Biometrics 62, 361–371
June 2006

DOI: 10.1111/j.1541-0420.2005.00489.x

Abundance-Based Similarity Indices and Their Estimation When There Are Unseen Species in Samples

Anne Chao,[1,*] Robin L. Chazdon,[2] Robert K. Colwell,[2] and Tsung-Jen Shen[3]

[1]Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan
[2]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269-3043, U.S.A.
[3]Department of Applied Mathematics, National Chung Hsing University, Tai-Chung 402, Taiwan
[*]*email:* chao@stat.nthu.edu.tw

Summary. A wide variety of similarity indices for comparing two assemblages based on species incidence (i.e., presence/absence) data have been proposed in the literature. These indices are generally based on three simple incidence counts: the number of species shared by two assemblages and the number of species unique to each of them. We provide a new probabilistic derivation for any incidence-based index that is symmetric (i.e., the index is not affected by the identity ordering of the two assemblages) and homogeneous (i.e., the index is unchanged if all counts are multiplied by a constant). The probabilistic approach is further extended to formulate abundance-based indices. Thus any symmetric and homogeneous incidence index can be easily modified to an abundance-type version. Applying the Laplace approximation formulas, we propose estimators that adjust for the effect of unseen shared species on our abundance-based indices. Simulation results show that the adjusted estimators significantly reduce the biases of the corresponding unadjusted ones when a substantial fraction of species is missing from samples. Data on successional vegetation in six tropical forests are used for illustration. Advantages and disadvantages of some commonly applied indices are briefly discussed.

Key words: Beta diversity; Biodiversity; Forest succession; Species overlap.

1. Introduction

In comparing species composition and biodiversity of two or more assemblages in taxonomic and ecological research, similarity (or overlap) or dissimilarity (complementarity, turnover, beta diversity, or distance) indices provide quantitative bases of assessment (Magurran, 2004). In this article, we focus on similarity indices for comparing two assemblages. All derivations and arguments can be readily applied to dissimilarity indices as well.

A large number of similarity indices based on presence/absence (incidence) data have been proposed in the literature. The two classic and the most widely used ones are the Jaccard and Sørensen indices (Ludwig and Reynolds, 1988; Magurran, 2004). Hubalek (1982) and Gower (1985) provided comprehensive reviews. Lennon et al. (2001) have given new interpretation and application to an old index originally proposed by Simpson (1943). For simplicity, this index hereafter is referred to as the Lennon et al. index. The Jaccard, Sørensen, Lennon et al., and many other incidence-type indices for comparing two assemblages are generally functions of three incidence counts: the number of species shared by two assemblages and the number of species unique to each.

Despite their simplicity and wide application in ecological studies, the incidence-based indices do not take species abundance into account, thus abundant and rare species are treated equally. These indices, when estimated from samples, do not perform well (Wolda, 1981, 1983; Colwell and Coddington, 1994; Fisher, 1999; Plotkin and Muller-Landau, 2002). The estimates are generally biased downward and the bias increases when sample sizes are small or species richness is large. A major statistical concern is that, based on incidence data, bias correction and measurements of variances are impossible. (See Section 4.1 for details.) Consequently, the interpretation of any incidence-based index becomes especially difficult for comparing two (or more) diverse assemblages that contain numerous rare species, particularly when limited to analysis of data from small samples. More discussion follows in Sections 4 and 6.

Abundance-type indices of compositional similarity have received relatively less attention by researchers in biodiversity research. A modified version of the Sørensen index was developed by Bray and Curtis (1957), based on abundance data (also known as the Sørensen abundance index; Magurran, 2004). The Bray–Curtis similarity index is widely used to generate distance matrices in vegetation ordination studies (Gotelli and Ellison, 2004). Another widely used abundance-based index is the Morisita-type index (Magurran, 2004, p. 175). The Bray–Curtis and Morisita indices will be further discussed in Section 4. A generalized form of the Morisita index was proposed by Grassle and Smith (1976). Smith, Solow, and Preston (1996) considered a Jaccard-type abundance

index using a modified beta-binomial model. Yue, Clayton, and Lin (2001) subsequently extended it to a general case. Plotkin and Muller-Landau (2002) developed a Sørensen-type similarity index for abundance counts using a parametric approach that relies on a gamma distribution to characterize species abundance structure. To our knowledge, the effect of unseen shared species on any index has not previously been addressed in the literature.

This research was motivated by analyzing plant data to compare tree and seedling (or tree and sapling) size classes in successional rain forest plots in Costa Rica. The abundance data were collected by Chazdon and colleagues as part of a larger study of secondary succession of tropical rain forests following pasture abandonment (Redondo, Vilchez, and Chazdon, 2001; Chazdon, Redondo, and Vilchez, 2005). During early stages of succession, a forest site is usually dominated by fast-growing and shade-intolerant colonizing tree species that are abundant in all size classes (trees, saplings, and seedlings). As the forest canopy is established, these shade-intolerant tree species drop out of the seedling and sapling pool, and some shade-tolerant species begin to colonize. The late colonizing tree species are represented by seedlings and saplings, but have few or no canopy trees present, gradually augmenting tree species richness as the forest matures (Guariguata et al., 1997, Table 4). As secondary forests mature, we would therefore predict that compositional similarity between tree species and seedling or sapling species would initially be high, but would quickly decline to a minimum during intermediate stages of succession and then begin to increase later in succession as shade-tolerant trees reach reproductive maturity and produce seedlings that can establish, grow, and survive. The detailed abundance data and analysis will be discussed in Section 5.

It is clear that abundance counts play important roles in describing forest compositional changes with time. Also, due to sampling limitation, complete inventories are impractical and even impossible, especially for smaller size classes (seedlings and saplings) (Chazdon et al., 1998), and similar limitations apply to many animal studies, especially in the tropics (e.g., Longino, Coddington, and Colwell, 2002). It can be expected that some species that are present in a forest stand will be missing from the samples. In such incomplete surveys, no existing abundance-based index incorporates the effect of unseen species to adjust for this undersampling bias. We were thus motivated to derive abundance-based indices that incorporate the effect of unseen shared species. See Chao et al. (2005) for the ecological counterpart paper.

We begin by reviewing a class of incidence-based similarity indices that are symmetric (i.e., the index is not affected by the identity ordering of the two assemblages) and homogeneous (i.e., the index is unchanged if all counts are multiplied by a constant). We then develop a new probabilistic approach that can be applied to any member of the class. This approach is further extended to formulate abundance-based indices that do not require any assumptions about the statistical distribution of species abundance. Thus any symmetric and homogeneous incidence index can be generalized to an abundance-type version. Using the Laplace approximation formulas, we propose estimators that adjust for the effect of unseen shared species for the abundance-based indices. Simulations from known plant assemblages are used to demonstrate that the adjusted estimators work well for inferring similarity between hyper-diverse assemblages for which a large portion of species might not be seen in samples.

In Section 2.1, a class of incidence-based indices is reviewed. In Section 2.2, we provide a probabilistic approach to interpreting the incidence indices. The approach is extended in Section 3.1 to formulate the corresponding abundance-based version. The estimation of the proposed abundance-based indices when there are unseen species in samples is briefly described in Section 3.2. (Derivation details are given on the *Biometrics* website.) Discussion of the pros and cons of some commonly used indices is provided in Section 4. Application to the rain forest data is presented in Section 5. Simulation results based on generated data from plant assemblages are reported in Section 6 to investigate the performance of our estimators. Concluding remarks and discussion are provided in Section 7.

2. A Probabilistic Approach to Incidence-Based Indices

2.1 *A Class of Incidence-Based Indices*

Most incidence-based indices for comparing two assemblages depend on three incidence counts: the number of species shared by two assemblages and the number of species unique to each of them. In the ecological literature, it has become traditional to refer to these counts as a, b, and c, respectively (Table 1). Assume that there are $S_1 > 0$ species in Assemblage 1 and there are $S_2 > 0$ species in Assemblage 2. Let the number of shared species be S_{12}. As shown in Table 1, the incidence counts a, b, and c correspond to $a = S_{12}, b = S_1 - S_{12}$, and $c = S_2 - S_{12}$.

Based on Gower (1985), Wolda (1981), Hubalek (1982), and Lennon et al. (2001), we select in the second column of Table 2 a class of similarity indices that satisfy

(i) Symmetry: b and c are interchangeable in the index; and

(ii) Homogeneity: if all counts (a, b, c) are multiplied by a constant K, then the index for the counts (Ka, Kb, Kc) is exactly the same as that for the counts (a, b, c). See Janson and Vegelius (1981) for discussion of this property.

All these incidence indices can be reexpressed as functions of S_1, S_2, S_{12} as shown in the third column of Table 2. This type of expression has the advantage of linking incidence and abundance indices, as will be seen in Section 3.

Table 1
Species classification counts used in incidence indices

		Assemblage 2	
		Presence	Absence
Assemblage 1	Presence	$a = S_{12}$	$b = S_1 - S_{12}$
	Absence	$c = S_2 - S_{12}$	–

Table 2
A class of incidence- and abundance-based indices and their relations

Index	Incidence based in terms of a, b, c	Incidence based in terms of S_1, S_2, S_{12}	Abundance based (see Section 3)
Jaccard	$\frac{a}{a+b+c}$	$\frac{S_{12}}{S_1+S_2-S_{12}}$	$\frac{UV}{U+V-UV}$
Sørensen; Dice	$\frac{2a}{(2a+b+c)}$	$\frac{2S_{12}}{S_1+S_2}$	$\frac{2UV}{U+V}$
Ochiai	$\frac{a}{[(a+b)(a+c)]^{1/2}}$	$\frac{S_{12}}{(S_1S_2)^{1/2}}$	$(UV)^{1/2}$
Anderberg	$\frac{a}{a+2(b+c)}$	$\frac{S_{12}}{2S_1+2S_2-3S_{12}}$	$\frac{UV}{2U+2V-3UV}$
Kulczynski	$\frac{a}{b+c}$	$\frac{S_{12}}{S_1+S_2-2S_{12}}$	$\frac{UV}{U+V-2UV}$
Kulczynski; Cody	$\frac{a}{2(a+b)}+\frac{a}{2(a+c)}$	$\frac{1}{2}\left(\frac{S_{12}}{S_1}+\frac{S_{12}}{S_2}\right)$	$\frac{1}{2}(U+V)$
Lennon et al. (2001)	$\frac{a}{a+\min(b,c)}$	$\frac{S_{12}}{S_{12}+\min(S_1-S_{12},S_2-S_{12})}$	$\frac{UV}{UV+\min(U-UV,V-UV)}$

2.2 *A Probabilistic Approach to the Incidence-Based Indices*

To extend the incidence indices to take account of the relative abundance of species, we must first provide a probabilistic derivation for incidence indices. Suppose we randomly select one *species* from Assemblage 1 and one *species* from Assemblage 2 and then classify each member of the pair according to whether it is a shared species or not. The corresponding probabilities are specified in Table 3.

Comparing Tables 1 and 3, we note the following relations: Table 3 is obtained by first interchanging $b(= S_1 - S_{12})$ and $c(= S_2 - S_{12})$ in Table 1 and then multiplying all counts by a constant $K = S_{12}/(S_1S_2)$ provided that $S_{12} > 0$. Then any symmetric and homogeneous index based on the counts (a, b, c) in Table 1 is exactly the same as that based on the counts (A, B, C) in Table 3. This can also be verified by direct calculations. For example, for the Jaccard index, we have

$$\begin{aligned}&\frac{A}{A+B+C}\\ &= \frac{\frac{S_{12}}{S_1}\frac{S_{12}}{S_2}}{\frac{S_{12}}{S_1}\frac{S_{12}}{S_2}+\frac{S_{12}}{S_1}\left(1-\frac{S_{12}}{S_2}\right)+\left(1-\frac{S_{12}}{S_1}\right)\frac{S_{12}}{S_2}}\\ &= \frac{S_{12}}{S_1+S_2-S_{12}} = \frac{a}{a+b+c}.\end{aligned}$$

Similar verification applies to the other indices. This probabilistic approach lays the groundwork for developing abundance-based indices, which in turn allow for the estimation of indices that adjust for the effect of unseen shared species.

Table 3
Probabilistic derivation of the incidence indices

		Select any *species* from Assemblage 2	
		Shared	Nonshared
Select any *species* from Assemblage 1	Shared	$A = \frac{S_{12}}{S_1} \times \frac{S_{12}}{S_2}$	$B = \frac{S_{12}}{S_1} \times (1 - \frac{S_{12}}{S_2})$
	Nonshared	$C = (1 - \frac{S_{12}}{S_1}) \times \frac{S_{12}}{S_2}$	

3. Extension to Abundance-Based Indices and Estimation

3.1 *Abundance-Based Indices*

Let the probabilities of species discovery in Assemblages 1 and 2 be denoted, respectively, by $(p_1, p_2, \ldots, p_{S_1})$ and $(\pi_1, \pi_2, \ldots, \pi_{S_2})$, where $p_i > 0$, $\pi_i > 0$, and $\sum_{i=1}^{S_1} p_i = \sum_{i=1}^{S_2} \pi_i = 1$. We no longer treat all species equally because some species are common and some are rare. Instead, the basic idea for handling abundance counts is that we treat all *individuals* equally. Adapting the approach in Section 2, we randomly select one *individual* from Assemblage 1 and also one *individual* from Assemblage 2. For each selected individual of the pair, note whether it is a shared species or not.

Without loss of generality, we assume the first S_{12} species are shared species, that is, the shared species are indexed by $1, 2, \ldots, S_{12}$. In Assemblage 1, let U denote the total relative abundances associated with the *shared* species, $U = p_1 + p_2 + \cdots + p_{S_{12}}$. Likewise, in Assemblage 2, let V denote the total relative abundances of the *shared* species, $V = \pi_1 + \pi_2 + \cdots + \pi_{S_{12}}$. Then Table 3 for incidence-type indices can be generalized to Table 4 for abundance-type indices.

As we construct the class of incidence-based indices from Tables 1 and 3, now we can formulate the corresponding abundance-based indices from Tables 1 and 4. By replacing a, b, and c in Table 2 with UV, $U(1-V)$, and $V(1-U)$, we obtain abundance-based indices as given in the last column of Table 2. For example, the Jaccard abundance-type index is $UV/(U+V-UV)$ and the Sørensen abundance-type index becomes $2UV/(U+V)$. Similarly for the other indices, their

Table 4
Probabilities for individual-based counts

		Select any *individual* from Assemblage 2	
		Shared	Nonshared
Select any *individual* from Assemblage 1	Shared	$A = U \times V$	$B = U \times (1 - V)$
	Nonshared	$C = (1 - U) \times V$	–

corresponding abundance-based indices are directly formulated in Table 2. Since U and V represent the total abundances of the *shared* species in Assemblages 1 and 2, respectively, it is clear that all abundance indices in Table 2 yield a maximum value of 1 when all species are shared (i.e., no unique species in both assemblages; $U = V = 1$). Also, all indices tend to a minimum value of 0 for disjoint assemblages (i.e., no shared species in both assemblages); e.g., for the Jaccard abundance index, we have $UV/(U + V - UV) = 1/\{(1/V) + (1/U) - 1\}$, which tends to 0 if both U and V tend to 0 (i.e., disjoint assemblages).

From Table 2, we can see that there is a nice link between the proposed abundance index (in terms of U, V, and UV) and the incidence form (in terms of S_1, S_2, S_{12}). That is, our abundance index is directly obtained by replacing S_1, S_2, S_{12}, respectively, by U, V, and UV. In the special case of equal abundance, that is, $p_1 = p_2 = \cdots = p_{S_1} = 1/S_1$ and $\pi_1 = \pi_2 = \cdots = \pi_{S_2} = 1/S_2$, then $U = p_1 + p_2 + \cdots + p_{S_{12}} = S_{12}/S_1$ and $V = \pi_1 + \pi_2 + \cdots + \pi_{S_{12}} = S_{12}/S_2$. It is clear that each cell in Table 4 is a generalization of that in Table 3. Consequently, in the special case of equal abundance, our abundance indices are reduced to their corresponding incidence indices.

3.2 *Estimation When There Are Unseen Species*

Assume that a random sample of n individuals (Sample 1) is taken from Assemblage 1 and a random sample of m individuals (Sample 2) is taken from Assemblage 2. Samples are taken with replacement. Denote the species frequencies in the samples by $(X_1, X_2, \ldots, X_{S_1})$ and $(Y_1, Y_2, \ldots, Y_{S_2})$, respectively. Assume that D_1 and D_2 species are respectively observed in samples 1 and 2. (If a species is missing from a sample, then X_i or Y_i will be zero.) Thus, the pair frequencies for the S_{12} shared species are $(X_1, Y_1)(X_2, Y_2) \ldots (X_{S_{12}}, Y_{S_{12}})$. Assume that D_{12} of the S_{12} shared species are actually observed, and their frequencies are the first D_{12} pairs. Note that an additional $S_{12} - D_{12}$ species are shared by the two assemblages, but absent from one or both of the samples. We refer to these as *unseen shared species*, even if present in only one of the two samples.

Recall that U and V denote, respectively, the total relative abundances associated with the *shared* species in Assemblages 1 and 2. A direct approach to obtaining an estimator that is unadjusted for the effect of unseen shared species, for any index, is to replace U and V, respectively, by $\tilde{U} = \sum_{i=1}^{D_{12}} X_i/n$ and $\tilde{V} = \sum_{i=1}^{D_{12}} Y_i/m$ (i.e., replace each species abundance by its sample abundance). Because $E(\tilde{U}) = \sum_{i=1}^{S_{12}} p_i\{1 - (1 - \pi_i)^m\} < \sum_{i=1}^{S_{12}} p_i = U$, it implies that $\tilde{U}$ underestimates U. Similarly $\tilde{V}$ underestimates V. Therefore, it follows from the well-known multivariate Jensen's inequality that the bias of an unadjusted estimator for any index listed in the last column of Table 2 is always negative, as will also be seen in the simulations. This underestimation arises because unseen shared species are ignored. Correction for the effect of unseen shared species leads to significant bias reduction when there are unseen shared species in samples. The key idea for bias correction is based on the boundary-mode Laplace approximation formula (Erkanli, 1994, 1997; Goutis and Casella, 1999). It turns out that we can use the frequencies of *observed* rare shared species to obtain an appropriate adjustment term for U and V to account for the effect of *unseen* shared species and thus remove the first-order bias. The assumptions and conclusion are outlined as follows. (A sketch of the detailed derivation is given on the *Biometrics* website.)

We assume that the number of species in each assemblage is finite and the discovery probabilities are assumed to be bounded below. Otherwise, if there were infinitely many undetectable or "invisible" shared species in hyper-diverse assemblages, then it would be impossible to obtain an accurate estimate of similarity indices. The bounded-below assumption is critical for dealing with heterogeneous assemblages; see, e.g., Huggins (2001, 2002). Let $f_{1+} = \sum_{i=1}^{D_{12}} I(X_i = 1, Y_i \geq 1)$ be the observed number of *shared* species that occur once $(X_i = 1)$ in Sample 1 (these species must be present in Sample 2, but may have any frequency). Now, let f_{2+} be the observed number of *shared* species that occur twice $(X_i = 2)$ in Sample 1. Similarly, we define f_{+1} and f_{+2} to be the observed number of shared species that occur, respectively, once $(Y_i = 1)$ and twice $(Y_i = 2)$ in Sample 2. Then the proposed estimator for U is

$$\hat{U} = \sum_{i=1}^{D_{12}} \frac{X_i}{n} + \frac{(m-1)}{m} \frac{f_{+1}}{2f_{+2}} \sum_{i=1}^{D_{12}} \frac{X_i}{n} I(Y_i = 1). \qquad (1)$$

Note that the first term in the right-hand side of equation (1) is $\tilde{U}$, the unadjusted estimator; the second term corrects for the effect of unseen shared species. Similarly, we have

$$\hat{V} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m} + \frac{(n-1)}{n} \frac{f_{1+}}{2f_{2+}} \sum_{i=1}^{D_{12}} \frac{Y_i}{m} I(X_i = 1). \qquad (2)$$

When $f_{+2} = 0$ or $f_{2+} = 0$, we suggest replacing f_{+2} and f_{2+} by $f_{+2} + 1$ and $f_{2+} + 1$, respectively. If the value of $\hat{U}$ or $\hat{V}$ is greater than 1 (which occasionally happens for highly overlapped communities), then it is replaced by 1. Replacing U and V in Table 2 by estimators $\hat{U}$ and $\hat{V}$ given in (1) and (2), we obtain our adjusted estimators for all indices. For example, the proposed abundance-based Jaccard and Sørensen estimators are, respectively, $\hat{U}\hat{V}/(\hat{U} + \hat{V} - \hat{U}\hat{V})$ and $2\hat{U}\hat{V}/(\hat{U} + \hat{V})$. The variances for the adjusted estimators are derived by a bootstrap method; details are given on the *Biometrics* website.

4. Pros and Cons of Some Indices

4.1 *Incidence-Based Indices*

When species inventories or surveys are nearly complete, incidence-based indices provide simple and intuitive overlap measures to compare two species lists, disregarding species abundance. Only species presence/absence data are required. When sample size is not sufficiently large to observe all species, it is well known (e.g., Wolda, 1981, 1983; Magurran, 2004, p. 175) that all incidence-based indices are biased and the biases are likely to be substantial for assemblages with high species richness and a large fraction of rare species. The widely used Jaccard and Sørensen indices are generally biased downward (e.g., Fisher, 1999 and this study). These indices become upward biased for the following situations: (i) shared species are relatively abundant whereas endemic species are rare; and (ii) one assemblage contains only a few or no endemic species but the other assemblage contains many rare

endemic species. Until now, we have found only one case of upward bias in our tested data sets.

One might suppose that, as long as the sampling fractions (the proportion of the total number of individuals present in an assemblage that actually appear in the samples) for two assemblages are equal, the biases associated with all incidence-based indices should be negligible. This supposition is not true, however. Theoretical bias formulas will be given below to justify our statement. We will also show by simulation (Section 6) that the biases do not diminish under equal sampling fractions. One might also conjecture that all incidence indices should work satisfactorily for the special case of equal abundance. This conjecture is not true, either. For example, the Sørensen index based on a sample is $2D_{12}/(D_1 + D_2)$; see Table 2 and Section 3 for notation. Equal abundance means that $p_1 = p_2 = \cdots = p_{S_1} \equiv p$ and $\pi_1 = \pi_2 = \cdots = \pi_{S_2} \equiv \pi$. If we approximate the expectation of a ratio by the ratio of expectations and ignore smaller-order terms, then

$$E\left(\frac{2D_{12}}{D_1 + D_2}\right) \approx \frac{2E(D_{12})}{E(D_1) + E(D_2)}$$
$$= \frac{2S_{12}\{1-(1-p)^n\}\{1-(1-\pi)^m\}}{S_1\{1-(1-p)^n\} + S_2\{1-(1-\pi)^m\}} < \frac{2S_{12}}{S_1 + S_2}.$$

The last inequality can be easily checked. Therefore, the Sørensen index underestimates even in the simplest case when all species are equally abundant. The bias cannot be reduced or removed for equal sampling fractions, neither for equal sample sizes nor for equal effort. Similarly, the Jaccard index and all the other five incidence indices listed in Table 2 exhibit the same behavior in the equal-abundance case.

For general heterogeneous cases, the bias formula for the Sørensen index is

$$\text{Bias} \approx \frac{2\sum_{i=1}^{S_{12}}\left\{1-(1-p_i)^n\right\}\left\{1-(1-\pi_i)^m\right\}}{\sum_{i=1}^{S_1}\left\{1-(1-p_i)^n\right\} + \sum_{i=1}^{S_2}\left\{1-(1-\pi_i)^m\right\}} - \frac{2S_{12}}{S_1 + S_2}. \quad (3)$$

It can be shown that this bias is negative under several species abundance models listed by Magurran (2004, Table 2.1). Not only the bias but also the variance formulas depend on species abundances (not incidence alone) and on the true index; thus, it is impossible to correct for the bias or to estimate errors without using abundance data. Even when abundance data are available, finding accurate estimators for the bias in (3) and variance is not easy. Except for the Lennon et al. (2001) index, the bias formulas for the other incidence-based indices in Table 2 can be analogously derived for heterogeneous cases and similar difficulty arises in assessing variances. The theoretical bias for the Lennon et al. (2001) index is analytically intractable. This index is not meaningful when there are no unique species in one of the two samples, because it always yields a maximum value of 1 no matter how many shared species are observed.

4.2 *Abundance-Based Indices*

The widely used Morisita-type abundance-based index (Krebs, 1999, p. 390) for a complete assemblage can be written as the following function of discovery probabilities:

$$C_M = \frac{\sum_{i=1}^{S_{12}} p_i\pi_i}{\frac{1}{2}\left(\sum_{i=1}^{S_1} p_i^2 + \sum_{i=1}^{S_2}\pi_i^2\right)}. \quad (4)$$

Morisita's original index based on pair frequency data (Krebs, 1999, p. 391) has been found to be nearly independent of sample size (e.g., Wolda, 1981), but it may exceed 1 and in such cases may lead to misleading interpretation. For example, two identical communities might yield an index value less than 1 whereas two different communities might result in a value greater than 1. To eliminate this drawback, Horn (1966) proposed the following Morisita–Horn index:

$$\hat{C}_{MH} = \frac{2\sum_{i=1}^{D_{12}}\frac{X_i}{n}\frac{Y_i}{m}}{\sum_{i=1}^{D_1}\left(\frac{X_i}{n}\right)^2 + \sum_{i=1}^{D_2}\left(\frac{Y_i}{m}\right)^2}, \quad (5)$$

which is always between 0 and 1 with the maximum value attained by two identical communities.

The index in (4) has an important probabilistic interpretation. Note that if one individual is selected randomly from each assemblage, then the probability that the two selected individuals belong to *the same shared* species is $\sum p_i\pi_i$, the numerator in (4). The denominator in (4) represents a normalized constant, which is the average of two such probabilities for two individuals drawn from the same assemblages. In contrast, our probabilistic approach (see Tables 3 and 4) is concerned with a (normalized) probability that both individuals, one from each of the two assemblages, belong to *shared* species (not necessarily to the same shared species).

Our extensive simulation has shown that the Morisita–Horn index systematically underestimates similarity (see also Ricklefs and Lau, 1980), whereas the Morisita original index slightly overestimates it. When sample sizes are sufficiently large, both indices generally perform satisfactorily. Nevertheless, a major drawback for the Morisita-type index (for many studies) is that it is highly sensitive to the most abundant species (Wolda, 1983; Magurran, 2004, p. 174). This is intuitively understandable from the preceding probabilistic interpretation because the most abundant species would contribute the major part of the probability that two randomly selected individuals belong to the *same* species. As a result, in a hyper-diverse assemblage, the index is dominated by a few abundant species and the relatively rare species (even if there are many of them) have little effect. In this case, while being relatively insensitive to compositional differences due to rarer species, the index is likely to be resistant to undersampling, because the influential abundant species are always present in samples.

Denote the total number of individuals for each species in the two complete assemblages by $(N_1, N_2, \ldots, N_{S_1})$ and $(M_1, M_2, \ldots, M_{S_2})$, respectively. The Bray–Curtis index is

$$C_{BC} = \frac{2\sum_{i=1}^{S_{12}} \min(N_i, M_i)}{\sum_{i=1}^{S_1} N_i + \sum_{i=1}^{S_2} M_i} \tag{6}$$

and the sample version is

$$\hat{C}_{BC} = \frac{2\sum_{i=1}^{D_{12}} \min(X_i, Y_i)}{\sum_{i=1}^{D_1} X_i + \sum_{i=1}^{D_2} Y_i}. \tag{7}$$

This index is also called the "quantitative Sørensen index" because it reduces to the Sørensen incidence index if for all i, $N_i = M_i = 1$. From (6), it is interesting to note that there is a probabilistic interpretation for $0.5C_{BC}$ as follows. Assume that all $\sum N_i + \sum M_i$ individuals from the two complete assemblages are *pooled* and *one* individual is randomly drawn from the pool. The index $0.5C_{BC}$ is exactly the probability that this individual belongs to a shared species and is from the assemblage in which this species has fewer individuals than in the other assemblage. When the index is applied to samples, the *observed* number of individuals for any species depends on the sampling fraction, so this index becomes meaningless for unequal sampling fraction cases. To further justify our claim, assume the sampling fraction is, respectively, α and β in Assemblages 1 and 2. Then we have $E(X_i) = \alpha N_i$ and $E(Y_i) = \beta M_i$. Comparing (6) and (7), it is readily seen that the sample Bray–Curtis makes sense only if the same sampling fractions are equal. Otherwise, the index performs erratically. Consider an extreme case in which the two assemblages are nearly identical (i.e., $N_i \approx M_i$ for all i), leading to $\hat{C}_{BC} = 2\min(\alpha, \beta)/(\alpha + \beta)$, which is surprisingly dependent only on sampling fractions, not on data at all. In general, this index has very large bias under unequal sampling fractions, as will be shown in simulations. For these reasons, from a statistical viewpoint, this index cannot be recommended unless sampling fractions are known to be equal. Given the unlikely prospect of establishing such conditions for field data, the Bray–Curtis index seems rarely to be an acceptable choice for such data. Note that equalizing the *number of individuals* in two or more samples by rarefaction before calculating the Bray–Curtis index, as suggested by Horner-Devine et al. (2004), does not equalize sampling fractions unless the assemblages themselves can be reasonably assumed to have the same total number of individuals susceptible to sampling.

5. Application to Rain Forest Succession Data

We apply our proposed indices and estimators to compare four second-growth forests and two old-growth forests in Costa Rica based on abundance data collected in 2000. The complete names, acronyms, and ages of the six forests are shown in Table 5, in which the observed species richness along with the corresponding number of individuals is also given. The summary data in Table 5 include records for canopy tree species only; shrubs, treelets, and midstory trees were excluded.

Table 5
The observed species richness in four second-growth and two old-growth sites. Forest names: Lindero Sur (LSUR), Lindero El Peje (LEP), Tirimbina (TIR), and Cuatro Rios (CR) (numbers in parentheses denote the number of individuals).

Site	Age	Seedlings	Saplings	Canopy trees
LSUR second growth	15	45 (421)	68 (1917)	12 (88)
TIR second growth	18	49 (817)	74 (1003)	16 (99)
LEP second growth	23	47 (551)	67 (1199)	24 (169)
CR second growth	28	57 (699)	91 (1297)	33 (211)
LSUR old growth	>200	47 (300)	101 (508)	37 (119)
LEP old growth	>200	69 (557)	102 (729)	43 (111)

As briefly described in the Introduction, the aim is to compare compositional similarity between different tree sizes as measured by diameter at breast height (DBH). Three sizes are considered: trees (≥25 cm in DBH), saplings (1–5 cm in DBH), and seedlings (>20 cm in height, but <1 cm in DBH). All trees and saplings were marked and measured for diameter within a 1 ha plot in each forest. Seedlings were sampled in 144 1 × 5 m quadrats within the 1 ha plot, for a total area sampled of 0.072 ha. Table 5 shows that young sites have fewer canopy tree species per hectare and fewer tree sapling species compared to old-growth forests, but differences in tree seedling species richness were less pronounced. We focus here on comparing species composition between trees and saplings (or seedlings). Table 6 shows the paired abundance counts for the LEP old-growth site only. Data for the other five sites are given on the *Biometrics* website.

Various similarity indices/estimates between seedlings and trees are shown in Table 7 and the corresponding results between saplings and trees are given in Table 8. In each table, we present three incidence-based indices (classic Jaccard, classic Sørensen, and Lennon et al.) and six abundance-based indices (Bray–Curtis, Morisita–Horn, unadjusted and adjusted Jaccard, and unadjusted and adjusted Sørensen). Except for our proposed adjusted indices, all the others ignore the effect of missing species. The estimated SEs based on 200 bootstrap replications are shown only for the adjusted abundance indices.

The classic Jaccard and Sørensen indices show low compositional similarity between trees and seedlings (Table 7) for the four second-growth forests compared to tree–seedling similarity for old-growth forests, with similarity decreasing slightly with forest age among the four second-growth forests. Similarity between trees and saplings (Table 8), in contrast, show gradual increases from the youngest forest to the older second-growth forest, continuing the trend to old-growth forests.

In both tables, the Lennon et al. (2001) index generally (except for the CR site in Table 8) decreases across the four second-growth forests. The youngest forest has the highest index. In all sites, there are always fewer unshared species among the trees than among individuals of smaller size classes. The Lennon et al. (2001) index thus becomes the

Table 6

Abundances of shared and unique species in the LEP old-growth (>200 years) site (69 seedling species, 102 sapling species, and 43 tree species)

(a) Abundance vectors for seedlings versus trees													
26 shared species	X_i	17	121	16	6	6	1	4	17	5	1	7	7
	Y_i	7	31	6	2	3	1	2	3	2	2	4	1
		20	5	3	1	1	3	1	2	3	1	2	1
		2	2	3	2	2	1	1	1	1	1	1	3
		1	3										
		1	1										
43 unique seedlings	X_i	14	10	5	73	2	11	4	9	2	4	6	17
		6	2	2	1	1	29	3	1	2	1	7	5
		1	1	1	5	3	3	1	1	1	1	1	1
		54	3	3	2	1	1	1					
17 unique trees	Y_i	5	1	1	1	3	1	1	1	1	2	2	1
		1	1	1	1	1							
(b) Abundance vectors for saplings versus trees													
32 shared species	X_i	48	37	30	23	21	19	17	17	16	15	14	12
	Y_i	7	31	6	2	3	1	2	3	2	2	4	5
		11	11	9	9	8	7	6	5	5	5	4	4
		1	2	1	2	1	3	2	1	2	1	1	1
		2	2	2	2	1	1	1	1				
		3	1	1	1	1	1	1	1				
70 unique saplings	X_i	38	27	24	19	19	16	16	15	14	13	9	8
		8	8	8	7	7	6	6	5	5	5	4	4
		4	4	3	3	3	2	2	2	2	2	2	2
		2	2	2	2	2	2	2	2	1	1	1	1
		1	1	1	1	1	1	1	1	1	1	1	1
		1	1	1	1	1	1	1	1	1	1		
11 unique trees	Y_i	2	2	1	1	1	1	1	1	3	1	1	

ratio of shared species richness to tree species richness. Among the four secondary forests, this ratio decreases almost monotonically with forest age. The number/proportion of unshared seedlings or saplings is completely ignored, though it is part of the hypothesis about succession.

The Bray–Curtis index for the four secondary forests shows little successional trend. From our arguments in Section 4 regarding the Bray–Curtis index, we have reservations about the use of this index here because the sampling fractions are unknown for these data.

For comparing seedlings versus trees (Table 7), the Morisita–Horn index yields very high similarity estimates (0.89 and 0.74) for the two old-growth forests because they are both dominated by a single, very abundant, shared species (*Pentaclethra macroloba*, the second species in the shared list in Table 6a). For example, in the LSUR old-growth site, the relative seedling abundance for this shared species is 36% and for trees is 48%, i.e., nearly half of the tree individuals belong to this species. Similarly, in the LEP old-growth site, the relative abundance for seedlings of the same species is 22% and for trees is 28%. As discussed in Section 4, the Morisita–Horn index is highly dependent on this highly dominant species. In contrast, between saplings and trees (Table 8) in the LSUR and LEP old-growth sites, no such dominant species exist, so the Morisita–Horn index drops, respectively, to 0.17 and 0.45. None of the three incidence-based nor either of the two previously published abundance-based indices (the Bray–Curtis and Morisita–Horn) exhibit the expected trend.

For our abundance-based Jaccard and Sørensen indices, the adjusted estimate is always higher than the corresponding unadjusted one because of the presence in sample pairs of observed, shared, rare species. In the old-growth sites, more data and more shared information are available than for the second-growth sites, so the adjusted effects are relatively more precise, as reflected by the smaller estimated SEs. For the secondary forests, due to less shared information, the estimated SEs are relatively large and the adjusted effects in some cases are less precise, especially in the CR site in Table 7 and the TIR site in Table 8. In these two sites, we note $f_{2+} = f_{+2} = 0$. More discussion is given in Section 7.

Except for the adjusted estimates in the TIR site in Table 8, both the unadjusted and adjusted Jaccard and Sørensen indices generally follow the expected pattern across the six forest stands. Compositional similarity between seedling and tree

Table 7

Various similarity indices/estimates for seedlings versus trees (SE are given only for the adjusted abundance-based Jaccard and Sørensen indices)

	Second-growth forest				Old-growth forest	
Index	LSUR (15 years)	TIR (18 years)	LEP (23 years)	CR (28 years)	LSUR (>200 years)	LEP (>200 years)
Incidence based						
Jaccard	0.19	0.16	0.16	0.14	0.25	0.30
Sørensen	0.32	0.28	0.28	0.24	0.40	0.46
Lennon et al. (2001)	0.75	0.56	0.42	0.33	0.46	0.60
Abundance based						
Bray–Curtis	0.13	0.11	0.16	0.12	0.40	0.24
Morisita–Horn	0.19	0.34	0.28	0.53	0.89	0.74
Unadjusted Jaccard	0.42	0.29	0.20	0.26	0.45	0.40
Adjusted Jaccard	0.45	0.37	0.22	0.44	0.50	0.48
(SE)	(0.18)	(0.19)	(0.11)	(0.23)	(0.16)	(0.14)
Unadjusted Sørensen	0.59	0.44	0.33	0.41	0.62	0.58
Adjusted Sørensen	0.62	0.54	0.36	0.61	0.66	0.65
(SE)	(0.19)	(0.19)	(0.14)	(0.22)	(0.15)	(0.13)

Table 8
Various similarity indices/estimates for saplings versus trees (SEs are given only for the adjusted abundance-based Jaccard and Sørensen indices)

	Second-growth forest				Old-growth forest	
Index	LSUR (15 years)	TIR (18 years)	LEP (23 years)	CR (28 years)	LSUR (>200 years)	LEP (>200 years)
Incidence based						
Jaccard	0.14	0.15	0.20	0.22	0.27	0.28
Sørensen	0.25	0.27	0.33	0.35	0.42	0.44
Lennon et al. (2001)	0.83	0.75	0.63	0.67	0.78	0.74
Abundance based						
Bray–Curtis	0.07	0.11	0.09	0.10	0.20	0.23
Morisita–Horn	0.19	0.14	0.15	0.30	0.17	0.45
Unadjusted Jaccard	0.39	0.20	0.17	0.24	0.33	0.46
Adjusted Jaccard	0.41	0.49	0.20	0.40	0.45	0.55
(SE)	(0.18)	(0.27)	(0.13)	(0.21)	(0.14)	(0.11)
Unadjusted Sørensen	0.56	0.33	0.29	0.38	0.50	0.63
Adjusted Sørensen	0.58	0.66	0.34	0.57	0.62	0.71
(SE)	(0.20)	(0.24)	(0.16)	(0.20)	(0.13)	(0.09)

assemblages and between sapling and tree assemblages was initially high in the youngest stand. As the forest matures, tree seedling and sapling pools become enriched by shade-tolerant species not represented as canopy trees, resulting in a decreasing compositional similarity between trees and younger stages that reached a minimum in the 23-year-old LEP stand. This minimum similarity represents a point in forest succession of maximum recruitment limitation for both seedlings and saplings. In the oldest second-growth plot, CR, our indices began to increase, reflecting the maturity and reproduction of shade-tolerant species. For comparing seedlings and trees (Table 7), the LSUR old-growth stand has the maximum similarity among the six forests, whereas the LEP old-growth attains the maximum similarity between saplings and trees (Table 8).

Based on the proposed adjusted estimates we find, except for the TIR site in Table 8, that old-growth forests have higher Jaccard and Sørensen similarity between trees and younger stages than any of the secondary sites. The compositional similarity in the 28-year-old second-growth forests is comparable to that observed within a 15-year-old second-growth forest; both are slightly less than those for old-growth sites but higher than those of intermediate ages.

6. Simulation

A simulation experiment was conducted to investigate the performance of the proposed indices and to compare them with the existing ones. We selected individuals with replacement from the abundance vectors. Table 6 shows the vectors for the LEP old-growth site. The abundance records are regarded as complete assemblages for the purposes of these simulations. We carried out a total of 12 simulation studies (seedlings versus trees and saplings versus trees for each of the six forests). The conclusions based on all studies are generally consistent and thus only results of seedlings versus trees for a site with more data (the LEP old-growth forest) are presented in Table 9.

From Table 5, in the LEP old-growth site, there were 69 seedling species (represented by 557 individuals) and 43 tree species (represented by 111 individuals). There were 26 species in common between seedlings and trees. We report here 10 combinations of sampling fraction. The first five are equal sampling-fraction cases (10% versus 10%, 30% versus 30%, ..., 90% versus 90%) and the others are unequal sampling-fraction cases (10% versus 60%, 20% versus 70%, ..., 50% versus 100%). All subsampling was done by selecting individuals with replacement. For example, in the case of 10% versus 10%, we randomly selected 56 individuals ($557 \times 10\% = 56$) from the seedling abundances and 11 individuals ($111 \times 10\% = 11$) from the tree abundances. Similarly for the unequal sampling fraction 50% versus 100% for comparing seedlings versus trees, we randomly selected 279 individuals ($557 \times 50\% = 279$) from the seedling abundances and 111 individuals ($111 \times 100\% = 111$) from the tree abundances. In each case, the two sets of selected individuals were then classified by species identity.

For each fixed sampling fraction, 5000 simulated sets of sample data were generated. Then for each simulated set, we calculated three incidence-based indices (Jaccard, Sørensen, and Lennon et al.) and six abundance-based indices (Bray–Curtis, Morisita–Horn, and unadjusted and adjusted Jaccard and Sørensen). Because each index is designed to measure different aspects of "similarity" and each has its own true value, comparison of the *absolute* magnitude under some traditional comparison criteria (bias, variance, and mean squared error) is statistically and biologically meaningless. Also, as will be discussed later, most indices are biased, thus comparison based on sample coefficient of variation is not statistically valid either. In this article, we adopt a percentage relative bias (relative to each true value) as our comparison criterion.

For each index/estimate, the average percentage of relative bias over 5000 simulation trials was calculated and given in Table 9. Also, the averages of the observed number of seedling, tree species, and the shared species are also shown in the

Table 9

Percentages (%) of average relative bias over 5000 *simulation trials generated from the LEP old-growth data of seedlings versus trees (the two assemblages include* 69 *seedlings,* 43 *trees, and* 26 *shared species; the three numbers under each fraction denote the average of observed saplings, trees, and shared species in samples)*

		Sampling fraction				
Index	True value	10% versus 10% (23, 8, 3)	30% versus 30% (40, 18, 8)	50% versus 50% (48, 24, 12)	70% versus 70% (54, 29, 15)	90% versus 90% (57, 32, 17)
Incidence based						
Jaccard	0.30	−63.6	−44.0	−31.7	−24.1	−19.0
Sørensen	0.46	−58.0	−38.0	−26.6	−19.8	−15.5
Lennon et al. (2001)	0.60	−36.0	−21.7	−15.0	−11.2	−8.8
Abundance based						
Bray–Curtis	0.24	−35.3	−20.4	−14.0	−11.2	−9.4
Morisita–Horn	0.74	−37.7	−16.3	−10.3	−7.3	−5.4
Unadjusted Jaccard	0.40	−47.6	−28.3	−18.9	−13.5	−10.2
Adjusted Jaccard	0.40	−30.2	0.6	3.9	3.7	2.7
Unadjusted Sørensen	0.58	−40.1	−22.3	−14.4	−10.1	−7.6
Adjusted Sørensen	0.58	−26.1	−1.3	1.7	2.0	1.6

		Sampling fraction				
Index	True value	10% versus 60% (23, 27, 7)	20% versus 70% (33, 29, 10)	30% versus 80% (40, 30, 13)	40% versus 90% (45, 32, 14)	50% versus 100% (48, 33, 16)
Incidence based						
Jaccard	0.30	−42.7	−32.3	−26.4	−22.5	−19.6
Sørensen	0.46	−36.9	−27.2	−21.9	−18.5	−16.0
Lennon et al. (2001)	0.60	−47.1	−38.9	−30.1	−24.1	−20.1
Abundance based						
Bray–Curtis	0.24	38.3	58.7	51.8	44.8	38.6
Morisita–Horn	0.74	−15.1	−10.1	−8.3	−7.0	−6.4
Unadjusted Jaccard	0.40	−29.1	−21.4	−17.2	−14.2	−12.1
Adjusted Jaccard	0.40	−5.3	0.0	1.4	2.2	2.6
Unadjusted Sørensen	0.58	−23.0	−16.5	−13.1	−10.7	−9.0
Adjusted Sørensen	0.58	−5.3	−0.9	0.3	1.0	1.4

same table. Note that our sampling was conducted *with* replacement, so there were some unseen species even in the case of 100% sampling fraction. For example, in the case of 100% fraction for trees (the last case in Table 9), there were, on average, only 33 tree species observed, out of 43 species in the sampling pool.

Based on Table 9 and on unreported simulation results for the other five sites, we summarize the following findings:

1. In the most severe undersampling case (10% versus 10%), all indices perform badly due to sparse information, as anticipated. The performance of each index improves when the sampling fraction is increased.
2. We confirmed previous findings that the three incidence-based indices typically underestimate the true values, especially for smaller sample sizes for which more species were missed in samples. Although the magnitude of biases decreases as more species are observed, it is clear that an equal-sampling scheme does not eliminate biases, as indicated in Section 4.
3. The Bray–Curtis index performs well when sampling fractions are equal, as the theory predicts in Section 4. However, this index exhibits unreasonably large positive biases when sampling fractions are not equal. It is suggested that this index be used with caution.
4. The results are consistent with the conclusion of Ricklefs and Lau (1980) that the Morisita–Horn index is negatively biased by undersampling. Setting aside our adjusted estimators, we see that this index is the least sensitive to sample size and performs well in terms of relative biases. Its general performance is superior to our unadjusted indices (at the expense of insensitivity to rarer species), but inferior to the adjusted ones.
5. In all cases, the adjustment for both Jaccard and Sørensen indices significantly reduces the bias of the corresponding unadjusted ones, especially for smaller sampling fractions. The improvement is evident for all sampling fractions. Except for the 10% versus 10% case, the adjusted estimates are generally resistant to undersampling as the relative biases are quite stable for all the other cases.
6. The two adjusted estimators on average have low relative biases. They are negatively biased in severe undersampling situations, whereas they are slightly positively biased as data improve. The adjusted Sørensen index tends to have smaller relative bias than the adjusted Jaccard index.

7. Concluding Remarks and Discussion

We have presented a class of abundance-based indices (in Table 2) based on a probabilistic approach. For these indices, estimators that adjust for the effect of unseen shared species

are proposed. Simulation results have generally shown that the adjustment is essential and significantly reduces bias associated with the corresponding unadjusted index when there is a large proportion of unseen shared species in samples, as in the tropical successional vegetation data set. Although we specifically deal with the selected indices listed in Table 2, our approach can be directly applied to generalize any symmetric and homogeneous incidence-based index to an abundance-based counterpart.

We remark that sufficient shared information is required to make a stable adjustment. How large should the samples be to provide sufficient information? Our experiences suggest that (1) the data should have at least 10 shared species and (2) there should be at least one shared species that occurs twice in one assemblage (i.e., data with $f_{2+} = f_{+2} = 0$ would generally lead to unstable estimates as these two statistics appear in the denominators of the adjusted effects; see equations (1) and (2)). Otherwise, the effect of unseen species cannot be accurately estimated and more data should be collected.

The advantages and disadvantages of some existing indices have also been briefly discussed in this article (Section 4). We emphasize that each index is derived from different theoretical justifications and each measures different aspects of assemblages. The incidence-based indices are useful for simply comparing species lists and are valid when samples are nearly complete. A simple stopping rule for a complete census is that all observed species occur at least twice (Colwell and Coddington, 1994). However, this requires more information than presence/absence only. For incomplete surveys, the incidence indices are typically biased downward. Accurate bias correction and variance assessment are generally impossible based on incidence data alone. Contrary to one's expectation, biases exist even under an equal-sampling scheme, under equal sampling effort, or for equal-abundance assemblages.

The Morisita-type and Bray–Curtis indices focus on comparing abundances, *species by species*, from the probabilistic interpretations provided in Section 4. The Morisita–Horn index has the advantage that it is not strongly sensitive to species richness and sample sizes. The Bray–Curtis index behaves satisfactorily for equal-sampling situations, but this index is not statistically meaningful in unequal-sampling cases. These two abundance-based indices primarily measure similarity in the composition of dominant species, so they are unavoidably strongly affected by a few dominant species and actually ignore the effect of rare species. Our proposed abundance indices are formulated by pooling shared abundances and are, thus, less likely to be dominated by particular species, but detailed species by species composition is not accounted for. We recommend that biologists consider carefully what it is that they mean by "similarity" or "overlap" and select an index accordingly, taking into account the nature and limitations of their data.

In this article, our estimation procedure for any abundance-based index is based on abundance data from each assemblage. With slight modification in model formulation and arguments, our procedure can be extended to deal with replicated incidence data. Assume that we take a set of replicated incidence samples from each assemblage. By replacing abundance frequencies and sample size by incidence frequencies and the number of replicated samples, generally parallel derivations yield the same types of estimators as those presented in this article (Section 3.2). Details are provided in the ecological counterpart paper (Chao et al., 2005).

The new similarity index estimators presented in this article (and their bootstrap variance estimators) are included in Version 7.5 of EstimateS (Colwell, 1994–2005), and in program SPADE (*S*pecies *P*rediction *A*nd *D*iversity *E*stimation); see Chao and Shen (2003–2005).

Acknowledgements

This work was supported by Taiwan National Science Council contract NSC92-2118-M007-013 to A. Chao and T.-J. Shen, by a grant from the Andrew W. Mellon Foundation to R. L. Chazdon, and by US-NSF grant DEB-0072702 to R. K. Colwell. We thank an associate editor and a reviewer for their helpful comments and suggestions.

References

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest assemblages of southern Wisconsin. *Ecological Monograph* **27,** 325–349.

Chao, A. and Shen, T. J. (2003–2005). Program SPADE (Species Prediction And Diversity Estimation). Program and User's Guide at the website. Available at `http://chao.stat.nthu.edu.tw`.

Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T.-J. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters* **8,** 148–159.

Chazdon, R. L., Colwell, R. K., Denslow, J. S., and Guariguata, M. R. (1998). Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. In *Forest Biodiversity Research, Monitoring and Modeling: Conceptual Background and Old World Case Studies*, F. Dallmeier and J. Comiskey (eds), 285–309. Paris: Parthenon Publishing.

Chazdon, R. L., Redondo Brenes, A., and Vilchez Alvarado, B. (2005). Effects of climate and stand age on annual tree dynamics in tropical second-growth rain forests. *Ecology* **86,** 1808–1815.

Colwell, R. K. (1994–2005). EstimateS: Statistical estimation of species richness and shared species from samples. URL `http://viceroy.eeb.uconn.edu/estimates`. Persistent URL `http://purl.oclc.org/estimates`.

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

Erkanli, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *Journal of the American Statistical Association* **89,** 250–258.

Erkanli, A. (1997). Boundary-mode approximations for posterior expectations. *Journal of Statistical Planning and Inference* **58,** 217–239.

Fisher, B. L. (1999). Improving inventory efficiency: A case study of leaf-litter ant diversity in Madagascar. *Ecological Applications* **9,** 714–731.

Gotelli, N. and Ellison, A. M. (2004). *A Primer of Ecological Statistics*. Sunderland, Massachusetts: Sinauer Associates.

Goutis, C. and Casella, G. (1999). Explaining the saddlepoint approximation. *American Statistician* **53,** 216–224.

Gower, J. C. (1985). Measures of similarity, dissimilarity and distance. In *Encyclopedia of Statistical Sciences*, Volume 5, S. Kotz and N. L. Johnson (eds), 397–405. New York: Wiley.

Grassle, J. F. and Smith, W. (1976). A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. *Oecologia* **25,** 13–22.

Guariguata, M. R., Chazdon, R. L., Denslow, J. S., Dupuy, J. M., and Anderson, L. (1997). Structure and floristics of secondary and old-growth forest stands in lowland Costa Rica. *Plant Ecology* **132,** 107–120.

Horn, H. S. (1966). Measurement of "overlap" in comparative ecological studies. *American Naturalist* **100,** 419–424.

Horner-Devine, M. C., Lage, M., Hughes, J. B., and Bohannan, B. J. M. (2004). A taxa–area relationship for bacteria. *Nature* **432,** 750–753.

Hubalek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews* **57,** 669–689.

Huggins, R. M. (2001). A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneous capture probabilities. *Statistics and Probability Letters* **54,** 147–152.

Huggins, R. M. (2002). A parametric empirical Bayes approach to the analysis of capture-recapture experiments. *Australian and New Zealand Journal of Statistics* **44,** 55–62.

Janson, S. and Vegelius, J. (1981). Measures of ecological association. *Oecologia* **49,** 371–376.

Krebs, C. J. (1999). *Ecological Methodology*, 2nd edition. Menlo Park, California: Benjamin/Cummings.

Lennon, J. J., Koleff, P., Greenwood, J. J. D., and Gaston, K. J. (2001). The geographical structure of British bird distributions: Diversity, spatial turnover and scale. *Journal of Animal Ecology* **70,** 966–979.

Longino, J. T., Coddington, J., and Colwell, R. K. (2002). The ant fauna of a tropical rain forest: Estimating species richness three different ways. *Ecology* **83,** 689–702.

Ludwig, J. A. and Reynolds, J. F. (1988). *Statistical Ecology: A Primer on Methods and Computing*. New York: Wiley.

Magurran, A. E. (2004). *Measuring Biological Diversity*. Oxford: Blackwell.

Plotkin, J. B. and Muller-Landau, H. C. (2002). Sampling the species composition of a landscape. *Ecology* **83,** 3344–3356.

Redondo Brenes, A., Vilchez Alvarado, B., and Chazdon, R. L. (2001). Estudio de la dinámina y composición de cuatro bosques secundarios en la región Huetar Norte, Sarapiquí, Costa Rica. *Revista Forestal Centroamericana* **36,** 21–26.

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

Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science* **241,** 1–31.

Smith, W., Solow, A. R., and Preston, P. E. (1996). An estimator of species overlap using a modified beta-binomial model. *Biometrics* **52,** 1472–1477.

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

Wolda, H. (1983). Diversity, diversity indices and tropical cockroaches. *Oecologia* **58,** 290–298.

Yue, J. C., Clayton, M. K., and Lin, F.-C. (2001). A nonparametric estimator of species overlap. *Biometrics* **57,** 743–749.

Received October 2004. *Revised July* 2005.
Accepted August 2005.