Ecology, 83(12), 2002, pp. 3344–3356

SAMPLING THE SPECIES COMPOSITION OF A LANDSCAPE

JOSHUA B. PLOTKIN[1,2,3] AND HELENE C. MULLER-LANDAU[2,4]

[1]*Institute for Advanced Study, Olden Lane, Princeton, New Jersey 08540 USA*
[2]*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544 USA*

Abstract. The abundances and spatial distribution of species is central to biogeography and conservation. Several theories have been offered to explain landscape-scale species distribution patterns. The verification of biogeographic theories, as well as conservation decisions, must be based upon empirical data gathered from necessarily restricted censuses. It is necessary, therefore, to understand the relationship between an underlying landscape-scale pattern and the corresponding pattern it produces upon sampling small subregions. The similarity of species composition between two samples depends not only on the species composition of the underlying landscape from which the samples are drawn, but also on the underlying distribution of species abundances, the degree of conspecific spatial clustering, and sample size. In this paper, we investigate how sampling expectations change depending upon species abundance distributions and upon spatial distributions. We derive analytical results for the expected species overlap between two sampled regions under a wide range of conditions. We compare these results with data from a 50-ha tropical forest census. These methodologies provide useful tools for assessing beta diversity, for testing macro-ecological theory, and for designing landscape-scale sampling schemes.

Key words: beta diversity; landscape-scale species distribution; negative binomial; sampling diversity; similiarity indices; spatial aggregation; tropical forest.

INTRODUCTION

Spatial patterns of species diversity are of great interest to ecologists. This interest is in part practical: to best conserve biodiversity, we must know where species richness is highest and how species assemblages change in space (Cody 1986). But diversity patterns are also of theoretical interest, providing material to test theories of why diversity varies among sites and how species turnover increases with intersite distance (Condit et al. 2002).

Species diversity in a landscape has long been divided into two parts: alpha diversity and beta diversity (Whittaker 1960). Most studies have focused on alpha diversity, the diversity of species within individual sites. Yet even the most diverse sites hold but a tiny fraction of the planet's species. The high diversity of species overall is due mainly to beta diversity, the change in species composition between sites. An understanding of the mechanisms and patterns of beta diversity is critical for understanding the local to landscape scaling of biodiversity in general.

Despite its obvious importance, beta diversity has received relatively little attention in the empirical literature. Those empirical studies that have examined species turnover across a landscape have used a variety of methods (not always comparable), and have focused on qualitative patterns (e.g., Routledge 1977, Cody 1986, Harrison et al. 1992, Harrison 1997, Potts et al. 2002). Recent contributions (Chave and Leigh, *in press*, Leigh et al. *in press*) make quantitative predictions regarding the spatial turnover of species composition under a neutral model of community dynamics (Hubbell 1995, 1997, 2001). Such work provides theoretical predictions of beta diversity patterns that can be tested with empirical data sampled from a landscape (Condit et al. 2002).

In order to assess landscape-scale species composition patterns with limited data, we must first understand the sampling distributions of our measured statistics. That is, given some underlying species similarity between two communities, what similarity do we expect to observe between two small samples from those communities? And thus what can we deduce about the underlying species similarity between two communities from the observed similarity between two small samples? Unfortunately, this question cannot in general be answered exactly. Previous studies have used computer simulations to investigate the robustness of community similarity measures to sample sizes (e.g., Morisita 1959, Ricklefs and Lau 1980, Wolda 1981). These studies show that similarity indices computed from small samples almost always underestimate the true similarity of the underlying communities from which the samples are drawn. The sampling distribution of similarity indices is so fundamental to biogeography that numerical approximations fitted to com-

Manuscript received 27 August 2001; revised 25 April 2002; accepted 3 May 2002.

[3] E-mail: plotkin@ias.edu

[4] Present address: National Center for Ecological Analysis and Synthesis, 735 State Street, Suite 300, Santa Barbara, California 93101-5504 USA.

puter simulations (Wolda 1981) have even been incorporated into popular textbooks on ecological methodology (Krebs 1999).

Previous studies of similarity indices assume that individuals are all sampled independently and randomly in space. For many types of communities and sampling practices this assumption is severely violated. For example, studies of contiguous plots in tropical tree communities reveal that nearly every species is spatially aggregated (He et al. 1997, Condit et al. 2000), and that this aggregation significantly affects large-scale ecological patterns such as species-area curves (Plotkin et al. 2000*b*). Clustering of conspecifics will further bias measures of species similarity between sites that are already biased by small sample sizes.

In this paper, we investigate how indices of community composition are affected by conspecific clustering within samples, as well as by abundance distributions and by sample size. We incorporate conspecific clustering into the theory of sampling distributions by using the negative binomial distribution. We derive analytic expressions for the expected value of the most basic similarity measure, the proportion of species in common to two samples, under a large variety of conditions. In order to test our results, we use information on conspecific clustering and the abundance distribution in a 50-ha tropical forest plot to calculate expected similarities between subplots, and compare these to observed similarities. Finally, we present numerical results for a variety of similarity indices, showing the effects of conspecific clumping on their sampling means and variances. We conclude by discussing the implications of our results for biogeographic theory and for landscape-scale sampling design.

Species Abundance Distributions

In this section, we review several common models of species abundances for which we will later compute the expected similarity between subsamples. The relative abundances of species in a large region will have a significant effect on the observed similarity between two small samples from the region. If a few common species dominate, then similarity in species composition between samples will be high; if there are many rare species, similarity will be low.

The distribution of abundances within ecological communities is typically described in one of two ways. The number of species in different abundance classes (often doubling classes) may be plotted as a histogram (Preston 1962). Alternatively, species may be ranked from most abundant to least abundant, and their abundance, generally on a log scale, plotted against their rank to produce a rank-abundance curve. These two alternatives are entirely equivalent and the connection between them is straightforward (May 1975). Fig. 1 illustrates five continuous probability distributions commonly used to model abundances of species. The differences between the distributions are more apparent when plotted as density functions rather than rank-abundance curves.

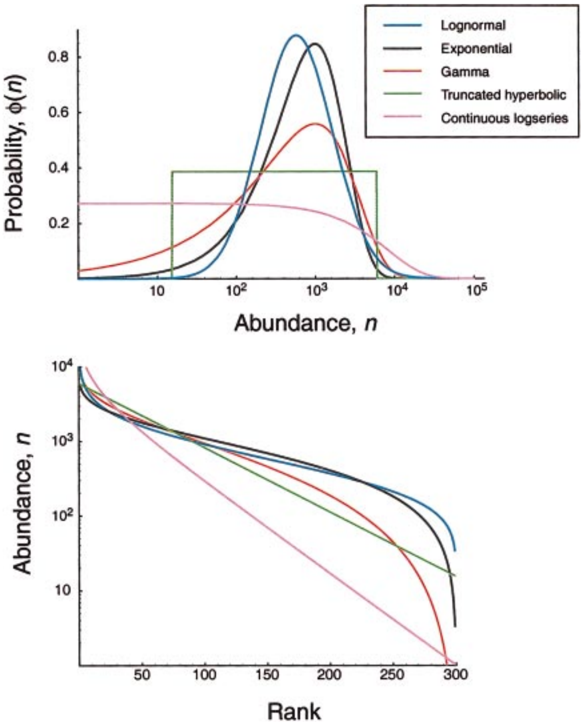

FIG. 1. Examples of five continuous abundance distributions represented as probability density functions on a log scale (top), and as rank–abundance curves (bottom). Parameter values are chosen so that all distributions have the same mean, 1000. The truncated hyperbolic, lognormal, and gamma distributions have the same variance, 2×10^6. The variance of the exponential distribution is 10^6. The variance of the continuous logseries distribution is 7.5×10^6. Exact parameter values are as follows: exponential $\lambda = 1/1000$; gamma $\beta = 1/2$ and $\lambda = 1/2000$; lognormal $\mu = 2 \log(1000) - \log(1000\sqrt{3})$ and $\sigma = [2 \log(1000\sqrt{3} - 2 \log(1000)]^{½}$; truncated hyperbolic $m = 15.2954$ and $M = 5984.7$; continuous logseries $x = 0.999882$.

For the purpose of analytically calculating the expected species overlap between sites, it is convenient to summarize abundances with a continuous probability density function $\phi(n)$. The expression $\phi(n)\ dn$ denotes the probability that a randomly chosen species from the community will have an abundance between n and $n + dn$. The continuous density function is normalized to total probability one: $\int_0^\infty \phi(n)\ dn = 1$. In reality, the distribution of species abundances is discrete. Nevertheless, a continuous approximation is useful and commonly employed in the ecological literature (May 1975). Furthermore, we will demonstrate that continuous abundance distributions yield almost precisely the same results as their discrete analogues. Table 1 summarizes the abundance distributions reviewed below. The table indicates the names of the distributions, the discrete analogues (where applicable), and the relevant equations.

TABLE 1. A summary of the species abundance distributions, with the discrete analogue of each continuous abundance distribution, where applicable, and the relevant equations for the probability density functions.

Continuous distribution	Discrete analogue	Eqs.
Lognormal		1
Truncated hyperbolic	geometric series	2
Exponential	broken stick	3
Gamma		4
Continuous logseries	logseries	5–7

Lognormal distribution

A popular model of species abundances is the lognormal distribution, for which the histogram of species numbers is normal when abundances are binned (plotted) on a log scale. The lognormal distribution takes the form

$$\phi(n) = \frac{1}{n\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln(n) - \mu)^2}{2\sigma^2}\right]. \quad (1)$$

This distribution is expected to arise if many independent factors act multiplicatively to determine abundances (May 1975). The lognormal distribution has two free parameters, μ and σ. A one-parameter subfamily called the canonical lognormal is also common in the literature (Preston 1962, May 1975).

Truncated hyperbolic distribution

Another common abundance distribution is the so-called geometric series distribution, whose rank–abundance curve is linear when the abundance axis is logarithmic (Fig. 1b). The continuous analogue of the geometric series is given by a truncated hyperbolic probability density function:

$$\phi(n) = \begin{cases} \dfrac{1}{n\ln\left(\dfrac{M}{m}\right)} & \text{for } m \le n \le M \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

This distribution has two parameters, m and M, which represent the minimal and maximal abundances in the community. This distribution arises when the community is dominated by a single resource factor, and if division of this resource proceeds in a hierarchical manner. The most successful species occupies a fraction of the resource, and the next most abundant species occupies the same fraction of the remaining resource, and so on. Ecologists have often referred to the truncated-hyperbolic as the ‘‘exponential’’ distribution, because of the shape of its rank–abundance curve. Nevertheless, we will reserve the term ‘‘exponential’’ for the abundance distribution given by Eq. 3.

Exponential distribution

We also use the exponential probability distribution to describe the abundances of species. The exponential distribution is the continuous analogue to MacArthur's discrete ‘‘broken stick’’ distribution (Longuet-Higgins 1971, May 1975). The exponential distribution has a single parameter, λ, and takes the form

$$\phi(n) = \lambda e^{-\lambda n}. \quad (3)$$

Note that this distribution is in no way related to the truncated hyperbolic distribution, which has been referred to as ‘‘exponential’’ by some authors. Because it has one parameter, the variance and mean of the exponential distribution cannot be varied independently.

Gamma distribution

We also investigate the gamma distribution, which has two parameters. The gamma distribution is very flexible and it matches empirical abundance distributions well, although it is rarely used in the ecological literature:

$$\phi(n) = \frac{\lambda^\beta n^{\beta-1}e^{-\lambda n}}{\Gamma(\beta)}. \quad (4)$$

When the shape parameter β equals one, the gamma distribution simplifies to the exponential distribution.

Logseries distribution

The four preceding abundance distributions are all continuous. Nevertheless, due its widespread use in ecology, we will also include the logseries distribution, which is discrete. The logseries was introduced by Fisher et al. (1943) and, much like the truncated hyperbolic, describes the abundances of species that compete for a single resource. The distribution is often characterized by two parameters, x and α. If we let $\phi(N)$ denote the number of species with exactly N individuals, then the logseries satisfies

$$\phi(N) = \frac{\alpha x^N}{N} \quad (5)$$

where N is a positive integer and the parameter x lies between zero and one. In this formulation, the discrete density function ϕ is not yet normalized; the total number of species present in the assemblage is given by $S = \Sigma\phi(N) = -\alpha \ln(1 - x)$. If the value of S is fixed, then x can be obtained from S, leaving only α as a free parameter.

For our purposes we do not need to know the total number of species present in the ambient large region. Hence, we will use the normalized form of the logseries distribution, which depends upon a single parameter x:

$$\phi(N) = \frac{x^N}{N\ln\left(\dfrac{1}{1-x}\right)}. \quad (6)$$

We refer to Eq. 6 as the logseries distribution.

The natural continuous analogue of the logseries distribution—just as the exponential distribution is the analogue of MacArthur's broken stick distribution (Longuet-Higgins 1971)—is given by the following density function:

$$\phi(n) = \begin{cases} \dfrac{x^n}{n\Gamma(0, -\ln(x))} & \text{for } n \geq 1 \\ 0 & \text{for } n < 1. \end{cases} \tag{7}$$

In this equation, $\Gamma(x,y)$ represents the partial gamma function defined by

$$\Gamma(x, y) = \int_{t=y}^{\infty} t^{x-1} e^{-t}\, dt.$$

We refer to Eq. 7 as the continuous logseries distribution. As we shall see in *Expected species overlap between samples*, the continuous analogue is an excellent approximation to the standard, discrete logseries.

Sampling Individuals from a Landscape

Given a species that occurs in a large region, the number of its individuals that occur in a small sample of the region depends on the total abundance of that species in the larger region, the size of the sample, and any conspecific spatial aggregation of the individuals. We now formulate analytic expressions that describe the probability of a species' occurrence in a small sample, given knowledge of its abundance in a larger ambient region. As is standard in such analyses (May 1975), we assume that the density of individuals of all species combined is uniform in space. This assumption is appropriate for many ecosystems, including closed-canopy forests (e.g., Hubbell 2001).

Random spatial distributions

If all individuals are located randomly in space, then the number of individuals of a species found in a sample follows a binomial distribution, and is well approximated by a Poisson distribution. (Equivalently, if a given number of individuals are sampled from an area by picking the individuals in random locations, then the number of individuals per species will follow a binomial distribution.) Consider a small sample whose area constitutes a proportion $a < 1$ of the large ambient region. According to the Poisson distribution, the probability $\psi(a,n)$ that a species will be encountered in the sample, given that it has abundance n in the larger region, is

$$\psi(a, n) = 1 - \exp(-an). \tag{8}$$

The Poisson distribution is the simplest model of sampling individuals from a landscape. In previous studies of similarity indices, authors have almost always assumed Poisson sampling (e.g., Morisita 1959, Ricklefs and Lau 1980, Wolda 1981).

Aggregated spatial distributions

The Poisson distribution stipulates that the probability that a sampled individual is of a particular species is independent of the probability that another individual in the sample is of that species. However, if samples are taken within a contiguous area, then conspecific aggregation often leads to nonindependence of individuals, elevating the probability of having more or fewer individuals of a particular species, depending on whether a clump is encountered or not.

Clustering of conspecifics within samples can be modelled using the negative binomial distribution in place of the Poisson. Specifically, if a species has abundance n in the large area, then its probability of occurrence in a sample of proportional size a may be represented by

$$\psi(a, n) = 1 - \left(1 + \frac{an}{k}\right)^{-k} \tag{9}$$

where k is a parameter that reflects the degree of overdispersion or clustering (Wright 1991, He and Gaston 2000). Parameter values $k > 0$ imply spatial aggregation. Notice that as $k \to \pm\infty$, the probability of occurrence approaches $1 - \exp(-an)$. In other words, as $k \to \pm\infty$, the spatial distribution approaches the Poisson (random) case. Parameter values $k < 0$ reflect spatial patterns that are more regular than random placement.

Extensive analyses in tropical forests reveal that nearly every species is aggregated, a few are random, and almost none exhibit spatial regularity (He et al. 1997, Condit et al. 2000, Plotkin et al. 2000*b*). We therefore assume that the clumping parameter k always exceeds zero.

Both Eqs. 8 and 9 and indicate that the probability of a species' occurrence in a sample of area zero is zero, as we would naturally expect. In addition, we would expect that a species that occurs in the large ambient region will also occur in the sample with probability one as the size of the sample approaches the size of the large region, $a \to 1$. Unfortunately, neither Eq. 8 nor 9 satisfies this criterion. This difficulty arises because the Poisson and negative binomial both represent sampling *with* replacement. After an individual is sampled, it is replaced back into the pool of individuals in the large region (even though, strictly speaking, it should be removed). Although the Poisson and the negative binomial can be altered to amend this shortcoming (He and Legendre 2002), the differences between sampling with replacement and without replacement are negligible when, as will always occur in practice, the sample size is significantly smaller than the ambient community from which it is drawn.

Expected Species Overlap Between Samples

General framework

Given a large region for which we know (or assume) an underlying abundance distribution of species, we now investigate the expected proportion of species in common between two small samples. The number of species in common divided by the average number of

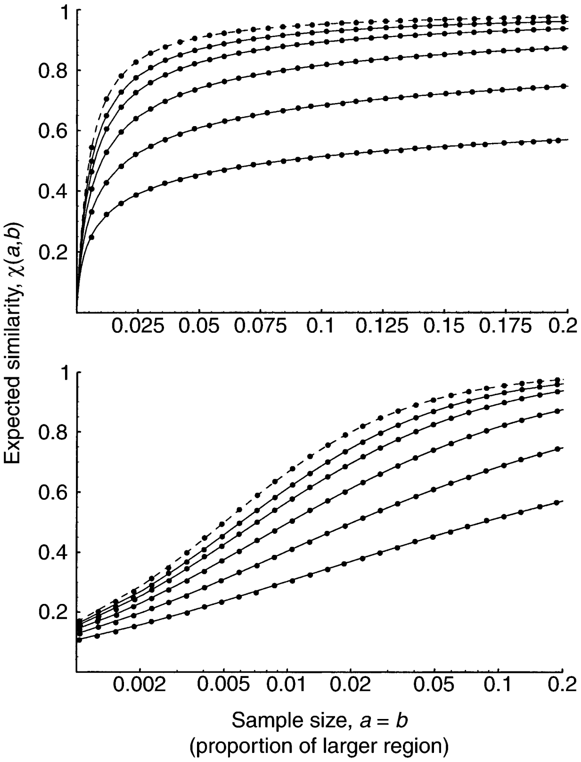

FIG. 2. The expected Sørenson similarity between subplots of increasing area ($a = b$) on linear axes (top) and log-linear axes (bottom). Areas a and b are measured as proportions of the large ambient ecosystem. We compare the analytic expressions in Eqs. 11 and 12 (lines) against the mean similarity of 100 simulated samples (dots). Species abundances were drawn from an exponential distribution with $\lambda = 0.01$. Individuals were sampled via either the Poisson distribution (dashed line) or the negative binomial distribution with a fixed clumping parameter k (solid lines). The similarity is shown for various values of the clumping parameter ($k = 0.2, 0.4, 0.8, 1.6, 3.2$; starting from bottom curve). The mean similarity found by numerical simulations of sampling with replacement (from a large region of 30 000 individuals) matches the analytic predictions. Note that spatial aggregation can significantly alter the similarity curve, but that as $k \to \infty$ the similarity curve approaches the Poisson (random) case.

species in two samples is called the Sørensen index of similarity (Legendre and Legendre 1998). We will also use the term species overlap. The Sørenson index is attractive and popular because of its simplicity and its relation to the species–area curve.

We will consider two sampled regions whose areas constitute proportions a and b, respectively, of a larger ambient region. We assume throughout that the possible species pool in the two samples is the same. (For discussion of cases where the species pool varies see the *Discussion* and the Appendix.) We examine the expected species overlap between the sampled regions under each of four abundance distributions and under both random (Poisson) and aggregated (negative binomial) spatial arrangements of individuals. Under these conditions, we derive an analytic expression for the Sørensen similarity, $\chi(a,b)$, in all cases except the lognormal abundance distribution. Using the same techniques, analytic results are also possible for most other measures of similarity based on species presence/absence.

As before, let $\phi(n)$ denote the distribution of species abundances in the large ambient region. Let $\psi(a, n)$ denote the probability of encountering a species in a sample that covers a proportion a of the larger region, given that the species has abundance n in the larger region. With these definitions, the expected Sørensen similarity between the two samples is given by

$$\chi(a, b) = \frac{\int_0^\infty \phi(n)\psi(a, n)\psi(b, n)\, dn}{\frac{1}{2}\left[\int_0^\infty \phi(n)\psi(a, n)\, dn + \int_0^\infty \phi(n)\psi(b, n)\, dn\right]}. \tag{10}$$

The numerator represents the number of species in common to the two samples, and the denominator represents the average of the number of species present in each sample. Notice that the denominator of Eq. 10 is essentially a species–area relation. Although the denominator contains a species–area relation, note that the dependence of the Sørenson index on sample area is *not* simply equivalent to the scaling of the species diversity with area.

For any abundance distribution ϕ of the species in the large region and for any type of sampling of individuals ψ (e.g., randomly placed individuals or aggregated individuals), Eq. 10 determines the expected number of species in common to two small subregions whose areas constitute proportions a and b of the larger ambient region. Notice that Eq. 10 does not involve the total number of species contained in the ambient region. Counter to our intuition, we can compute the expected similarity between two samples without knowledge of the total diversity contained in the ambient region. This fact justifies our use of a normalized abundance distribution $\phi(n)$ whose total species diversity is left unspecified.

Strictly speaking, Eq. 10 is only approximately correct because it treats the numerator and denominator of Sørensen's index as independent quantities. Nevertheless, when the samples are small compared to the ambient region (as will always be the case in practice) and when the ambient ecosystem is diverse enough so that its abundance distribution is approximately continuous, then the simplifying assumption of independence is justified and Eq. 10 is extremely accurate. (The accuracy of Eq. 10, even for samples as large as 20% of the ambient region, is demonstrated below by comparison with simulations.)

Analytical results

We now integrate Eq. 10 under the assumptions of various abundance and spatial distributions. We start with the situation in which two samples are drawn from the same ambient region whose species abundance distribution is exponential (Eq. 3). Assuming that the spatial distribution of individuals is random (or, equivalently, that individuals are chosen for sampling at random), then integrating Eq. 10 with Eq. 8 yields the expected Sørensen similarity between two samples

$$\chi(a, b) = \frac{2ab(a + b + 2\lambda)}{(a + b + \lambda)(2ab + a\lambda + b\lambda)}. \qquad (11)$$

Next we address expected sample similarity when the individuals are aggregated in space and samples are contiguous. For a species with abundance n and conspecific clustering parameter k, we use $\psi(a, n)$ given by the negative binomial formulation (Eq. 9). We start by assuming that all species are clustered in the same way, i.e., that they all have the same value of the parameter k, and that this parameter is invariant across spatial scales. (Below we address variation in the cluster parameter k.) In this case, Eq. 10 is integrable provided that the two sampled regions have the same area (i.e., $a = b$), yielding

$$\begin{aligned}\chi(a, a) &= \{(k - 1)\Gamma(k)[2H^k\Gamma(1 - k, H) \\ &\qquad - H^{2k}\Gamma(1 - 2k, H) - \exp(-H)]\} \\ &\quad \div \{[\Gamma(k + 1) - \Gamma(k)] \\ &\qquad \times [H^k\Gamma(1 - k, H) - \exp(-H)]\} \qquad (12)\end{aligned}$$

where $H = k\lambda/a$. Eq. 12 matches simulated data exactly, and shows that similarity is lower when individuals are aggregated as opposed to random (Fig. 2).

When the underlying abundance distribution is truncated hyperbolic (Eq. 2) and the spatial distribution is Poisson (Eq. 8), then the expected Sørensen similarity between two subplots is given by

$$\begin{aligned}\chi(a, b) = &\Bigg\{\ln\left(\frac{M}{m}\right) - \Gamma(am) + \Gamma(aM) - \Gamma(bm) \\ &\quad + \Gamma(bM) + \Gamma[(a + b)m] - \Gamma[(a + b)M]\Bigg\} \\ &\div \left[\ln\left(\frac{M}{m}\right) - \frac{1}{2}\Gamma(am) + \frac{1}{2}\Gamma(aM) - \frac{1}{2}\Gamma(bm) \right. \\ &\quad \left. + \frac{1}{2}\Gamma(bM)\right]. \qquad (13)\end{aligned}$$

Under negative binomial sampling, we find a closed-form solution for equal-sized samples: $\chi(a, a) = I/J$ where

$$\begin{aligned}I = k \ln\left(\frac{M}{m}\right) &- 2\left(\frac{k}{am}\right)^k F\left(k, k; k + 1; \frac{k}{am}\right) \\ &+ 2\left(\frac{k}{aM}\right)^k F\left(k, k; k + 1; \frac{k}{aM}\right) \\ &+ \frac{1}{2}\left(\frac{k}{am}\right)^{2k} F\left(2k, 2k; 2k + 1; \frac{k}{am}\right) \\ &- \frac{1}{2}\left(\frac{k}{aM}\right)^{2k} F\left(2k, 2k; 2k + 1; \frac{k}{aM}\right) \\ J = k \ln\left(\frac{M}{m}\right) &- \left(\frac{k}{am}\right)^k F\left(k, k; k + 1; \frac{k}{am}\right) \\ &+ \left(\frac{k}{aM}\right)^k F\left(k, k; k + 1; \frac{k}{aM}\right). \qquad (14)\end{aligned}$$

In these equations, F denotes the generalized hypergeometric function (Gradshtein and Rhyzhik 2000).

For gamma-distributed abundances (Eq. 4) and Poisson sampling, the expected Sørensen similarity is given by

$$\begin{aligned}\chi&(a, b) \\ &= \frac{2[(a + \lambda)^{-\beta} + (b + \lambda)^{-\beta} - (a + b + \lambda)^{-\beta} - \lambda^{-\beta}]}{(a + \lambda)^{-\beta} + (b + \lambda)^{-\beta} - 2\lambda^{-\beta}}. \\ &\qquad (15)\end{aligned}$$

In the case of gamma-distributed abundances and conspecific spatial aggregation (Eq. 9), the expected similarity between two subplots of proportional area a equals $\chi(a, a) = X/Y$ where

$$\begin{aligned}X = 1 &+ \{[\Gamma(k)\Gamma(2k - \beta)F_1(\beta, 1 + \beta - 2k, H) \\ &\qquad - 2\Gamma(2k)\Gamma(k - \beta)F_1(\beta, 1 + \beta - k, H)] \\ &\qquad \div [\Gamma(k)\Gamma(2k)H^{-\beta}]\} \\ &+ \{[H^k\Gamma(\beta - 2k)F_1(2k, 1 - \beta + 2k, H) \\ &\qquad - 2\Gamma(\beta - k)F_1(k, 1 - \beta + k, H)] \\ &\qquad \div [\Gamma(\beta)H^{-k}]\} \\ Y = 1 &- \frac{H^\beta\Gamma(k - \beta)F_1(\beta, 1 + \beta - k, H)}{\Gamma(k)} \\ &- \frac{H^k\Gamma(\beta - k)F_1(k, 1 - \beta + k, H)}{\Gamma(\beta)} \qquad (16)\end{aligned}$$

where $H = k\lambda/a$ and F_1 denotes the Kummer confluent hypergeometric function. Even through the analytic expressions above are complicated, they match simulations exactly (not shown) and they also accurately match empirical data (see *Comparison with Empirical Data*). As usual, aggregated spatial distributions show distinctly lower similarity than random (Poisson) spatial distributions.

Unlike all of the other abundance distributions we consider, the standard logseries distribution is discrete. In this case, the equivalent discrete version of Eq. 10 takes the form

$$\chi(a, b) = \frac{\sum_{N=1}^{\infty} \phi(N)\psi(a, N)\psi(b, N)}{\frac{1}{2}\left[\sum_{N=1}^{\infty} \phi(N)\psi(a, N) + \sum_{n=1}^{\infty} \phi(N)\psi(b, N)\right]}. \qquad (17)$$

In general, it is much more difficult to reduce Eq. 17 into closed form than it is integrate the quantities in Eq. 10. Nevertheless, for the logseries distribution (Eq. 5 or 6) when individuals are distributed randomly in space (Eq. 8), then we have the following closed form solution for the expected species overlap:

$$\chi(a, b) = \{2[\ln(1 - xe^{-a}) + \ln(1 - xe^{-b}) - \ln(1 - x) - \ln(1 - xe^{-a-b})]\} \div [\ln(1 - xe^{-a}) + \ln(1 - xe^{-b}) - 2\ln(1 - x)]. \qquad (18)$$

If we use the continuous analogue of the logseries (Eq. 7), then under Poisson sampling we obtain

$$\chi(a, b) = \{2[\Gamma(0, -\ln(x)) + \Gamma(0, a + b - \ln(x)) - \Gamma(0, a - \ln(x)) - \Gamma(0, b - \ln(x))]\} \div [2\Gamma(0, -\ln(x)) - \Gamma(0, a - \ln(x)) - \Gamma(0, b - \ln(x))]. \qquad (19)$$

In Fig. 3, below, we see that the continuous solution (Eq. 19) agrees almost perfectly with the exact discrete solution (Eq. 18), and that both solutions agree with numerical simulations. This result validates our use of continuous abundance distributions instead of their discrete analogues (e.g., the exponential distribution instead of the broken stick distribution, the truncated hyperbolic instead of the geometric, etc.).

Comparison with previous literature

Previous studies have investigated species overlap between samples. Wolda (1981) analyzed the situation in which abundances are distributed according to the logseries distribution, and individuals are sampled randomly. (Wolda also investigated many other similarity indices.) By fitting numerical simulations, Wolda obtained an approximate equation relating expected species overlap with sample sizes. Wolda gives approximations for four different parameter values of the logseries distribution. In Fig. 3, we compare Wolda's approximate equations with our exact results and with simulations (which match the exact results).

As seen in Fig. 3, Wolda's approximations are fairly accurate within the range of sample sizes and for the particular values of the logseries abundance distribution that he considered. Unfortunately, his numerical results cannot be generalized to other situations. Eqs. 11 through 19, on the other hand, provide exact similarity predictions for a large range of abundance distributions, for arbitrary abundance-distribution parameters, and for both random and aggregated spatial distributions.

Variation in aggregation across species and across spatial scales

So far, we have treated spatial aggregation via a negative binomial sampling function whose clumping parameter, k, does not change with spatial scale nor depend upon species. If the clumping parameter k depends upon area, as it will in many ecosystems, then we can simply replace the parameter k in Eqs. 11–19 with the dependency on area, $k(a)$ and $k(b)$. Thus we can easily accommodate for changes in the intensity of aggregation with spatial scale.

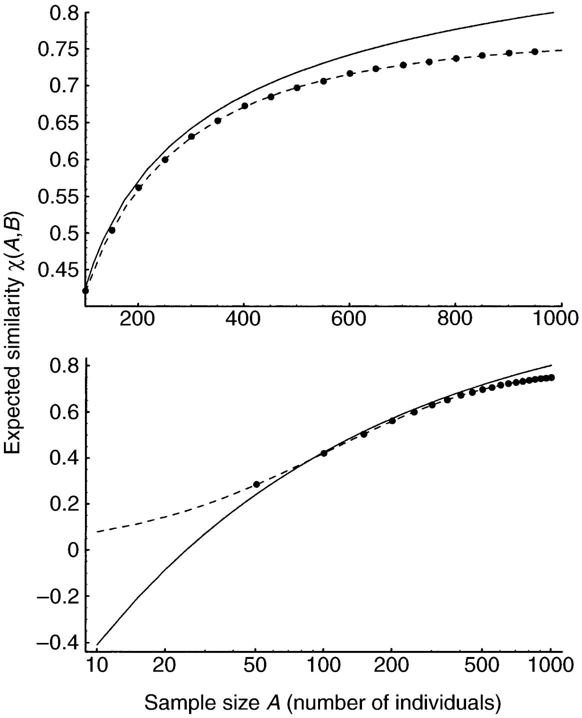

FIG. 3. The expected Sørensen similarity between subplots of increasing sizes on linear axes (top) and log-linear axes (bottom). The dots show the simulation results, the dashed line shows our analytic prediction, and the solid line shows the approximation of Wolda (1981). Samples were drawn randomly from a community that matches one of the exact cases simulated by Wolda (1981): a community of 100 000 individuals in which the abundance distribution of $S = 580$ species is logseries with $\alpha = 81.54$. Wolda (1981) states that in this case the Sørensen similarity should vary as $\chi(A, B) = 1.137 - 3.375A^{-0.347} - 281 \times 10^{-7} B$, where A is the number of individuals in the smaller sample and B is the number in the larger sample. We fix the size of the larger sample, $B = 1000$, and plot similarity as a function of the smaller sample's size, A. Numerical results (dots) represent the mean of 100 simulated samples. Our analytic formulae (Eq. 18, based up on the discrete logseries, and Eq. 19, based upon the continuous logseries) are indistinguishable from each other, and they match the numerical simulations exactly.

Studies have found that within a given species, k often varies with the mean number of individuals sampled. Empirical estimation of k at multiple spatial scales from data on Barro Colorado Island shows that k increases approximately linearly with the area sampled. In this case, there is less of a difference between the negative binomial and the Poisson at large areas, and more of a difference at small areas.

Species may also differ from one another in their tendency to aggregate. Extensive analyses of tropical forests have demonstrated that species exhibit a range of clumping intensities (Condit et al. 2000, Plotkin et al. 2000*b*). Moreover, the range of clumping parameters cannot be collapsed to a single mean value without losing information about the accumulation of species diversity with area (Plotkin et al. 2000*a*, Plotkin and Levin 2001). Tropical forest analyses do not reveal any significant correlation between a species' tendency to aggregate and its abundance (Plotkin et al. 2000*b*). This *does not* in itself imply that a species' clumping parameter is uncorrelated with its abundance, because of possible sampling biases in estimating k. Nevertheless, it is a relatively safe assumption that k and abundance are only weakly correlated.

We can incorporate interspecific variation in clumping into our analysis of similarity indices. If the distribution of clumping parameters is denoted by $\gamma(k)$, independent of abundance, then the expected similarity between subplots a and b is given by simply

$$\chi(a, b) = \int_{k=0}^{\infty} \gamma(k)\chi(a, b, k)\, dk. \tag{20}$$

For most distributions of k values, Eq. 20 is not easily integrated into a closed form solution. Nevertheless, this expression can certainly be evaluated numerically for any given distribution of clumping parameters. As we will see in the next section, however, we can often safely ignore interspecific variation in k and still predict intersite similarity with accuracy.

Comparison with Empirical Data

The methodologies developed above are intended for applications to landscape-scale patterns of species turnover. Nevertheless, we can evaluate the accuracy of our analytic results by using data from the 50-ha tropical forest census on Barro Colorado Island, Panama (Hubbell et al. 1995). In this case, we consider the BCI census as the "large ambient region," and we sample multiple subplots within the region, computing their observed species similarities. Within the 50-ha plot on BCI every free-standing woody stem >1 cm in diameter has been spatially mapped to 1 m accuracy and identified to species. The 50-ha plot contains 229 070 such individuals comprising 300 species (1995 census year).

In order to apply the techniques developed here to BCI, we must first choose a probability density function to model the abundances of species within the BCI census. We find that the abundances are best modelled by using a gamma distribution. In our case, we require that the mean abundance of the best-fit distribution agree with the empirically observed mean abundance (763.57 individuals), so that the total number of individuals represented by 300 species agrees with the observed total number of individuals. Therefore, we require that $\beta/\lambda = 763.57$. Among this one-parameter subfamily of gamma distributions, we choose the parameters ($\beta = 0.2452$, $\lambda = 0.0003211$) that minimize the sum of the squared differences from the observed abundances, when binned (plotted) on a logarithmic (base two) scale (Fig. 4).

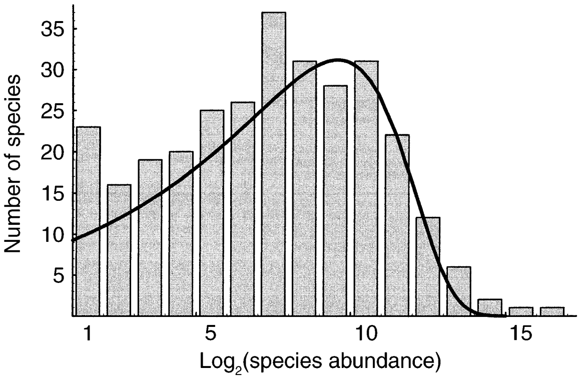

FIG. 4. A histogram of species abundances in the 50-ha plot on Barro Colorado Island, with the best-fit gamma distribution ($\beta = 0.245$, $\lambda = 0.000321$) . The first histogram bar represents the number of species with n individuals, $1 \le n < 2$. The second bar represents those species with $2 \le n < 4$, etc.

Given the best-fit gamma abundance distribution, Eq. 16 predicts the expected similarity between two randomly chosen, equal-sized subplots, under the assumption of conspecific spatial clustering. In order to apply these equations, we must first measure the negative binomial parameter, k, at various quadrat areas (Krebs 1999). Several empirical studies have investigated the relationship between clustering k and quadrat area. In our data, the scaling of k is, in general, best described by a power law of the form

$$k(A) = cA^z + d \tag{21}$$

where A denotes quadrat area in square meters and c, d, and z are constants.

The equations derived above (Eqs. 12, 14, and 16) assume that every species is described by the same clumping parameter $k(A)$, which may vary with area A. In reality, however, there is considerable interspecific variation in clumping parameters. Using the observed abundances found in multiple quadrat draws from BCI, we have estimated the best-fit parameter $k(A)$ for each species and for each quadrat area A via a maximum likelihood procedure (Krebs 1999). Alternatively, we can use the same maximum likelihood method to determine a best fit parameter $k(A)$ that holds across all species simultaneously. Fig. 5 shows the best-fit cluster parameters for two representative species at BCI, as well as the best-fit cluster parameter for all species overall. The scaling of the overall cluster parameter $k(A)$ is accurately described by Eq. 21 with $d = 0.8604$, $c = 0.002923$, and $z = 0.5450$ ($r^2 > 0.99$). The area is measured in square meters.

Fig. 6 shows the observed subplot Sørensen similarities at BCI as well as the predicted similarity using a gamma abundance distribution and a power-law relationship between clustering and area. We sample square subplots of equal areas $a = b$ constituting, at most, 12.5% of the entire 50-ha region.

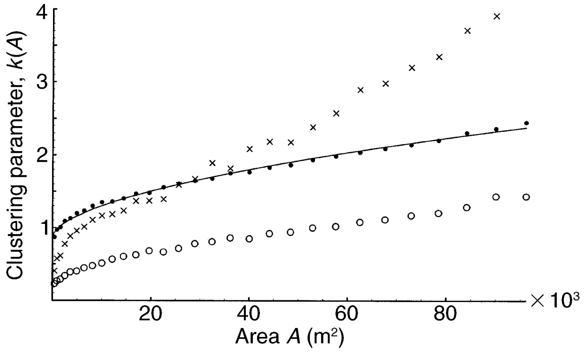

FIG. 5. The observed relationship between spatial scale and the clustering parameter k at BCI. All BCI species are more clustered (lower k) at small scales than at large areas. There is, however, considerable interspecific variation in the scaling of k with area. For example, the x's in the figure show k values of *Calophyllum longifolium* ($n = 1000$), whereas the open circles show *Xylopia macrantha* ($n = 1133$). The dots show the clumping parameter at each scale that is the best fit to all species simultaneously; the solid line shows the best-fit power-law model of $k(A) = 0.8604 + 0.002923 \times A^{0.545}$ ($r^2 > 0.99$).

As seen in Fig. 6, the negative binomial model matches the observed similarity curve in the BCI 50-ha plot fairly well. The predicted similarity is far more accurate when we account for aggregation than when we assume a Poisson model (Fig. 6). The Poisson model consistently overestimates similarity, but falls within two standard errors of the data as the subplot area becomes >5% of 50 ha. The Poisson approximation cannot be rejected at larger areas because the k values increase with area (Fig. 5). In all cases, however, the negative binomial model predicts mean similarity more accurately than the Poisson model.

Although it accurately describes the empirical data, even the negative binomial model seems to overestimate similarity slightly. Such errors generally arise because we are testing the theory at a relatively small (50-ha) scale instead of a true landscape scale. There are three distinct potential sources of error between our analytic theory and the observed BCI similarity data. First, we have modelled the species abundances at BCI by using the gamma distribution (Fig. 4). Although the gamma distribution fits the BCI data better than any other well-known distribution, any discrepancies will cause errors in the predicted similarity between samples. Second, we have assumed that every species at BCI has the same clumping parameter $k(A)$, and that this parameters scales as a power-law in area. But in reality we know that there is a range of clustering parameters across species. Third, we have assumed that the spatial distribution of a species can be modelled by the negative binomial sampling distribution. Although the negative binomial is a reasonable and popular model of spatial aggregation, it need not match the observed sampling distribution $\psi(a, n)$ exactly. In particular, as we have discussed before, the negative binomial formulation assumes sampling with replacement, whereas there is no replacement when sampling in reality.

We can test the relative importance of these three sources of error by using numerical simulations. For instance, we can use the true observed abundances to query the extent of error attributable to the gamma distribution assumption. Such simulations (not shown) indicate that all three potential sources of error do, indeed, contribute small amounts to the observed discrepancy between model and data. Despite these minor sources of error, Fig. 6 demonstrates a remarkably good fit.

NUMERICAL RESULTS AND OTHER SIMILARITY INDICES

Exact results such as the ones presented above can be easily derived only for the expected value (the sampling mean) of presence/absence similarity indices, such as Sørenson's index. Analogous results are more difficult to obtain for quantitative similarity indices that incorporate information on species' abundances. Sampling variances are also difficult to calculate analytically. Thus, further exploration of the sampling distributions of similarity indices requires numerical simulations.

Several studies have used simulations to examine a variety of similarity indices and their dependence upon sample size, as well as on species diversity, always under the assumption of random sampling (Morisita 1959, Ricklefs and Lau 1980, Wolda 1981, Smith and Zaret 1982, Mueller and Altenberg 1985, Smith 1985). In this section, we examine the effects of spatial ag-

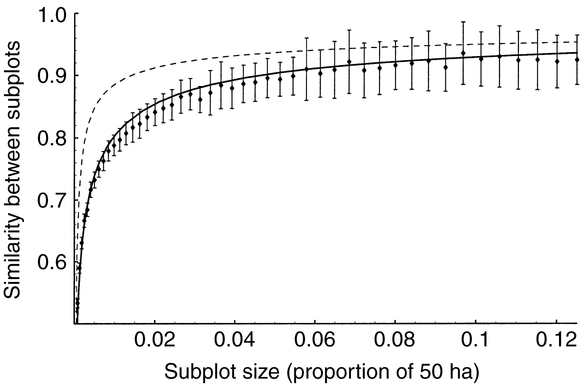

FIG. 6. Observed mean Sørenson similarity between equal-sized subplots of the 50-ha plot on Barro Colorado Island (dots ± 2 SE), compared with the similarity predicted by two analytic models. The dashed line shows similarity predicted according to Poisson sampling (Eq. 15), and the solid line shows negative binomial sampling (Eq. 16). Predictions are based upon the fitted gamma abundance distribution ($\beta = 0.245$, $\lambda = 0.000321$) and the scale-dependent clumping parameter $k(A)$, as in Fig. 5. Observed similarity was calculated, for each subplot size, as the average over many (at least 64) randomly paired, disjoint subplots. The negative binomial model accurately predicts the observed similarity, whereas the Poisson model overestimates similarity.

gregation on sampling means and variances of a presence/absence and a quantitative similarity index.

As before, we will use Sørensen's index as a typical presence/absence index. We use the Renkonen index (Renkonen 1938) to illustrate the general behavior of quantitative similarity indices. Whereas the Sørensen's index is essentially the number of species in common to two samples divided by the average number of species in each sample, the Renkonen index may be thought of as essentially the number of individuals in common divided by the average number of individuals in each sample. The Renkonen index is defined as

$$\zeta(a, b) = \frac{\sum_{i=1}^{S} \min(n_a(i), n_b(i))}{\frac{1}{2}\sum_{i=1}^{S} n_a(i) + \frac{1}{2}\sum_{i=1}^{S} n_b(i)} \qquad (22)$$

where $n_a(i)$ and $n_b(i)$ are the abundances of species i in sample a and sample b, respectively, and S is the total number of species. The value of $\zeta(a, b)$ always lies between 0 and 1; $\zeta(a, b)$ equals 1 only if the two samples a and b contain the same set of species with exactly the same abundances. The Renkonen index has also been referred to as the Steinhaus index (Motyka 1947), the ''Sørensen index with cover,'' the ''percent similarity,'' and the ''coefficient of community.'' The one-complement of the Renkonen index is often called the Bray-Curtis coefficient (Bray and Curtis 1957) or Odum's percentage difference (Odum 1950).

Conspecific aggregation decreases expected similarity and increases sample variance in both Sørensen and Renkonen similarity. This behavior is apparent from Fig. 7, which shows simulation results using parameter values consistent with the BCI 50-ha plot. Note that the Renkonen index is much more sensitive to conspecific aggregation, in both its mean and variance. This sensitivity is a general phenomenon of quantitative similarity indices that are responsive to the number of individuals in each sample, rather than simply their presence or absence. The degree to which quantitative and presence/absence indices differ depends on the diversity of the community: for less skewed abundance distributions and more species-rich communities, quantitative indices behave more like presence/absence indices. Quantitative similarity indices are also more sensitive to variation in clumping among species within the community (not shown).

While most presence/absence similarity indices show the same general response to sample size as the Sørensen, and most quantitative similarity measures behave like the Renkonen, there are a few notable exceptions. The probability that an individual randomly drawn from area a is the same species as an individual randomly drawn from area b is unaffected by sample size; this index, called Morisita's index (Morisita 1959), has been used in recent analytic work (Chave and Leigh, *in press*, Leigh et al., *in press*) and empirical tests thereof (Condit et al. 2002). Under Poisson sampling, the sampling mean of Morisita's index is unaffected by sample size, while the variance decreases steadily with increased sample size. Negative binomial sampling decreases the value of Morisita's index to a level that is dependent on k alone and independent of area; thus, under negative binomial sampling, Morisita's index will increase with area only if k increases with area. The other quantitative similarity indices examined, Horn's information theory index (Horn 1966), Euclidean distance (Ricklefs and Lau 1980), and Morisita's index as adjusted by Horn, show lower similarity under negative binomial sampling than under Poisson, and within each sampling scheme, exhibit decreasing similarity at smaller sample sizes (not shown).

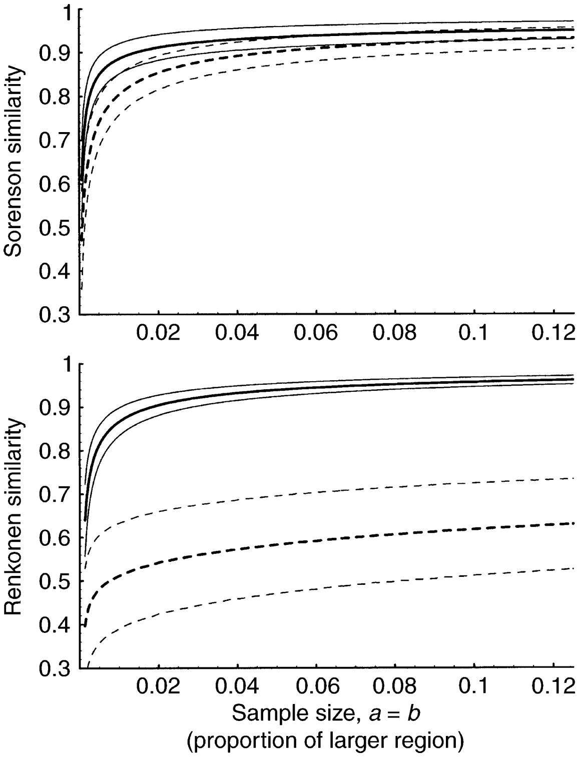

FIG. 7. The sampling means ± 2 SD of the Sørensen (top) and the Renkonen (bottom) similarity indices under Poisson (solid) and negative binomial (dashed) sampling. We simulated 1000 replicate communities, each containing 300 species whose abundances were drawn independently from a gamma distribution with $\lambda = 0.245$ and $\beta = 0.000321$ (as found on BCI). Negative binomial sampling assumed that k was constant for all species and varied with sample size a as in Fig. 5. Replicate similarity indices for a given sample size and type were distributed normally; hence the ±2 SD interval shown encompasses 95% of the variation.

DISCUSSION

Knowledge of the sampling distributions of similarity indices is a prerequisite for testing biogeographic theories of landscape-scale community variation, especially because such theories must be tested from small samples (e.g. Condit et al. 2000, Potts et al. 2002). In this paper, we have developed analytic techniques that synthesize information about the underlying species abundance distribution and conspecific spatial aggregation to yield exact mathematical predictions for species overlap between sampled subregions. Such results can be used to detect and analyze variation in underlying species composition, i.e., species turnover, from small censuses spread across a landscape.

We have demonstrated that local clustering of conspecifics significantly reduces the similarity between sampled subregions. Clustering also increases the variance of sampled similarity indices. Nevertheless, we can incorporate the effect of aggregation into analytic formulae for the expected species overlap between samples. These results are important in light of the spatial aggregation that is nearly ubiquitous in tropical forests and other ecosystems. In fact, without accounting for spatial aggregation, we cannot accurately explain the observed species overlap between subplots of the 50-ha tropical forest census at BCI.

We have limited our analysis to the simple, null case when two samples are drawn from the same underlying region. In other words, we have assumed that a species' abundance in the two sample locations is perfectly correlated, which is a reasonable assumption when the samples are close together. If, instead, the samples are from two very different, distant regions, and species abundances in one region are not correlated with their abundances in the other, then the species overlap between the samples is a function of the joint abundance distribution in the two regions; the integrals over abundances in Eq. 10 would be replaced by double integrals over the abundances in the two regions (see the Appendix). Solutions for two special cases of joint distributions are given in the Appendix. The appropriate form of such joint abundance distributions for intermediate cases in which the distributions in the two regions are partially correlated has not yet been addressed and is an important topic for further research.

Our results, especially those regarding spatial aggregation, have implications for the sampling design of beta diversity studies. If the degree of spatial contagion at the scale of measurement is known, then it can be incorporated into expected similarity as we have shown. Alternatively, sampling efforts can be designed to minimize the influence of local clustering by either (1) arranging long narrow plots (cf. Gentry 1988), or (2) spreading the same total sampled area over a greater region. The disadvantage of such irregular sampling schemes is that they compromise our ability to characterize any remaining effects of aggregation. Furthermore, it would be difficult to compare similarity measures calculated between samples taken in different ways; sampling should above all be done consistently. When the objective is to assess similarity between different regions, the fail-safe way to account for contagion and sample size is to include replicate plots in each habitat or region being compared. In this way, similarity between plots in the same region can be contrasted with similarity of plots from different regions.

We have primarily focused on the Sørenson index of similarity. This simple index measures the proportion of species that overlap between two samples. Although we have seen that the Sørenson index is not as sensitive as other, quantitative indices, Sørenson has several distinct advantages. Foremost, the Sørenson index only requires knowledge of species presence/absence in a census, which is often the extent of information collected in a landscape survey (Gaston 1994). Moreover, the species overlap between samples is closely related to the species–area curve in the underlying region. Novel theories that unify species overlap and species–area curves (Harte et al. 1999) are being developed.

Despite the vast literature on biogeography, there are relatively few unified theories of community assembly, based on biological mechanisms such as speciation and dispersal, that predict the change in species composition across a landscape (Hubbell 2001). The most recent contributions provide theoretical predictions for one similarity index (Chave and Leigh, *in press*, Leigh et al., *in press*), but predictions for species overlap require further research. Such biogeographic theories must eventually be tested against empirical data, bearing in mind the effects of sample size on measured statistics. Similarly, the design and placement of refugia for the preservation of landscape-scale biodiversity must be guided by a sound understanding of the sampling distributions for species turnover.

Acknowledgments

Foremost, we thank Rick Condit for posing the problem of Sorenson similarity to us, and for his conceptual input. We also thank Fangliang He and Jonathan Dushoff for stimulating discussions, and John Harte, Jess Green, and an anonymous reviewer for helpful comments. We are indebted to the principal investigators behind the 50-ha Forest Dynamics Plot on Barro Colorado Island: Robin Foster, Steve Hubbell and Rick Condit. The Forest Dynamics Plot of Barro Colorado Island has been made possible through the generous support of the U.S. National Science Foundation, The John D. and Catherine T. MacArthur Foundation, and the Smithsonian Tropical Research Institute, and through the collaborative efforts of many scientists over the past two decades. The BCI Forest Dynamics Plot is part of the Smithsonian's Center for Tropical Forest Science, a global network of demographic tree plots. Finally, we gratefully acknowledge support to J. B. Plotkin in the form of a National Science Foundation graduate fellowship, and from the Teresa and H. John Heinz III Foundation and the Burroughs Wellcome Fund, and to H. C. Muller-Landau from the Smithsonian Tropical Research Institue, the Mellon Foundation, and the Princeton Environmental Insititute.

Literature cited

Bray, R. J., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs **27**:325–349.

Chave, J., and E. G. Leigh, Jr. *In press.* A spatially-explicit neutral model of beta-diversity in tropical forests. Theoretical Population Biology.

Cody, M. L. 1986. Diversity, rarity, and conservation in Mediterranean-climate regions. Pages 123–152 *in* M. E. Soulé, editor. Conservation Biology. Sinauer Associates, Sunderland, Massachusetts, USA.

Condit, R., et al. 2000. Spatial patterns in the distribution of tropical tree species. Science **288**:1414–1418.

Condit, R., et al. 2002. Beta diversity in tropical forest trees. Science **295**:666–669.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relations 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. 1994. Rarity. Chapman and Hall, London, UK.

Gradshtein, I. S., and I. M. Ryzhik. 2000. Table of integrals, series, and products. Sixth edition. Academic Press, San Diego, California, USA.

Harrison, S. 1997. How natural habitat patchiness affects the distribution of diversity in Californian serpentine chaparral. Ecology **78**:1898–1906.

Harrison, S., S. J. Ross, and J. H. Lawton. 1992. Beta-diversity on geographic gradients in Britain. Journal of Animal Ecology **61**:151–158.

Harte, J., K. Taylor, A. P. Kinzig, and M. Fisher. 1999. Estimating species–area relationships from plot to landscape scale using species spatial-turnover data. Oikos **86**:45–54.

He, F., and K. J. Gaston. 2000. Estimating species abundance from occurrence. American Naturalist **156**:553–559.

He, F., and P. Legendre. 2002. Species diversity patterns derived from species–area models. Ecology **83**:1185–1198.

He, F., P. Legendre, and J. LaFrankie. 1997. Distribution patterns of tree species in a Malaysian tropical rain forest. Journal of Vegetation Science **8**:105–114.

Horn, H. 1966. Measurement of ''overlap'' in comparative ecological studies. American Naturalist **100**:419–424.

Hubbell, S. P. 1995. Towards a theory of biodiversity and biogeography on continuous landscapes. Pages 171–199 *in* G. Carmichael, G. Folk, and J. Schnoor, editors. Preparing for global change: a Midwestern perspective. SPB Academic, Amsterdam, The Netherlands.

Hubbell, S. P. 1997. A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. Coral Reefs **16**:S9–S21.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.

Hubbell, S. P., R. Foster, R. Condit, S. Lao, and R. Perez. 1995. Demographic tree data from the 50-ha Barro Colorado Island Forest Dynamics Plot, 1982–1995. Center for Tropical Forest Science, Forest Dynamics Plot Data Series. Panama City, Republic of Panama. CD-ROM. Smithsonian Institution, Washington, D.C., USA.

Krebs, C. J. 1999. Ecological methodology. Second edition. Benjamin/Cummings, Menlo Park, California, USA.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam, The Netherlands.

Leigh, E. G., Jr., R. Condit, and S. Loo de Lao. In press. Null models, sample plots, and tropical forest diversity. *In* E. C. Losos, R. Condit, J. V. LaFrankie, and E. G. Leigh, editors. Tropical forest diversity and dynamism: findings from a network of large-scale tropical forest plots. University of Chicago Press, Chicago, Illinois, USA.

Longuet-Higgins, M. S. 1971. On the Shannon-Weaver index of diversity, in relation to the distribution of species in bird censuses. Theoretical Population Biology **2**:271–289.

May, R. M. 1975. Patterns of species abundance and diversity. Pages 81–120 *in* M. L. Cody and J. M. Diamond, editors. Ecology and evolution of communities. Belknap Press of Harvard University Press, Cambridge, Massachusetts, USA.

Morisita, M. 1959. Measuring of interspecific association and similarity between communities. Memoirs of the Faculty of Science of Kyushu University, Series E. Biology **3**:65–80.

Motyka, J. 1947. O zadaniach i metodach badan geobotanicznych. Sur les buts et les méthodes des recherches géobotaniques. Annales Universitatis Mariae Curie-Sklodowska (Lublin, Polonia), Sectio C, Supplementum **I**.

Mueller, L. D., and L. Altenberg. 1985. Statistical-inference on measures of niche overlap. Ecology **66**:1204–1210.

Odum, E. P. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. Ecology **31**:587–605.

Plotkin, J. B., and S. A. Levin. 2001. The spatial distribution and abundances of species: lessons from tropical forests. Comments on Theoretical Biology **6**:251–278.

Plotkin, J. B., M. D. Potts, N. Leslie, N. Manokaran, J. LaFrankie, and P. S. Ashton. 2000*a*. Species–area curves, spatial aggregation, and habitat specialization in tropical forests. Journal of Theoretical Biology **207**:81–99.

Plotkin, J. B., M. D. Potts, D. W. Yu, S. Bunyavejchewin, R. Condit, R. Foster, S. Hubbell, J. LaFrankie, N. Manokaran, L. H. Seng, R. Sukumar, M. A. Nowak, and P. S. Ashton. 2000*b*. Predicting species diversity in tropical forests. Proceedings of the National Academy of Sciences (USA) **97**: 10850–10854.

Potts, M. D., P. A. Ashton, L. Kaufman, and J. B. Plotkin. 2002. The effect of habitat and distance on tropical tree species: a floristic comparison of 105 plots in Northwest Borneo. Ecology **83**, In press.

Preston, F. W. 1962. The canonical distribution of commonness and rarity: Part II. Ecology **43**:410–432.

Renkonen, O. 1938. Statisch-ökologische Untersuchungen über die terrestiche Kaferwelt der finnischen Bruchmoore. Annales Zoologici Societatis Zoologicae–Botanicae Fennicae 'Vanamo' **6**:1–123.

Ricklefs, R. E., and M. Lau. 1980. Bias and dispersion of overlap indices: results of some Monte Carlo simulations. Ecology **61**:1019–1024.

Routledge, R. D. 1977. On Whittaker's components of diversity. Ecology **58**:1120–1127.

Smith, E. P. 1985. Statistical comparison of weighted overlap measures. Transactions of the American Fisheries Society **114**:250–257.

Smith, E. P., and T. M. Zaret. 1982. Bias in estimating niche overlap. Ecology **63**:1248–1253.

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

Wolda, H. 1981. Similarity indices, sample size and diversity. Oecologia **50**:296–302.

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

APPENDIX

Expected Species Overlap when Abundances are not Perfectly Correlated

In general, the expected species overlap between two samples is a function of the joint distribution, $\phi(m, n)$ of species abundances in the regions from which they are sampled, where m is the abundance in the region from which proportion a is sampled, and n is the abundance in the region from which proportion b is sampled:

$$\chi(a, b) = \left[\int_0^\infty \int_0^\infty \phi(m, n)\psi(a, m)\psi(b, n)\, dm\, dn\right] \div \left\{\frac{1}{2}\left[\int_0^\infty \int_0^\infty \phi(m, n)\psi(a, m)\, dm\, dn + \int_0^\infty \int_0^\infty \phi(m, n)\psi(b, n)\, dn\, dm\right]\right\}. \qquad \text{(A.1)}$$

When the samples are taken from the same region such that $m = n$ for all species, then this equation reduces to Eq. 10. It follows also that if samples are taken from regions that share a proportion p of their species, with those species having the same abundances in both regions and other species not shared, then the expected species overlap between the samples is simply $p \times \chi(a, b)$, where χ is calculated as in Eqs. 11–19.

If the species pool is identical at the two samples, and the abundance distributions have the same exact shape in both regions, but the abundances of individual species in the two regions are completely uncorrelated, then the expected similarity is given by

$$\chi(a, b) = \left[\int_0^\infty \phi(n)\psi(a, n)\, dn \int_0^\infty \phi(m)\psi(b, m)\, dm\right] \div \left\{\frac{1}{2}\left[\int_0^\infty \phi(n)\psi(a, n)\, dn + \int_0^\infty \phi(m)\psi(b, m)\, dm\right]\right\}. \qquad \text{(A.2)}$$

This situation is somewhat unrealistic, since it suggests that exactly the same species are present in the two communities, but that a species's abundance in one location is completely uncorrelated with its abundance in the other. Nevertheless, an examination of the extreme case of completely uncorrelated abundances may be useful in indicating the limitations of similarity indices at detecting amounts of correlation.

When the two samples each comprise the same proportional area of their respective ambient regions, $a = b$, then Eq. A.2 reduces to $\chi(a, a) = \int_0^\infty \phi(n)\, \psi(a, n) dn$. In other words, in this case the expected species overlap between the two samples is simply the expected proportion of all the species in one ambient region that occur in a sample of size a. So, in this special case, the similarity between two subregions scales linearly with the species–area curve, regardless of the abundance or spatial distribution.

Below we present closed-form integrations of Eq. A.2 for a range of abundance and spatial distributions, parallel to our results in Eqs. 11–19.

Truncated hyperbolic abundance distribution

For Poisson sampling and uncorrelated abundance distributions, Eq. A.2 reduces to the following form:

$$\chi(a, b) = \left\{2\left[\ln\left(\frac{M}{m}\right) + \Gamma(0, am) - \Gamma(0, aM)\right] \times \left[\ln\left(\frac{M}{m}\right) + \Gamma(0, bm) - \Gamma(0, bM)\right]\right\} \div \left\{\ln\left(\frac{M}{m}\right)\left[2 \ln\left(\frac{M}{m}\right) + \Gamma(0, am) - \Gamma(0, aM) + \Gamma(0, bm) - \Gamma(0, bM)\right]\right\}. \qquad \text{(A.3)}$$

For negative binomial sampling and uncorrelated abundances with constant k:

$$\chi(a, b) = \left[\ln\left(\frac{M}{m}\right) - k\left(\frac{k}{am}\right)^k F\left(k, k; k + 1; \frac{k}{am}\right) + k\left(\frac{k}{aM}\right)^k F\left(k, k; k + 1; \frac{k}{aM}\right)\right] \div \left[\ln\left(\frac{M}{m}\right)\right]. \qquad \text{(A.4)}$$

Exponential abundance distribution

When the abundances are uncorrelated and spatial distribution is random, we have

$$\chi(a, b) = \frac{2ab}{2ab + b\lambda + a\lambda}. \qquad \text{(A.5)}$$

Under negative binomial sampling, we find

$$\chi(a, b) = 2 \times \left(\frac{a}{a - e^{k\lambda/a} k\lambda E\left(k, \frac{k\lambda}{a}\right)} + \frac{b}{b - e^{k\lambda/b} k\lambda E\left(k, \frac{k\lambda}{b}\right)}\right)^{-1} \qquad \text{(A.6)}$$

where $E(x, y) = \int_1^\infty e^{-yt}\, t^{-x}\, dt$ is the exponential integral.

Gamma abundance distribution

Under Poisson sampling, with uncorrelated abundances we find that

$$\chi(a, b) = \frac{-2(\lambda^{-\beta} - (a + \lambda)^{-\beta} - (b + \lambda)^{-\beta} + (a + b + \lambda)^{-\beta})}{(-2\lambda^{-\beta} + (a + \lambda)^{-\beta} + (b + \lambda)^{-\beta})}. \qquad \text{(A.7)}$$

Under negative binomial sampling, we find that $\chi(a,a) = N/D$ where

$$N = \{\Gamma(\beta) + a^{-2k}(k\lambda)^k \times [(k\lambda)^k F(2k, 1 - \beta + 2k, H)\Gamma(\beta - 2k) - 2a^k F(k, 1 - \beta + k, H)\Gamma(\beta - k)]\}/[\Gamma(\beta)] - \frac{2a^{-\beta}(k\lambda)^\beta F(\beta, 1 + \beta - k, H)\Gamma(k - \beta)}{\Gamma(k)} + \frac{a^{-\beta}(k\lambda)^\beta F(\beta, 1 + \beta - 2k, H)\Gamma(2k - \beta)}{\Gamma(2k)}$$

$$D = 1 - \frac{\Gamma(\beta)^{-1} a^{-k}(k\lambda)^k F(k, 1 - \beta + k, H)\Gamma(\beta - k)}{\Gamma(\beta)} - \frac{a^{-\beta}(k\lambda)^\beta F(\beta, 1 + \beta - k, H)\Gamma(k - \beta)}{\Gamma(k)}$$

where $H = k\lambda/a$.

Logseries abundance distribution

Under Poisson sampling, with uncorrelated abundances we find that

$$\chi(a, b) = \{2[\ln(1 - xe^{-a}) - \ln(1 - x)] \times [\ln(1 - x) - \ln(1 - xe^{-b})]\} \div \{\ln(1 - x)[\ln(1 - xe^{-a}) + \ln(1 - xe^{-b}) - 2 \ln(1 - x)]\}. \qquad \text{(A.8)}$$