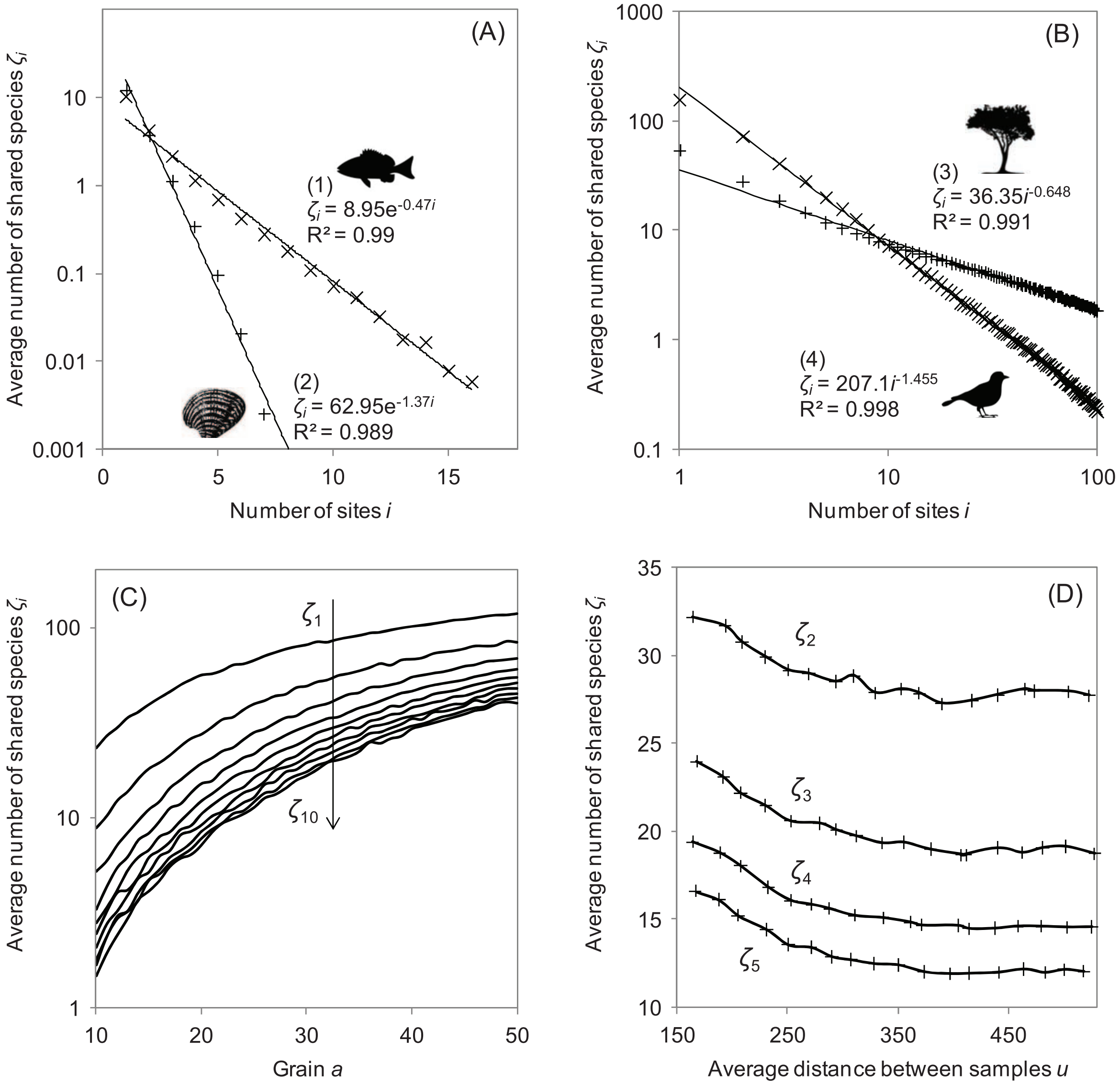

Figure 2: *A*, *B*, The two dominant forms of ζ decline (the relationship between ζ and sample number i): the exponential (*A*) and the power law (*B*). *A*, Freshwater fish in the Greenbrier River, West Virginia, with 35 species occurring in 30 sites (1; Hocutt et al. 1978); 41 marine fouling organisms on 12 tile plates (2; Sutherland and Karlson 1977). *B*, 20 × 20-m quadrate samples of 307 tree species in the 50-ha plot on Barro Colorado Island (3; Condit 1998; Hubbell et al. 1999, 2005); quarter-degree cells of 761 bird species in southern Africa (4; Harrison et al. 1997). *C*, The ζ scaling relationship (ζ-diversity as a function of sampling grain a [in m]). *D*, The distance decay of ζ diversity (average distance between samples u [in m] due to expanding sampling extent) for data set 3 in *B*. Lines from top to bottom in *C*, indicated by the arrow, are for ζ_1–ζ_{10}; those in *D* are for ζ_2–ζ_5 at a grain of 20 m × 20 m.

munities would be expected to have a power-law form of ζ diversity decline (fig. 2*B*).

Zeta Diversity Scaling with Sample Grain

As with all biodiversity metrics, ζ diversity is sensitive to the scale at which a study is conducted, that is, the grain and extent (Scheiner et al. 2011). When sampling grain increases, the species richness in each sample and the number of species shared by multiple samples will increase (fig. 2*C*). This incremental pooling of samples to form larger sampling grains, so that samples at the finest grain are nested within samples forming larger grain sizes, is termed a "hierarchical sample design." Crist et al. (2003) and Crist and Veech (2006) used such a design in their framework for β diversity partitioning. The general form of ζ diversity when pooling m samples to form n larger grain clusters in this way (see "Zeta Diversity Scaling with Sample Grain and Hierarchical Diversity Partitioning" in the appendix) is

$$\zeta_n(m) = \sum_{k=n}^{n\times m} \frac{\sum_{x_j \geq 1, \Sigma_{x_j=k}} \prod_{j=1}^{n} C_m^{x_j}}{C_{n\times m}^k} F_{k, n\times m}. \tag{4}$$

This provides the general form of the relationship between ζ diversity and sample grain, or the "ζ diversity scaling relationship" (fig. 2*C*). The number of species shared by two clusters, that is, the hierarchical partitioning of β diversity, where the first cluster is formed by pooling m_1 samples and the second by pooling another m_2 samples (see "Zeta Diversity Scaling with Sample Grain and Hierarchical Diversity Partitioning" in the appendix) is

$$\zeta_2(m_1, m_2) = \sum_{k=2}^{m_1+m_2} \frac{\sum_{i=1}^{k-1} C_{m_1}^i C_{m_2}^{k-i}}{C_{m_1+m_2}^k} F_{k, m_1+m_2}. \tag{5}$$

If we define $\zeta_1(m_1) = S_{m_1}$ and $\zeta_1(m_2) = S_{m_2}$, then Crist and Veech's (2006) hierarchical β diversity partitioning can be expressed using ζ diversity. Both β and ζ diversity components are nonindependent across hierarchical levels, or nested sample grains, because they represent part-to-whole associations (sensu Scheiner et al. 2011). Nonetheless, ζ diversity scaling can also be used for nonhierarchical sampling schemes, to better understand how diversity changes across spatial scales and, for example, the contribution of different habitats to regional diversity (Veech and Crist 2010).

Zeta Diversity Decay with Distance

An increase in the distance between samples or the average distance between random samples (see "Zeta Diversity Decay with Distance" in the appendix) results in a decline in the similarity of species composition (fig. 2*D*). This is known as the "distance decay of similarity" in applied ecology and biogeography (Nekola and White 1999; related to the n-point correlation function in physics [for $n = 2$]; e.g., Weinberg 1996). Distance decay relationships are valuable for estimating the rate of species turnover with distance and the importance of dispersal in driving the similarity of species assemblages at various scales (Qian and Ricklefs 2012). Traditionally, β diversity, for example, using the Jaccard index, is plotted against distance. ζ diversity can be used in a similar way and is likely to provide a more accurate estimate of the rate of species turnover with distance.

To formulate how ζ diversity decays with the increase in the mean distance between random samples, we used pair approximation from statistical physics. This is used to convert the spatial structures of species distributions to correlations between adjacent samples. As in Hui et al. (2006), the number of species shared by two random samples an average distance of u apart is

$$\zeta_2(u) = \zeta_1 Q(u), \tag{6}$$

where $Q(u)$ is an iterative function with $Q(0) = 1$ and $Q(1) = \zeta_2/\zeta_1$ (see "Zeta Diversity Decay with Distance" in the appendix for the full formulation). A direct formulation of $\zeta_n(u)$ for $n \geq 3$ is rather formidable, and indeed higher-order n-point correlation functions tend to be used only in complicated fields of physics (e.g., in quantum field theory; Weinberg 1996). Instead, we formulate higher-order $\zeta_n(u)$ (see "Zeta Diversity Decay with Distance" in the appendix; fig. 2*D*), using the Bayesian rule for inferring the presence/absence of a species in additional samples, given its occurrence in known samples.

In summary, the scale dependence of species turnover is well known, but the specific forms of these scaling relationships are not well established and remain central to understanding diversity dynamics and its context dependence (Soininen et al. 2007; McGlinn and Hurlbert 2012). Here we provide the general form of ζ diversity relationships with scale. The ζ diversity component $\zeta_i(u)$ declines with both the number of sites i (ζ diversity decline) and the average distance between random sites u (distance decay of similarity) and increases with grain (ζ diversity scaling; figs. 2, A2).

Zeta Diversity and Incidence-Based Biodiversity Patterns

ζ diversity can be used to derive several familiar and commonly used biodiversity descriptors and macroecological relationships. These include (1) the species accumulation curve (SAC), used to estimate species richness and the number of samples needed to achieve reliable richness estimates and richness comparisons (Colwell et al. 2004), (2) the endemics-effort relationship (EER), used to quantify the level of endemism and the sensitivity of local extinctions to habitat loss (Green and Ostling 2003; Storch et al. 2012), and (3) the occupancy frequency distribution (OFD), used to examine interspecific patterns in species range sizes (McGeoch and Gaston 2002; Hui and McGeoch 2007*a*). To illustrate the value of ζ diversity, below we show how ζ diversity informs debates about these relationships and the patterns in biodiversity that they represent.

Species Accumulation Curves

The SAC based on ζ components (eq. [1]) not only provides a general formula for forecasting species discovery with increasing effort but also illustrates how species turn-

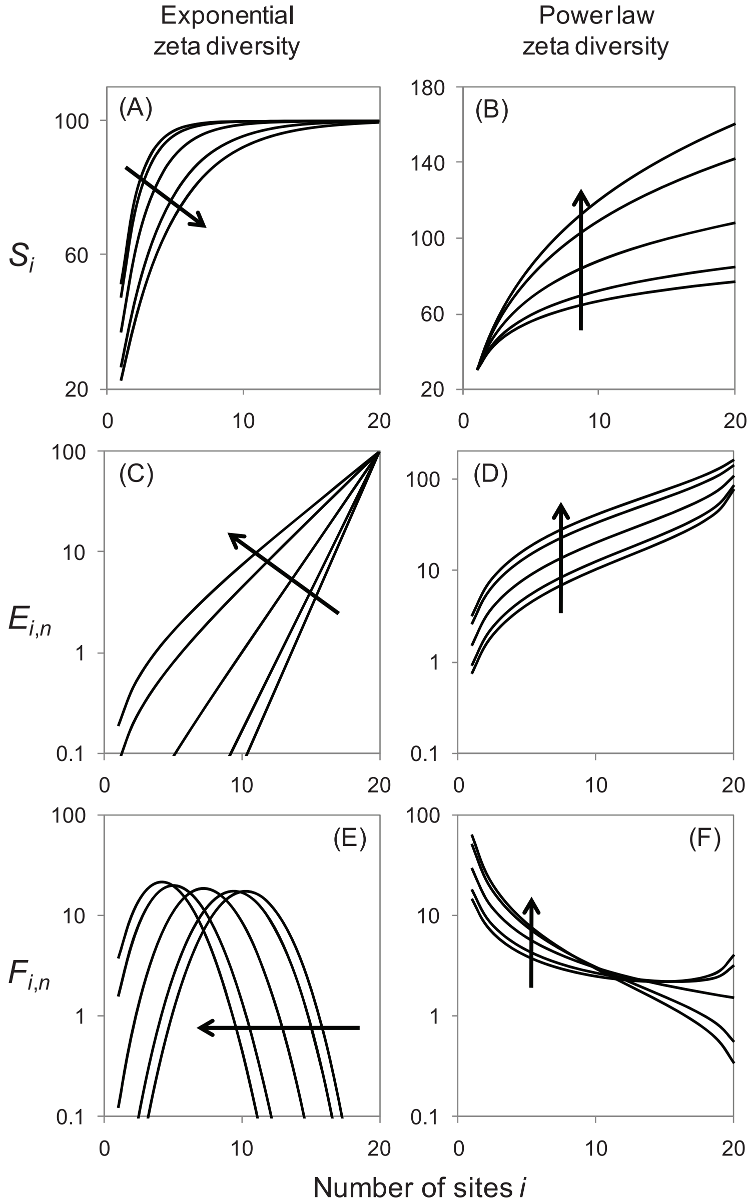

Figure 3: Species accumulation curves S_i (*A*, *B*), endemics-effort relationships $E_{i,\,n}$ (*C*, *D*), and occupancy frequency distributions $F_{i,\,n}$ (*E*, *F*) derived for the negative exponential form (*A*, *C*, *E*) and the power-law form (*B*, *D*, *F*) of ζ diversity based on equations (1), (8), and (2), respectively, with $a = 100$, $c = 30$, $n = 20$, and the values of *b* or *d* being 2/3, 3/4, 1, 4/3, and 3/2 for curves moving along directions of the arrow in each plot.

over affects the exact form of the SAC. ζ diversity provides a general estimator of richness for any particular number of samples (or areas; see below) without the need to assume the existence of an asymptote. The parameter S_n provides the SAC for *n* samples (or sites; eq. [1]). With an exponential form of ζ diversity decline, the series $<F_{1,\,n}/n>$ is convergent and the SAC has an asymptote (fig. 3*A*). Under this specific form of ζ diversity decline, the regional species richness can be specified with equation (3), which gives

$$S_\infty = S_n + \frac{n-1}{n}\frac{F_{1,n}^2}{2F_{2,n}} = S_{\text{Chao2}}, \qquad (7)$$

where S_{Chao2} is the Chao2 estimator that provides the regional species richness for *n* samples (sites; Chao 1984).

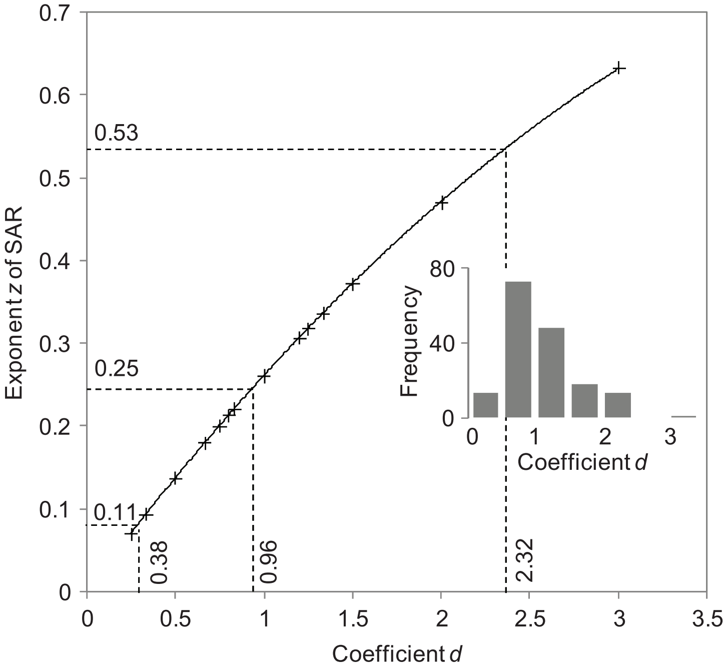

Figure 4: Relationship between the coefficient d of the power-law form of ζ diversity and the exponent z for the Arrhenius species-area relationship (SAR). Crosses represent estimates of z based on equation (1) for d ranging from 1/4 to 3. The z-d relationship can be complicated but resembles a polynomial function, $z = -0.024d^2 + 0.286d$, for large values of n (see "Zeta Diversity Decline and the Species-Area Relationship" in the appendix, available online). The inset shows the histogram of 167 estimates of d for the power-law form of ζ diversity, selected from 291 empirical species-by-assemblage matrices (supplementary table, available online) with the adjusted R^2. Dashed lines indicate the geometric mean of the inset histogram (0.96) and the corresponding z exponents (0.25) and the lower and upper 95% confidence intervals (details of estimates and related statistics are available in the supplementary table).

The general estimator of regional species richness (eq. [3]) therefore includes the Chao2 estimator as a special case. By contrast, the conditions of the power-law form of ζ diversity imply that the series $<F_{1,n}/n>$ is not convergent, and therefore there is no asymptote to the regional SAC (i.e., S_∞ is unbounded; fig. 3*B*). The difference between ζ diversity and nonparametric richness estimators such as Chao2 is that with ζ diversity, estimates are based on changes in the composition of both common and rare species across samples and not only on the frequencies of rare species.

When the SAC is derived from a hierarchical sample design, it is equivalent to the species-area relationship (SAR; Chase and Knight 2013). Efforts to describe the shape of empirical SARs (Drakare et al. 2006; Dengler 2009) have to date mostly involved curve fitting (but see He and Legendre 1996, 2002). A general formula for sample-based SARs that describes the range of forms of the relationship has been lacking (Gotelli and Colwell 2011), except under specific conditions (Colwell et al. 2004). With exponential ζ diversity decline, the SAR follows Fisher et al.'s (1943) "limiting form": $S_n = a[1 - (1 - e^{-b})^n]$ (fig. 3*A*). When ζ diversity decline follows a power law ($\zeta_i = c \cdot i^{-d}$; see "Zeta Diversity Decline and the Species-Area Relationship" in the appendix), the SAR resembles Arrhenius's (1920) power-law form, $S_n \sim n^z$. Based on the 167 empirical species-by-site matrices fitting the power-law form of ζ diversity (supplementary table), the coefficient d ranges from 0.38 to 2.32, with a geometric mean at 0.96. From this we estimate the SAR exponent z to range between 0.11 and 0.53, with a mean of 0.25 (fig. 4), consistent with empirical results (Drakare et al. 2006). The SAR generated from ζ diversity therefore encompasses both Fisher et al.'s (1943) negative exponential form and Arrhenius's (1920) power-law form, with the power-law exponent predicted to center around 0.25.

ζ diversity can also be used to examine the empirical relationship between the SAR and species turnover (as Sizling et al. 2011 and Grilli et al. 2012 did for β diversity). The relationship between the exponent z and the coefficient d of ζ diversity follows the general form $z = \ln(S_n/S_{n+1})/\ln(n/(n+1))$, from equation (1). It is complicated to specify the exact form for larger n (fig. A3), but the general form resembles a polynomial function (fig. 4). Tjørve and Tjørve's (2008) formula of pairwise species turnover as a function of the exponent (z), $\zeta_2/\zeta_1 = 2 -$

2^z, becomes a special case of the general form using ζ for $n = 1$ (see "Zeta Diversity Decline and the Species-Area Relationship" in the appendix). This particular expression is dependent on the number of samples and applies only when species distributions are scale invariant (McGlinn and Hurlbert 2012).

The Endemics-Effort Relationship

Patterns in species endemism, essential to effective conservation planning (Green and Ostling 2003; Sandel et al. 2011), can also be examined with ζ diversity. The endemics-area relationship was originally formulated as the number of species confined to smaller patches within a larger biome (Kinzig and Harte 2000). Here we express the number of endemics, using ζ diversity as a function of the number of sites sampled rather than of area per se. As we have shown above, area, grain, or distance can also be used. Let $E_{i,n}$ be the number of locally endemic species. These are species that occur only in the selected i sites within the total n sites surveyed. When n approaches infinity, $E_{i,n}$ will converge toward the number of globally endemic species, E_i (He and Legendre 2002; Green and Ostling 2003). Because of logistic constraints on sampling effort (which increase as n approaches infinity), the endemics-effort relationship (EER) is usually referred to as the number of locally endemic species (i.e., $E_{i,n}$), which is a function of the number of selected i sites. For selected i sites within a total of n sites surveyed, the number of local endemics (i.e., the local EER) can be expressed as

$$E_{i,n} = \sum_{k=1}^{i} C_i^k \cdot \frac{F_{k,n}}{C_n^k} = S_n - S_{n-i} = \sum_{k=n-i+1}^{n} \frac{F_{1,k}}{k} \qquad (8)$$

(see also "Incidence-Based Zeta Diversity Partitioning" in the appendix). When n is much greater than i, the number of local endemics approximates i times the derivative of S_n, $E_{i,n} \approx i \cdot \dot{S}_n$. This means that the local EER is approximately linear for low sampling effort. The roughly linear form of the EER has strong empirical support at large spatial scales (Storch et al. 2012). The multiple, specific forms possible for the EER across scales based on ζ diversity are specified in equation (8).

When ζ diversity decline is exponential (fig. 2*A*), the local EER follows

$$E_{i,n} = a(1 - e^{-b})^{n-i} \cdot [1 - (1 - e^{-b})^{i}]. \qquad (9)$$

The number of local endemics increases monotonically with i, exponentially for small values of b and following a power law for large values of b (fig. 3*C*). When ζ diversity decline follows a power law, the number of local endemics largely also follows a power law, with an exponent close to, but slightly greater than, 1 (fig. 3*D*). This is consistent with empirical results for endemics-area relationships (Storch et al. 2012). ζ diversity can therefore be used to compare empirical EERs within and across regions to better understand the consequences of environmental change for endemic species diversity.

Occupancy Frequency Distributions

Although the frequency of singleton and doubleton species in assemblages is useful for regional diversity estimation (Colwell and Coddington 1994), this represents only the rare species in an assemblage. A more general pattern that captures the frequency of species across the full range of occurrences in an assemblage (common, intermediate, or rare), is the occupancy frequency distribution (OFD; Gaston 1996; McGeoch and Gaston 2002). The OFD is the frequency distribution of the numbers of species occupying different numbers of sites and is used to quantify assemblage range patterns (McGeoch and Gaston 2002). The shape of OFDs has been used to formulate hypotheses about the mechanisms driving assemblage structure (Gaston and Blackburn 2000; Jenkins 2011).

Empirical OFDs have a diverse range of forms (McGeoch and Gaston 2002; Gaston and He 2011). The prevalence of bimodality in OFDs, that is, with modes for rare and common species and comparatively few species with intermediate occupancies, has been of particular interest (Raunkiær 1934; Hanski and Gyllenberg 1993). Several mechanisms have been proposed to explain this bimodality. One explanation is the core-satellite hypothesis, which explains bimodality as a division of the assemblage into groups of species with different stochastic immigration and extinction rates (Hanski 1982; but see Gotelli and Simberloff 1987; Gaston and Lawton 1989; Magurran and Henderson 2003). Explanations based on sampling (Nee et al. 1991; Papp and Izsák 1997) and the scale dependence of species occupancy (Conlisk et al. 2007; He and Condit 2007; Hui and McGeoch 2007*a*, 2007*b*) are also able to account for this bimodality. However, these explanations all make the implicit and unrealistic assumption of species independence. In other words, the occurrence of species A does not affect the occurrence of species B, and an OFD can be constructed by simply overlaying each species distribution independently of all others. Species turnover is thus assumed to be stochastic. ζ diversity does not presume species independence or stochastically driven species turnover, and it provides a mechanistic link between the shape of the OFD and the rates of turnover under which bimodality is possible.

The OFD of n samples, $F_{i,n}$, is provided by equation (2). To produce a bimodal OFD, the inequalities $F_{1,n} > F_{2,n}$ and $F_{n-1,n} < F_{n,n}$ must, at a minimum, be satisfied, from which we have $n < 2$ if ζ_i follows the exponential form. This suggests that bimodal OFDs are not possible under the conditions of the exponential form of ζ diversity decline. Instead, the OFD is unimodal, with the mode shifting to the left with an increase in the exponent b (fig. 3*E*). On the other hand, with the power-law form of ζ diversity decline, the OFD becomes bimodal if the inequality $d < \ln((n+1)/n)/\ln(n/(n-1))$ is satisfied. For instance, to ensure a bimodal OFD, values of $d < 0.98$ for $n = 50$ and $d < 0.99$ for $n = 100$ are needed. For large numbers of sites (>100), $d < 1$ ensures the presence of bimodality in the OFD (fig. 3*F*).

Low species turnover among the more common species in the assemblage produces a shallow slope for ζ diversity decline and a bimodal OFD. When environmental change has a disproportionally detrimental effect on rare species, the slope of ζ diversity decline will become shallower (shifting toward a bimodal OFD). When common species are more severely affected than rare ones, a steeper ζ diversity decline is expected, along with a right-skewed, unimodal OFD. The loss of common versus rare species has significantly different consequences for biodiversity (Gaston 2010). Because ζ diversity is sensitive to changes across species occupancy classes, comparisons of the form of ζ diversity decline can be used to signify ecologically relevant differences in the mechanics of species turnover.

The Relationship between ζ and β Diversity

ζ diversity can, of course, be used to calculate the range of existing incidence-based, pairwise β diversity and multiple-assemblage metrics. For example, Jaccard's (1900) similarity index is $\zeta_2/(2\zeta_1 - \zeta_2)$, and Sørensen's (1948) index is ζ_2/ζ_1. For multiple-assemblage similarity metrics, Koch's (1957) index of dispersity (i.e., taxonomic homogeneity) is $(\zeta_1/S_n - 1/n)/(1 - 1/n)$, and Diserud and Ødegaard's (2007) index is $(n - S_n/\zeta_1)/(n - 1)$. Clearly, pairwise indices are a combination of ζ_1 and ζ_2 only and do not consider higher-order ζ components (ζ_i where $i \geq 3$). Existing multiple-assemblage similarity metrics concern only ζ_1 and S_n and do not consider intermediate ζ components. For large n, if ζ_i is exponential, Koch's dispersity approaches e^{-b} and Diserud and Ødegaard's index approaches 1; if ζ_i follows a power law with $d \geq 1$, Koch's dispersity approaches 0 and Diserud and Ødegaard's index approaches 1. This means that multiple-assemblage metrics for large samples do not necessarily reflect assemblage similarity but rather are biased by the number of sites involved.

By contrast, the way in which ζ_i declines with i (i.e., coefficient $1/b$ or $1/d$) provides an unbiased measure of multiple-assemblage similarity. Higher coefficients of $1/b$ or $1/d$ suggest that assemblages are comparatively similar, with more shared species. Lower coefficients reflect substantially fewer shared species across sites. We propose ζ diversity here principally to represent species occupancy and turnover across independent samples. However, it can also be used for β diversity partitioning with aggregated samples in nested or hierarchical sampling schemes (such as Crist and Veech 2006), as depicted by equation (5). Finally, while pairwise β diversity does capture species losses and gains, it is insensitive to occupancy changes in common species (McGlinn and Hurlbert 2012). As shown above, ζ diversity is responsive to changes across the range of species occupancy classes.

Conclusion

We propose the use of ζ diversity components, that is, the average number of species shared by i assemblages, as a concept that (1) describes the structure of multispecies distributions and (2) unifies incidence-based assemblage patterns. ζ diversity is easy to calculate and provides a general framework from which other assemblage patterns can be explored and their distributions derived. We have shown this here for the SAC, the EER, the OFD, the scale dependence of diversity, and other indices of diversity partitioning. Unlike pairwise turnover measures, ζ diversity requires no assumptions to be made about site and species independence.

Perhaps the most important question is whether the use of ζ diversity will result in new insight or improved understanding of biodiversity pattern and process. We have already shown how ζ diversity partitioning informs a number of macroecological debates and how specific forms of ζ diversity relationships are directly related to particular ecological processes. With α and β as functions of ζ diversity, higher-order ζ components may better differentiate assemblages that are not distinguishable on the basis of α and β alone. ζ diversity may prove more sensitive to the detection of drivers of environmental change than pairwise turnover measures when used to examine changes in diversity across gradients or with declines in habitat quality. We suggest, for example, that the dynamics of rare and common species in assemblages and the mechanisms underpinning these may be better understood by using ζ components than by using pairwise turnover measures. In sum, ζ diversity provides a common currency for incidence-based biodiversity patterns and relationships, enabling direct and multivariate comparisons among them. It provides a measure of diversity and its scale dependence more comprehensive than existing metrics. As a result, future comparisons of ζ diversity within and across regions and systems may provide new insights on the processes that drive patterns in biodiversity.

Acknowledgments

We are grateful to S. L. Chown, T. O. Crist, K. J. Gaston, S. Harrison, B. D. Patterson, D. Srivastava, and an anonymous reviewer for constructive comments. C.H. is supported by the South African Research Chair Initiative and the Incentive Research Programme of the National Research Foundation and M.A.M. by a Monash University Larkins Fellowship. The Barro Colorado Island forest dynamics research project was made possible by National Science Foundation grants to S. P. Hubbell: DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, and DEB-7922197; by support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund, and numerous private individuals; and through the hard work of more than 100 people from 10 countries over the past 2 decades. The plot project is part the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

Literature Cited

Arrhenius, O. 1920. Distribution of the species over the area. Meddelanden från Vetenskapsakadmiens Nobelinstitut 4(7):1–6.

Atmar, W., and B. D. Patterson. 1995. The nestedness temperature calculator: a visual basic program, including 294 presence-absence matrices. AICS Research and the Field Museum, Chicago.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265–270.

Chao, A., L. Jost, S. C. Chiang, Y. H. Jiang, and R. L. Chazdon. 2008. A two-stage probabilistic approach to multiple-community similarity indices. Biometrics 64:1178–1186.

Chase, J. M., and T. M. Knight. 2013. Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. Ecology Letters 16:17–26.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society B: Biological Sciences 345:101–118.

Colwell, R. K., C. X. Mao, and J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology 85:2717–2727.

Condit, R. 1998. Tropical forest census plots. Springer, Berlin.

Conlisk, E., M. Bloxham, J. Conlisk, B. Enquist, and J. Harte. 2007. A new class of models of spatial distribution. Ecological Monographs 77:269–284.

Crist, T. O., and J. A. Veech. 2006. Additive partitioning of rarefaction curves and species-area relationships: unifying α-, β- and γ-diversity with sample size and habitat area. Ecology Letters 9:923–932.

Crist, T. O., J. A. Veech, J. C. Gering, and K. S. Summerville. 2003. Partitioning species diversity across landscapes and regions: a hierarchical analysis of α, β, and γ diversity. American Naturalist 162:734–743.

Dengler, J. 2009. Which function describes the species-area relationship best? a review and empirical evaluation. Journal of Biogeography 36:728–744.

Diserud, O. H., and F. Ødegaard. 2007. A multiple-site similarity measure. Biology Letters 3:20–22.

Drakare, S., J. J. Lennon, and H. Hillebrand. 2006. The imprint of the geographical, evolutionary and ecological context on species-area relationships. Ecology Letters 9:215–227.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12:42–58.

Gaston, K. J. 1996. Species-range-size distributions: patterns, mechanisms and implications. Trends in Ecology and Evolution 11:197–201.

———. 2010. Valuing common species. Science 327:154–155.

Gaston, K. J., and T. M. Blackburn. 2000. Pattern and process in macroecology. Blackwell Science, Oxford.

Gaston, K. J., and F. L. He. 2011. Species occurrence and occupancy. Pages 141–151 *in* A. E. Magurran and B. J. McGill, eds. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford.

Gaston, K. J., and J. H. Lawton. 1989. Insect herbivores on bracken do not support the core-satellite hypothesis. American Naturalist 134:761–777.

Gotelli, N. J., and R. K. Colwell. 2011. Estimating species richness. Pages 39–54 *in* A. E. Magurran and B. J. McGill, eds. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford.

Gotelli, N. J., and D. Simberloff. 1987. The distribution and abundance of tallgrass prairie plants: a test of the core-satellite hypothesis. American Naturalist 130:18–35.

Green, J. L., and A. Ostling. 2003. Endemics-area relationships: the influence of species dominance and spatial aggregation. Ecology 84:3090–3097.

Grilli, J., S. Azaele, J. R. Banavar, and A. Maritan. 2012. Spatial aggregation and the species-area relationship across scales. Journal of Theoretical Biology 313:87–97.

Hanski, I. 1982. Dynamics of regional distribution: the core and satellite species hypothesis. Oikos 38:210–221.

Hanski, I., and M. Gyllenberg. 1993. Two general metapopulation models and the core-satellite species hypothesis. American Naturalist 142:17–41.

Harrison, J. A., D. G. Allan, L. G. Underhill, M. Herremans, A. J. Tree, V. Parker, and C. J. Brown. 1997. The atlas of southern African birds. BirdLife South Africa, Johannesburg.

He, F. L., and R. Condit. 2007. The distribution of species: occupancy, scale, and rarity. Pages 32–50 *in* D. Storch, P. A. Marquet, and J. H. Brown, eds. Scaling biodiversity. Cambridge University Press, Cambridge.

He, F. L., and P. Legendre. 1996. On species-area relations. American Naturalist 148:719–737.

———. 2002. Species diversity patterns derived from species-area models. Ecology 83:1185–1198.

Hocutt, C. H., R. F. Denoncourt, and J. R. Stauffer. 1978. Fishes of the Greenbrier River, West Virginia, with drainage history of the central Appalachians. Journal of Biogeography 5:59–80.

Hubbell, S. P., R. Condit, and R. B. Foster. 2005. Barro Colorado Forest census plot data. Center for Tropical Forest Science. https://ctfs.arnarb.harvard.edu/webatlas/datasets/bci.

Hubbell, S. P., R. B. Foster, S. T. O'Brien, K. E. Harms, R. Condit, B. Wechsler, S. J. Wright, and S. L. de Lao. 1999. Light-gap disturbances, recruitment limitation, and tree diversity in a Neotropical forest. Science 283:554–557.

Hui, C., and M. A. McGeoch. 2007*a*. Modeling species distributions by breaking the assumption of self-similarity. Oikos 116:2097–2107.

———. 2007*b*. A self-similarity model for the occupancy frequency distribution. Theoretical Population Biology 71:61–70.

———. 2008. Does the self-similar species distribution model lead to unrealistic predictions? Ecology 89:2946–2952.

Hui, C., M. A. McGeoch, and M. Warren. 2006. A spatially explicit approach to estimating species occupancy and spatial correlation. Journal of Animal Ecology 75:140–147.

Jaccard, P. 1900. Contribution au problème de l'immigration postglaciaire de la flore alpine. Bulletin de la Société Vaudoise des Sciences Naturelles 36:87–130.

Jenkins, D. G. 2011. Ranked species occupancy curves reveal common patterns among diverse metacommunities. Global Ecology and Biogeography 20:486–497.

Jost, L. 2007. Partitioning diversity into independent alpha and beta components. Ecology 88:2427–2439.

Jost, L., A. Chao, and Chazdon, R. L. 2011. Compositional similarity and β (beta) diversity. Pages 66–84 *in* A. E. Magurran and B. J. McGill, eds. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford.

Kinzig, A. P., and J. Harte. 2000. Implications of endemics-area relationships for estimates of species extinctions. Ecology 81:3305–3311.

Koch, L. F. 1957. Index of biotal dispersity. Ecology 38:145–148.

Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. Oikos 76:5–13.

Magurran, A. E., and P. A. Henderson. 2003. Explaining the excess of rare species in natural species abundance distributions. Nature 422:714–716.

McGeoch, M. A., and K. J. Gaston. 2002. Occupancy frequency distributions: patterns, artefacts and mechanisms. Biological Reviews 77:311–331.

McGill, B. J. 2010. Towards a unification of unified theories of biodiversity. Ecology Letters 13:627–642.

McGlinn, D. J., and A. H. Hurlbert. 2012. Scale dependence in species turnover reflects variance in species occupancy. Ecology 93:294–302.

Munoz, F., P. Couteron, and B. R. Ramesh. 2008. Beta diversity in spatially implicit neutral models: a new way to assess species migration. American Naturalist 172:116–127.

Nee, S., R. D. Gregory, and R. M. May. 1991. Core and satellite species: theory and artifacts. Oikos 62:83–87.

Nekola, J. C., and P. S. White. 1999. The distance decay of similarity in biogeography and ecology. Journal of Biogeography 26:867–878.

Papp, L., and J. Izsák. 1997. Bimodality in occurrence classes: a direct consequence of lognormal or logarithmic series distribution of abundances: a numerical experimentation. Oikos 79:191–194.

Qian, H., and R. E. Ricklefs. 2012. Disentangling the effects of geographic distance and environmental dissimilarity on global patterns of species turnover. Global Ecology and Biogeography 21:341–351.

Raunkiær, C. 1934. Life forms and statistical plant geography. Oxford University Press, Oxford.

Sandel, B., L. Arge, B. Dalsgaard, R. G. Davies, K. J. Gaston, W. J. Sutherland, and J. C. Svenning. 2011. The influence of Late Quaternary climate-change velocity on species endemism. Science 334:660–664.

Scheiner, S. M., A. Chiarucci, G. A. Fox, M. R. Helmus, D. J. McGlinn, and M. R. Willig. 2011. The underpinnings of the relationship of species richness with space and time. Ecological Monographs 81:195–213.

Scheiner, S. M., and M. R. Willig. 2008. A general theory of ecology. Theoretical Ecology 1:21–28.

Sizling, A. L., W. E. Kunin, E. Sizlingova, J. Reif, and D. Storch. 2011. Between geometry and biology: the problem of universality of the species-area relationship. American Naturalist 178:602–611.

Soininen, J., J. J. Lennon, and H. Hillebrand. 2007. A multivariate analysis of beta diversity across organisms and environments. Ecology 88:2830–2838.

Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter 5:1–34.

Storch, D., P. Keil, and W. Jetz. 2012. Universal species-area and endemics-area relationships at continental scales. Nature 488:78–81.

Sutherland, J. P., and R. H. Karlson. 1977. Development and stability of the fouling community at Beaufort, North Carolina. Ecological Monographs 47:425–446.

Tjørve, E., and K. M. C. Tjørve. 2008. The species-area relationship, self-similarity, and the true meaning of the *z*-value. Ecology 89:3528–3533.

Veech, J. A., and T. O. Crist. 2010. Toward a unified view of diversity partitioning. Ecology 91:1988–1992.

Weinberg, S. 1996. The quantum theory of fields. Cambridge University Press, Cambridge.

Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30:279–338.

Associate Editor: Susan Harrison
Editor: Judith L. Bronstein

© 2014 by The University of Chicago. All rights reserved. DOI: 10.1086/678125

Appendix from C. Hui and M. A. McGeoch, “Zeta Diversity as a Concept and Metric That Unifies Incidence-Based Biodiversity Patterns”
(Am. Nat., vol. 184, no. 5, p. 684)

Zeta Diversity Partitioning Unifies Species Accumulation, Endemism, and Assemblage Occupancy Patterns

Incidence-Based Zeta Diversity Partitioning

Here we derive the equations presented in “Zeta Diversity Partitioning” and “The Endemics-Effort Relationship.” Recall the definition of ζ_i being the average number of shared species among a number i of sites. Let C_n^i $(= n!/(i!(n - i)!))$ be the number of combinations of choosing i from n sites. For $n = 1$, it is straightforward to have

$$S_1 = \zeta_1 = C_1^1\zeta_1, \tag{A1}$$

$$F_{1,1} = \zeta_1 = C_1^1C_0^0\zeta_1. \tag{A2}$$

For $n = 2$, according to the Venn diagram in figure 1, we have

$$S_2 = 2\zeta_1 - \zeta_2 = C_2^1\zeta_1 - C_2^2\zeta_2, \tag{A3}$$

$$F_{1,2} = 2\zeta_1 - 2\zeta_2 = C_2^1(C_1^0\zeta_1 - C_1^1\zeta_2), \tag{A4}$$

$$F_{2,2} = \zeta_2 = C_2^2C_0^0\zeta_2. \tag{A5}$$

For $n =3$, we have

$$S_3 = 3\zeta_1 - 3\zeta_2 + \zeta_3 = C_3^1\zeta_1 - C_3^2\zeta_2 + C_3^3\zeta_3, \tag{A6}$$

$$F_{1,3} = 3\zeta_1 - 6\zeta_2 + 3\zeta_3 = C_3^1(C_2^0\zeta_1 - C_2^1\zeta_2 + C_2^2\zeta_3), \tag{A7}$$

$$F_{2,3} = 3\zeta_2 - 3\zeta_3 = C_3^2(C_1^0\zeta_1 - C_1^1\zeta_2), \tag{A8}$$

$$F_{3,3} = \zeta_3 = C_3^3C_0^0\zeta_3. \tag{A9}$$

Comparing S_1, S_2, and S_3, we can have equation (1),

$$S_n = \sum_{k=1}^{n}(-1)^{k+1}\cdot C_n^k\cdot\zeta_k. \tag{A10}$$

Comparing $F_{1,1}$, $F_{1,2}$, $F_{2,2}$, $F_{1,3}$, $F_{2,3}$, and $F_{3,3}$, we can derive equation (2),

$$F_{i,n} = C_n^i\sum_{k=1}^{n-i+1}(-1)^{k+1}\cdot C_{n-i}^{k-1}\cdot\zeta_{i+k-1}. \tag{A11}$$

Now, considering the difference between S_n and S_{n-1}, we have

$$\begin{aligned} S_n - S_{n-1} &= \sum_{k=1}^{n} (-1)^{k+1} \cdot C_n^k \cdot \zeta_k - \sum_{k=1}^{n-1} (-1)^{k+1} \cdot C_{n-1}^k \cdot \zeta_k \\ &= \sum_{k=1}^{n-1} (-1)^{k+1} (C_n^k - C_{n-1}^k) \zeta_k + (-1)^{n+1} C_n^n \zeta_n \\ &= \sum_{k=1}^{n-1} (-1)^{k+1} C_{n-1}^{k-1} \zeta_k + (-1)^{n+1} C_{n-1}^{n-1} \zeta_n \\ &= \sum_{k=1}^{n} (-1)^{k+1} C_{n-1}^{k-1} \zeta_k \\ &= \frac{F_{1,n}}{C_n^1} = \frac{F_{1,n}}{n}. \end{aligned} \tag{A12}$$

That is, adding one extra site to a survey with $n - 1$ sites will add $F_{1,n}/n$ new species. Therefore, we have

$$\begin{aligned} S_{n+m} &= S_{n+m-1} + \frac{F_{1,n+m}}{n+m} \\ &= S_{n+m-2} + \frac{F_{1,n+m-1}}{n+m-1} + \frac{F_{1,n+m}}{n+m} \\ &\cdots \\ &= S_n + \sum_{k=n+1}^{n+m} \frac{F_{1,k}}{k}. \end{aligned} \tag{A13}$$

When m approaches infinity, we have equation (3),

$$S_\infty = S_n + \sum_{k=n+1}^{\infty} \frac{F_{1,k}}{k}. \tag{A14}$$

For i selected sites within a total of n sites surveyed, the local endemics represent species that occur only in k ($= 1, 2, 3, ..., i$) sites of these selected i sites. These local endemics with occupancy k are also species occupying only k sites among the n sites. Therefore, the number of local endemics with an occupancy of k is $C_i^k F_{k,n}/C_n^k$. Alternatively, we can also count the number of local endemics by subtracting those species that do not occur exclusively in the $n - i$ sites from the total surveyed species, $S_n - S_{n-i}$, or we can sum the number of new species discovered when sequentially adding the i sites to the rest of $n - i$ sites, $\sum_{k=n-i+1}^{n} F_{1,k}/k$. These three different ways are mathematically equivalent, and thus we have equation (8),

$$E_{i,n} = \sum_{k=1}^{i} C_i^k \cdot \frac{F_{k,n}}{C_n^k} = S_n - S_{n-i} = \sum_{k=n-i+1}^{n} \frac{F_{1,k}}{k}. \tag{A15}$$

Given a binary (0/1) species-by-site matrix as a .csv file named test.csv in the working directory of R, with rows representing different species and columns different sites, ζ diversity can be calculated with the following R script:[1]

```
data<-read.csv(“test.csv”,header=FALSE)
x<-dim(data)[2]
zeta<-numeric()
u<-numeric()
for(j in 1:x){
for(z in 1:1000){
sam<-sample(1:x,j,replace=FALSE)
u[z]<-sum(apply(data[sam],1,prod))}
zeta[j]<-mean(u)}
plot(1:x,zeta)
write.csv(zeta,file=“zeta.csv”)
```

[1] Code that appears in the *American Naturalist* has not been peer-reviewed, nor does the journal provide support.

Results of this R script will be presented as a figure (as fig. 2*A*, 2*B*) in the RGui, and the exact values of ζ diversity, ζ_i for $i = 1, ..., n$ (the number of sites), will be exported to a file named zeta.csv in the working directory. A fully functioning R package with all incidence-based metrics and macroecological patterns, together with ζ diversity calculation, is currently under development.

Form of the Relationship between Zeta Diversity and Sample Number (Zeta Decline)

We compared the performance of seven parametric models of ζ diversity decline, that is, ζ_i as a function of *i*, for the 291 species-by-site matrices listed in the supplementary table. Specifically, we tested all six models that have been proposed in literature for abundance or occupancy rank curves (Jenkins 2011), plus the power-law model, which satisfies the positive and monotonically declining trajectory of ζ diversity:

Exponential convex:

$$\zeta_i = c_1 + c_2 \cdot \exp(c_3 \cdot i), \tag{A16}$$

linear:

$$\zeta_i = c_1 + c_2 \cdot i, \tag{A17}$$

sigmoidal symmetric (sig_sym):

$$\zeta_i = \frac{c_1}{1 + \exp(c_2 + c_3 \cdot i)}, \tag{A18}$$

sigmoidal asymmetric (sig_asym):

$$\zeta_i = c_1(1 + \exp(c_2 \cdot i^{c_3})), \tag{A19}$$

exponential (= lognormal = exponential concave):

$$\zeta_i = c_1 \exp(c_2 \cdot i), \tag{A20}$$

power law (power):

$$\zeta_i = c_1 \cdot i^{c_2}, \tag{A21}$$

logarithmic (log):

$$\zeta_i = c_1 + c_2 \cdot \log(i). \tag{A22}$$

In the above parametric models, c_1, c_2, and c_3 are model parameters. The nonlinear fit of the model to each of the 291 species-by-site matrices was tested with “NonlinearModelFit” in Mathematica 8.0 (Wolfram Research). For each matrix, we report in the supplementary table the estimate, standard error, *t* statistics, and *P* value for each parameter and the adjusted R^2 and Akaike Information Criteria (AIC) scores for each model. For the 291 matrices, adjusted R^2 suggests that the power-law model is the best fit for 167 matrices, the exponential model for 76 matrices, the sig_sym model for 25 matrices, the sig_asym model for 9 matrices, the logarithmic model for 12 matrices, the linear model for 2 matrices, and the exponential convex model for no matrices (see fig. A1 for a summary). Use of the AIC produced very similar results, suggesting that the power-law model is the best fit for 166 matrices, the exponential model for 72 matrices, the sig_sym model for 29 matrices, the sig_asym model for 12 matrices, the logarithmic model for 10 matrices, the linear model for 2 matrices, and the exponential convex model for no matrices (see the supplementary table). The power-law and exponential models are the best-fitting parametric models for ζ diversity scaling, together representing more than 80% of the matrices. We therefore discuss only these two specific forms of ζ diversity scaling and their incidence-based diversity patterns in detail.

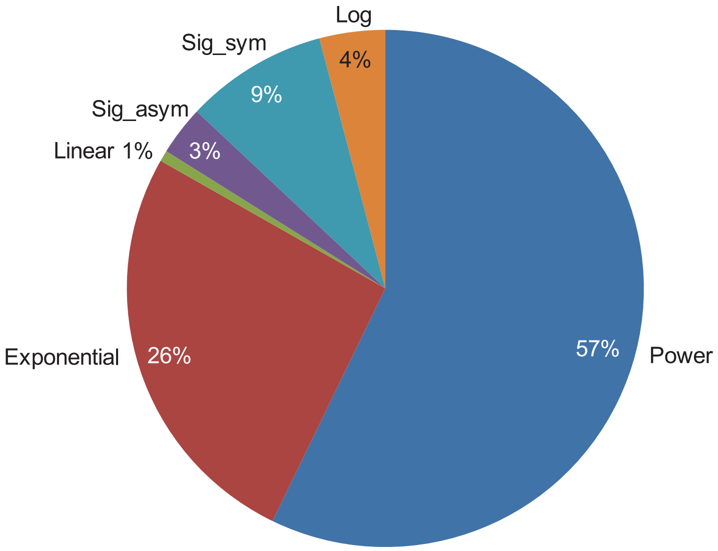

Figure A1: Percentage of different parametric models that best fit the relationship between ζ diversity and sample number (ζ decline) for 291 species-by-site matrices (see the supplementary table for details of the results), according to the adjusted R^2. Sig_sym = sigmoidal symmetric; sig_asym = sigmoidal asymmetric.

Zeta Diversity Scaling with Sample Grain and Hierarchical Diversity Partitioning

As with all diversity metrics, ζ diversity changes with sampling grain; when the sampling grain increases, the number of species in each sample increases. This is a consequence of the scale dependency of species distributions, as demonstrated by numerous models based on either point-process or probability theory (e.g., Nachman 1981; Wright 1991; Hanski and Gyllenberg 1997; Kunin 1998; He and Gaston 2000, 2003; He et al. 2002; Hui et al. 2006; Hui 2009; Zillio and He 2010; Azaele et al. 2012; see a review by Barwell et al. [2014]). Similarly, when a collection of samples are pooled to form increasingly larger sample grains, the α, β, and δ diversity components of this hierarchical sample design will also change. Crist et al. (2003) and Crist and Veech (2006) provide a framework for α and β diversity partitioning across hierarchical levels. Here we derive the general form of ζ diversity partitioning when pooling multiple samples to form increasingly larger sample grain in the same way.

Consider the simplest scenario where two samples are pooled, that is, the sampling grain increases from the original α to 2α. We have

$$\zeta_1(2\alpha) = S_2(\alpha) = 2\zeta_1(\alpha) - \zeta_2(\alpha). \tag{A23}$$

ζ diversity $\zeta_2(2\alpha)$ represents the number of shared species between two clusters, with each cluster formed by two samples. The number of shared species between two clusters can be expressed by a combination of species with specific occupancies and occurring in both clusters,

$$\zeta_2(2\alpha) = \frac{C_2^1C_2^1}{C_4^2}F_{2,4} + \frac{C_2^1C_2^2 + C_2^2C_2^1}{C_4^3}F_{3,4} + \frac{C_2^2C_2^2}{C_4^4}F_{4,4}. \tag{A24}$$

Following the same logic, the number of shared species among three clusters can be estimated as follows,

$$\begin{aligned}\zeta_3(2\alpha) = {} & \frac{C_2^1C_2^1C_2^1}{C_6^3}F_{3,6} + \frac{C_2^1C_2^1C_2^2 + C_2^1C_2^2C_2^1 + C_2^2C_2^1C_2^1}{C_6^4}F_{4,6} \\ & + \frac{C_2^1C_2^2C_2^2 + C_2^2C_2^1C_2^2 + C_2^2C_2^2C_2^1}{C_6^5}F_{5,6} + \frac{C_2^2C_2^2C_2^2}{C_6^6}F_{6,6}.\end{aligned} \tag{A25}$$

Consider a slightly complicated scenario where four samples are pooled together as a cluster. We have the number of shared species between two clusters as follows:

$$\begin{aligned}\zeta_2(4\alpha) = {} & \frac{C_4^1C_4^1}{C_8^2}F_{2,8} + \frac{C_4^1C_4^2 + C_4^2C_4^1}{C_8^3}F_{3,8} + \frac{C_4^1C_4^3 + C_4^2C_4^2 + C_4^3C_4^1}{C_8^4}F_{4,8} \\ & + \frac{C_4^1C_4^4 + C_4^2C_4^3 + C_4^3C_4^2 + C_4^4C_4^1}{C_8^5}F_{5,8} + \frac{C_4^2C_4^4 + C_4^3C_4^3 + C_4^4C_4^2}{C_8^6}F_{6,8} \\ & + \frac{C_4^3C_4^4 + C_4^4C_4^3}{C_8^7}F_{7,8} + \frac{C_4^4C_4^4}{C_8^8}F_{8,8}.\end{aligned} \tag{A26}$$

By deduction, we have the general form of the number of shared species among *n* clusters when pooling *m* samples to form a cluster as (eq. [4])

$$\zeta_n(m) = \sum_{k=n}^{n\times m} \frac{\sum_{x_j\geq 1,\, \Sigma_{x_j}=k} \prod_{j=1}^{n} C_m^{x_j}}{C_{n\times m}^k} F_{k,n\times m}. \tag{A27}$$

This provides a general form that describes the relationship between ζ diversity and sample grain. Of particular interest to hierarchical diversity partitioning is the number of species shared by two clusters, where the first cluster is formed by pooling m_1 samples and the second cluster by pooling another m_2 samples, $\zeta_2(m_1, m_2)$. Slightly modifying the above equation, we have equation (5),

$$\zeta_2(m_1, m_2) = \sum_{k=2}^{m_1+m_2} \frac{\sum_{i=1}^{k-1} C_{m_1}^i C_{m_2}^{k-i}}{C_{m_1+m_2}^k} F_{k,m_1+m_2}. \tag{A28}$$

Together with $\zeta_1(m_1) = S_{m_1}$ and $\zeta_2(m_2) = S_{m_2}$, Crist and Veech's (2006) hierarchical diversity partitioning approach can readily be adopted for ζ. A clear message here is that both β and ζ diversity components are sensitive to sample grain and that their values are not independent across scales, echoing the scale-heritage assumption, as shown in Hui and McGeoch (2008).

Zeta Diversity Decay with Distance

Here we discuss how ζ diversity components are affected by the average distance between random samples. The average distance *u* between two random samples within a compact convex sampling area *A* can be estimated as

$$u = s \cdot \lambda(A) \tag{A29}$$

(Burgstaller and Pillichshammer 2009), where $\lambda(A)$ is the maximum distance between two samples within the area (or sampling extent), often called the diameter of *A*; *s* is a constant depending on the shape of *A* (e.g., $s = 1/3$ for a transect, $s = 0.452$ for a disk, and $s = 0.369$ for a square). Specifically, if we have *n* samples within the sampling area, with u_{ij} the distance between samples *i* and *j*, the average distance between two samples, u_2, is

$$u_2 = \frac{1}{C_n^2}\sum_{i\neq j} u_{ij}. \tag{A30}$$

The average of the distances between three random samples, u_3, is

$$u_3 = \frac{1}{C_n^3}\sum_{i\neq j\neq k} \frac{u_{ij} + u_{jk} + u_{ki}}{3} = u_2. \tag{A31}$$

Therefore, the average distance between a number *i* of random samples is a constant,

$$u_i = u, \tag{A32}$$

for $1 \leq I \leq n$. Consequently, when we increase the sampling extent, the average distance between samples will increase.

The distance decay of similarity is a pervasive characteristic of geographical and ecological patterns (Taylor 1971; Nekola and White 1999). On the basis of a comparison of 10 models that predict the downscaling of occupancy (that is, extrapolating fine-scale occupancy values from coarse-scale occupancy), Barwell et al. (2014) concluded that our model based on joint-count statistics and pair approximation performed best (Hui et al. 2006; Hui 2009). Therefore, here we derive the distance decay of similarity for ζ_2, using the same method. Let $\zeta_2(u)$ be the number of shared species between two random samples with the average distance between them being u, and let $\zeta_2(1) = \zeta_2$ be the number of shared species between two random samples that are one distance unit apart. Moreover, let p_+ and p_0 be the respective probabilities of finding a species present and absent in one randomly selected sample. Let $q_{0/+}$ be the conditional probability that, knowing a species is present (+) in a focal sample, the species is found absent (0) in a random sample one distance unit away from the focal sample. Similarly, we can further define $q_{+/+}$, $q_{+/0}$, and $q_{0/0}$. We have

$$p_0 = 1 - p_{+,} \tag{A33}$$

$$q_{0/+} = 1 - q_{+/+}, \tag{A34}$$

$$q_{+/0} = \frac{(1 - q_{+/+})p_+}{1 - p_+}, \tag{A35}$$

$$q_{0/0} = 1 - \frac{(1 - q_{+/+})p_+}{1 - p_+}. \tag{A36}$$

That is, all these probabilities and conditional probabilities can be expressed by p_+ and $q_{+/+}$. Let $Q(u)$ be the conditional probability of finding a species present in a sample to be also present in another sample distance u away. We have $Q(0) = 1$ and $Q(1) = q_{+/+}$. Following Hui et al. (2006), we can have

$$Q(2) = q_{0/+}q_{+/0} + q_{+/+}q_{+/+}. \tag{A37}$$

This can be explained as follows: there are three samples A, B, and C, aligned along a line consecutively, with A and B one distance unit apart, B and C one distance unit apart, and therefore A and C two distance units apart. The above formula depicts the probability that a species, already occurring in sample A, also occurs in sample C, regardless of whether it occurs in sample B. It equals the sum of probabilities for the occurrences of ABC being +0+ and +++, where + and 0 indicate presence and absence, respectively. The first term on the right of equation (A37) represents the first scenario and the second term the second scenario. Similarly, we have

$$Q(3) = q_{0/+}q_{0/0}q_{+/0} + q_{0/+}q_{+/0}q_{+/+} + q_{+/+}q_{0/+}q_{+/0} + q_{+/+}q_{+/+}q_{+/+.} \tag{A38}$$

The four terms on the right represent, in order, the scenarios +00+, +0++, ++0+, and ++++. It is worth noting that we already know that the species occurs in both the first and the last samples; thus, only the statuses of those samples in the middle are unknown (i.e., they can be either + or 0). Deductively, we have

$$Q(u) = q_{+/+} \cdot Q(u-1) + q_{0/+}q_{+/0}\sum_{i=1}^{u-1} q_{0/0}^{i-1} \cdot Q(u-1-i). \tag{A39}$$

Let $q_{+/+} = \zeta_2/\zeta_1$ and $p_+ = \zeta_1/S$, where S is the number of species in the sampling extent. We then have the following distance decay of similarity for ζ (eq. [6]),

$$\zeta_2(u) = \zeta_1 Q(u), \tag{A40}$$

where $Q(0) = 1$, $Q(1) = \zeta_2/\zeta_1$, and

$$Q(u) = \frac{\zeta_2}{\zeta_1}Q(u-1) + \frac{(\zeta_1 - \zeta_2)^2}{\zeta_1(S - \zeta_1)}\sum_{i=1}^{u-1}\left(1 - \frac{\zeta_1 - \zeta_2}{S - \zeta_1}\right)^{i-1} Q(u-1-i). \tag{A41}$$

Clearly, $Q(u)$ is a function of u, ζ_1, ζ_2, and S.

The direct deduction of $\zeta_n(u)$ for $n \geq 3$ is rather formidable. Indeed, $\zeta_n(u)$ is related to the n-point correlation function, which remains a hotly contested dilemma in theoretical physics, astronomics, quantum mechanics, and material science (e.g., Weinberg 1996; Baniassadi et al. 2012). Instead, following Hui et al.'s (2006) method, we here provide the Bayesian solution for higher orders of $\zeta_n(u)$. Given $n - 1$ samples an average distance u apart from each other and a known $\zeta_{n-1}(u)$, we have

$$\frac{\zeta_n(u)}{\zeta_{n-1}(u)} = \frac{p_+Q(u)^{n-1}}{p_+Q(u)^{n-1} + p_0Q'(u)^{n-1}}, \tag{A42}$$

where

$$Q'(u) = q_{+/0} \cdot Q(u-1) + q_{+/0}\sum_{i=1}^{u-1} q_{0/0}^{i} \cdot Q(u-1-i). \tag{A43}$$

As shown in figure A2, with the increase in sampling extent, the average distance between random samples u will increase, and thus ζ diversity components will decline. Clearly, according to the above equations and figure A2, $\zeta_i(u)$ declines with both the number of samples i (i.e., ζ diversity decline) and the average distance between samples u (i.e., distance decay). When $u = 0$, all samples collapse into one, and ζ diversity components become a constant ($=\zeta_1$); when $i = 2$, $\zeta_2(u)/\zeta_1$ becomes the typical distance decay of similarity, declining from 1 with an increase in distance u. Of course, other indices, such as Jaccard's index $J = \zeta_2(u)/(2\zeta_1 - \zeta_2(u))$, can also be used for depicting the distance decay of similarity, although the ζ diversity ratio $\zeta_2(u)/\zeta_1$ provides a normalized index for expressing assemblage similarity.

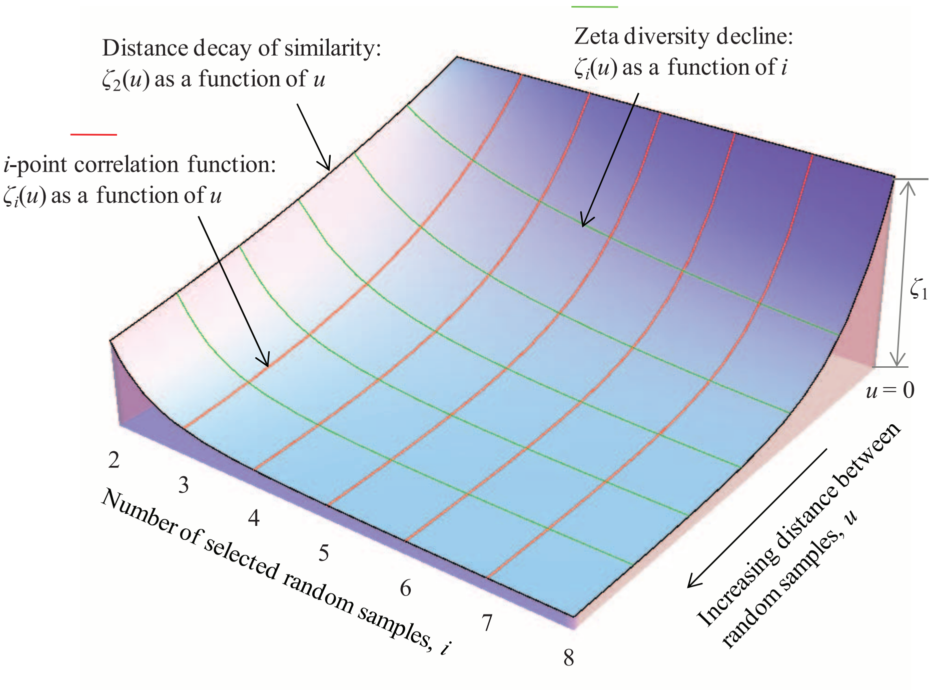

Figure A2: ζ diversity $\zeta_i(u)$ as a function of the number of random samples i (ζ decline; green lines) and the average distance between random samples u (distance decay of ζ diversity; red lines).

Zeta Diversity Decline and the Species-Area Relationship

Here we provide the detailed derivation for the relationship between the coefficient d of the power-law form of ζ decline ($\zeta_i = c \cdot i^{-d}$) and the exponent z of the power-law species-area relationship (SAR; $S_i \sim i^z$). For the power-law SAR, we have

$$\left(\frac{S_n}{S_{n+1}}\right)^{1/z} = \frac{n}{n+1}. \tag{A44}$$

Therefore, we have the general form of z-d relationship as follows:

$$z = \frac{\ln (S_n/S_{n+1})}{\ln (n/(n+1))}. \tag{A45}$$

Clearly, the z-d relationship is dependent on n. According to equation (1) of S_n and $\zeta_i = c \cdot i^{-d}$, we have, for $n = 1$,

$$2 - 2^z = \frac{\zeta_2}{\zeta_1}, \tag{A46}$$

that is,

$$z = \frac{\ln (2 - 2^{-d})}{\ln 2}. \tag{A47}$$

Note that ζ_2/ζ_1 represents the proportion of species shared between two areas. This special form has been derived by Tjørve and Tjørve (2008) and shown to be scale dependent by McGlinn and Hurlbert (2012), analogous to the n dependency in the above general formula (eq. [A45]) of the z-d relationship. For $n = 2$, we have

$$z = \frac{\ln((2 - 2^{-d})/(3 - 3 \cdot 2^{-d} + 3^{-d}))}{\ln(2/3)}. \tag{A48}$$

For larger values of n, the z-d relationship can become complicated but can still be readily calculated according to the above general formula (eq. [A45]; e.g., see the z-d relationship under different values of n in fig. A3). Practically, for larger values of n (>20), the z-d relationship resembles a polynomial function, $z = -0.024d^2 + 0.286d$, which can be estimated directly via nonlinear regression on the empirical data (as in fig. 4).

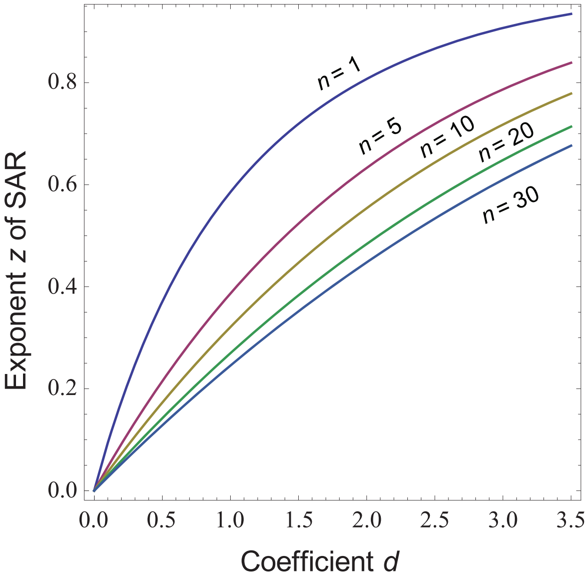

Figure A3: Relationship between coefficient d of the power-law form of ζ diversity and the exponent z for the Arrhenius species-area relationship (SAR), under different values of n in its general form $z = \ln(S_n/S_{n+1})/\ln(n/(n + 1))$.

Literature Cited Only in the Appendix

Azaele, S., S. J. Cornell, and W. E. Kunin. 2012. Downscaling species occupancy from coarse spatial scales. Ecological Applications 22:1004–1014.

Baniassadi, M., S. Ahzi, H. Garmestani, D. Ruch, and Y. Remond. 2012. New approximate solution for *N*-point correlation functions for heterogeneous materials. Journal of the Mechanics and Physics of Solids 60:104–119.

Barwell, L., S. Azaele, W. E. Kunin, and N. J. B. Isaac. 2014. Can coarse-grain patterns in insect atlas data predict local occupancy? Diversity and Distributions 20:895–907.

Burgstaller, B., and F. Pillichshammer. 2009. The average distance between two points. Bulletin of Australian Mathematical Society 80:353–359.

Hanski, I., and M. Gyllenberg. 1997. Uniting two general patterns in the distribution of species. Science 284:397–400.

He, F., and K. J. Gaston. 2000. Occupancy-abundance relationships and sampling scales. Ecography 23:503–511.

———. 2003. Occupancy, spatial variance, and the abundance of species. American Naturalist 162:366–375.

He, F., K. J. Gaston, and J. Wu. 2002. On species occupancy-abundance models. Ecoscience 9:119–126.

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

Kunin, W. E. 1998. Extrapolating species abundance across spatial scales. Science 281:1513–1515.

Nachman, G. 1981. A mathematical model of the functional relationship between the density and spatial distribution of a population. Journal of Animal Ecology 50:453–463.

Taylor, P. J. 1971. Distance transformation and distance decay functions. Geographical Analysis 3:221–238.

Wright, D. H. 1991. Correlations between incidence and abundance are expected by chance. Journal of Biogeography 18: 463–466.

Zillio, T., and F. He. 2010. Modeling spatial aggregation of finite populations. Ecology 91:3698–3706.