

# A Two-Stage Probabilistic Approach to Multiple-Community Similarity Indices

Anne Chao,<sup>1,\*</sup> Lou Jost,<sup>2</sup> S. C. Chiang,<sup>1</sup> Y.-H. Jiang,<sup>1</sup> and Robin L. Chazdon<sup>3</sup>

<sup>1</sup>Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043

<sup>2</sup>Via a Runtun, Baños, Tungurahua, Ecuador

<sup>3</sup>Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269-3043, U.S.A.

\**email*: chao@stat.nthu.edu.tw

**SUMMARY.** A traditional approach for assessing similarity among  $N$  ( $N > 2$ ) communities is to use multiple pairwise comparisons. However, pairwise similarity indices do not completely characterize multiple-community similarity because the information shared by at least three communities is ignored. We propose a new and intuitive two-stage probabilistic approach, which leads to a general framework to simultaneously compare multiple communities based on abundance data. The approach is specifically used to extend the commonly used Morisita index and NESS (normalized expected species shared) index to the case of  $N$  communities. For comparing  $N$  communities, a profile of  $N - 1$  indices is proposed to characterize similarity of species composition across communities. Based on sample abundance data, nearly unbiased estimators of the proposed indices and their variances are obtained. These generalized NESS and Morisita indices are applied to comparison of three size classes of plant data (seedling, saplings, and trees) within old-growth and secondary rain forest plots in Costa Rica.

**KEY WORDS:** Abundance data; Beta diversity; Morisita index; NESS index; Shared species; Species overlap.

## 1. Introduction

In taxonomic and ecological research, similarity or dissimilarity indices provide quantitative information for comparing species composition of two or more assemblages (Colwell and Coddington, 1994; Arita and Rodríguez, 2002; Magurran, 2004). We use the term “assemblage” and “community” interchangeably in this article.

There are many similarity indices based on presence/absence (incidence) data for comparing assemblages or studying spatial patterns of species diversity (Gower, 1985; Ludwig and Reynolds, 1988; Magurran, 2004). However, incidence-based measures do not take species abundance into account, giving an ecologically unrealistic view of similarity. A maple forest with one pine tree is not ecologically identical to a pine forest with one maple tree, but incidence-based similarity measures would find them identical. Furthermore, when sample size is not sufficiently large to observe all species, it is well known (e.g., Wolda, 1981, 1983; Magurran, 2004) that all estimators of the incidence-based indices are generally biased downward and the biases are likely to be substantial for assemblages with high species richness and a large fraction of rare species. A major statistical concern is that, based on incidence data, it is difficult to accurately correct for bias and to assess variance (see Chao et al. [2005, 2006] for relevant discussions).

An ecologically more meaningful and statistically more reliable diversity or similarity measure would take into account differences in species frequencies between assemblages (Smith

and Grassle, 1977; Yue, Clayton, and Lin, 2001). In contrast to previous analyses based on parametric approaches (e.g., Smith, Solow, and Preston, 1996; Plotkin and Muller-Landau, 2002), we focus here on nonparametric indices that do not depend on species abundance distribution. If species abundance data are available, a widely used nonparametric similarity index for comparing two communities is the Morisita index (Magurran, 2004). The Morisita index is based on the Simpson concentration (Simpson, 1949; Jost, 2006), and so is mainly sensitive to dominant species; relatively rare species have little effect. Although this property has sometimes been regarded as a drawback, it actually provides statistical advantages, and the Morisita index is therefore a useful tool for comparisons based on common species. Grassle and Smith (1976) proposed the NESS (normalized expected species shared) index for samples of size  $m$ , which is a generalized form of the Morisita index for comparing two communities. The NESS index takes rare species into account when  $m$  is large, and so is less dependent on dominant species.

A traditional approach for simultaneously comparing more than two communities in ecology is to use multiple pairwise comparisons. However, the similarity of three or more communities cannot be reduced to a set of pairwise similarities. A simple example is given in Table 1. The two sets of communities, Communities (1, 2, 3) and Communities (4, 5, 6) in Table 1 have the same *pairwise* overlap, but evidently the two sets are not *globally* similar, because their gamma diversities differ. In the first set, there is a species shared by all

**Table 1**

Two sets of communities with same pairwise similarity but different global similarity. (A sign “X” denotes existence of a species. In each community, there are two equally abundant species.)

Species	First set of communities			Second set of communities		
	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
1	X	X	X	X	—	X
2	X	—	—	X	X	—
3	—	X	—	—	X	X
4	—	—	X	—	—	—

three communities whereas in the second set, none are shared by all communities. Thus, multiple pairwise similarity indices do not completely characterize multiple-community similarity because the information shared by at least three communities is ignored.

Our research was motivated by comparing compositional similarity between different tree size classes in Costa Rican forests. Three sizes are considered: trees ( $\geq 25$  cm in diameter at breast height [DBH]), saplings (1–5 cm in DBH), and seedlings ( $> 20$  cm in height, but  $< 1$  cm in DBH); these are compared within second-growth forests and old-growth forests in Costa Rica. The abundance data for the three size classes of tree species were collected as part of a larger study of secondary succession of tropical rain forests following pasture abandonment (Chazdon et al., 1998; Chazdon, Brenes, and Alvarado, 2005). The sampling plot and detailed analysis will be discussed in Section 4. The full data sets are included as Web-based supplementary materials of this article.

Previous statistical analyses of the above forest data were limited to pairwise comparisons (Chao et al., 2005, 2006) due to lack of proper abundance-based multiple-assemblage indices at that time. When we compared pairwise similarities in a secondary forest with those in a primary forest based on the Morisita index, we found that the two forests had similar trees versus saplings similarity, the secondary forest has a much higher similarity for saplings versus seedlings, whereas the primary forest has a much higher similarity for trees versus seedlings (see Section 4). Is there a way to simultaneously compare trees, saplings, and seedlings? In the pairwise comparisons, the overlap information among three size classes (i.e., those species with all three size classes present in a plot) is ignored. We were thus motivated to find multiple-assemblage similarity measures by incorporating all overlap information.

In this article we explore a two-stage probabilistic approach, which leads to a general framework to construct multiple-assemblage similarity. This approach is used specifically to generalize the Morisita index and the NESS index to the case of comparing  $N$  ( $N > 2$ ) assemblages in which a profile of  $N - 1$  indices are needed to characterize similarity across assemblages. An additional probabilistic quantity  $g$ , which varies the sensitivity of these indices to rare and common species, emerges naturally from this approach and may be called the “order” of the similarity index, analogous to the “order” of a diversity index (Hill, 1973; Jost, 2006). These generalized Morisita and NESS indices are applied to our forest data.

In Section 2, we derive the Morisita and NESS indices for two assemblages using our two-stage probabilistic interpretation. In Section 3, we extend the probabilistic approach to the case of more than two assemblages. Section 3.1 presents the case for three assemblages and Section 3.2 extends it to a general case of  $N$  assemblages. Estimators and their variances of the proposed indices are briefly described. Application to our forest data is provided in Section 4. Concluding remarks and discussions are made in Section 5.

## 2. Comparing Two Assemblages

### 2.1 Indices for Assemblages

The goal of statistical analysis here is to describe similarity of two communities or assemblages based on sample data from each assemblage. In Section 2.1, we formulate theoretical similarity indices for two assemblages. In Section 2.2, estimation of similarity indices based on sample data is described.

Assume that the total number of species in Assemblage 1 is  $S_1$  (unknown) and there are  $S_2$  (unknown) species in Assemblage 2. Let the number of shared species be  $S_{12}$  (unknown). The total number of species in the combined assemblage is  $S = S_1 + S_2 - S_{12}$ . Assume that all these  $S$  species are indexed by  $1, 2, \dots, S$ . Using an index in the combined assemblage will simplify notation and formulas, as will be seen later. Let  $p_{ij}$ ,  $i = 1, \dots, S$ ,  $j = 1, 2$  denote the (unknown) probability that an individual chosen randomly from Assemblage  $j$  is of species  $i$ . When all individuals have the same probability of being selected, the two sets of probabilities  $\{p_{11}, p_{21}, \dots, p_{S1}\}$  and  $\{p_{12}, p_{22}, \dots, p_{S2}\}$  also represent the species relative abundances. If species  $i$  is a shared species, then  $p_{i1} > 0$  and  $p_{i2} > 0$ ; if any species  $i$  is a unique species in Assemblage 1, then  $p_{i2} = 0$ ; if species  $i$  is a unique species in Assemblage 2, then  $p_{i1} = 0$ .

The widely used Morisita abundance-based index (Krebs, 1999) can be written as

$$C_{22} = \frac{2 \sum_{i=1}^S p_{i1} p_{i2}}{\sum_{i=1}^S p_{i1}^2 + \sum_{i=1}^S p_{i2}^2}. \quad (1)$$

(The meaning of the subscript in  $C_{22}$  will be clearly seen later.) It is evident that in the above formula there are only  $S_{12}$ ,  $S_1$ , and  $S_2$  nonzero terms in the summation  $\sum p_{i1} p_{i2}$ ,  $\sum p_{i1}^2$ , and

$\sum p_{i2}^2$ , respectively. Here all indices can be summed from 1 to  $S$ , because species are indexed in the combined assemblage, rather than in each assemblage. In all formulas throughout the article, the index  $i$  (which denotes a species index in the combined assemblage) is always summed from 1 to  $S$ .

Horn (1966) gave a simple probabilistic interpretation separately for the numerator and denominator of the above index. In this article, we propose a two-stage probabilistic approach that lays the groundwork for constructing multiple-assemblage indices. At Stage I of this approach, we sequentially make two selections; each selection can be either Assemblage 1 or Assemblage 2 with equal probability. For such selections, we have four possible outcomes (1, 1), (2, 2), (1, 2), and (2, 1) with equal probability of 1/4. Here the outcome (1, 1) means that both selections result in Assemblage 1, (1, 2) means that first selection results in Assemblage 1 whereas the second selection results in Assemblage 2, and similar interpretation pertains to the other two outcomes. At Stage II, for the two assemblages selected in Stage I, we randomly select *one* individual from each of the two assemblages.

Now what is the conditional probability of the event  $Z$  that two individuals selected in Stage II belong to the same species given some specific outcomes in Stage I? We consider two disjoint sets of outcomes in Stage I: the two selections result in either the same assemblages or different assemblages. It follows from the definition of a conditional probability that we have

$$P(Z | \text{Stage I results in the same assemblages}) = \frac{(1/4) \sum p_{i1}^2 + (1/4) \sum p_{i2}^2}{(1/4) + (1/4)}, \tag{2a}$$

$$P(Z | \text{Stage I results in different assemblages}) = \frac{(1/4) \sum p_{i1}p_{i2} + (1/4) \sum p_{i2}p_{i1}}{(1/4) + (1/4)} = \sum p_{i1}p_{i2}. \tag{2b}$$

The Morisita index  $C_{22}$  defined in (1) is the ratio of these two conditional probabilities. Thus the first sub-index of  $C_{22}$  means that we select *two* assemblages in Stage I selection, and the second sub-index means we are comparing *two* target assemblages. We remark that in the numerator of (2a),  $\sum p_{i1}^2$  and  $\sum p_{i2}^2$  denote, respectively, the Simpson concentration index for Assemblages 1 and 2 (Magurran, 2004). For example, in Assemblage 1,  $\sum p_{i1}^2$  is the probability of picking two individuals that are from the same species if all individuals have the same probability of being selected.

The Morisita index is primarily sensitive to the abundant species (Wolda, 1983; Magurran, 2004). This is intuitively understandable from the preceding probabilistic interpretation because in Stage II the 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 hyperdiverse assemblage, the index is dominated by abundant species and the relatively rare species (even if there are many of them) have little effect.

Instead of selecting one individual in Stage II, we could select  $m$  individuals with replacement from each of two assemblages selected in Stage I. Denote the random vari-

able  $Y$  as the number of species that appear in both sets of  $m$  individuals. Define  $\mu_{i1}(m) = 1 - (1 - p_{i1})^m$  and  $\mu_{i2}(m) = 1 - (1 - p_{i2})^m$ . Analogous to (2a) and (2b), it follows that

$$E(Y | \text{Stage I results in the same assemblages}) = \frac{(1/4) \sum [\mu_{i1}(m)]^2 + (1/4) \sum [\mu_{i2}(m)]^2}{(1/4) + (1/4)}, \tag{3a}$$

$$E(Y | \text{Stage I results in different assemblages}) = \sum \mu_{i1}(m)\mu_{i2}(m). \tag{3b}$$

The classic NESS index (Grassle and Smith, 1976) for size  $m$  is the ratio of (3b) to (3a). Note (3a) and (3b) are obtained by replacing  $p_{i1}$  by  $\mu_{i1}(m)$  and  $p_{i2}$  by  $\mu_{i2}(m)$  in (2a) and (2b). Thus, the classic NESS index, which is denoted as  $NESS_{22}(m)$  in our approach, can be obtained directly from the Morisita index by substituting  $p_{i1}$  by  $\mu_{i1}(m)$  and  $p_{i2}$  by  $\mu_{i2}(m)$  in (1). That is,

$$NESS_{22}(m) = \frac{2 \sum \mu_{i1}(m)\mu_{i2}(m)}{\sum [\mu_{i1}(m)]^2 + \sum [\mu_{i2}(m)]^2}. \tag{4}$$

Note that when  $m = 1$ , we have  $\mu_{i1}(m) = p_{i1}$  and  $\mu_{i2}(m) = p_{i2}$ , thus  $NESS_{22}(1)$  reduces to the Morisita index  $C_{22}$ . The  $NESS_{22}(m)$  index takes rare species into account when  $m$  is large, and so is less dependent on dominant species.

### 2.2 Sample Data and Estimation

In practice, we need to estimate the Morisita and NESS indices from sample data. Assume that a random sample of  $n_1$  individuals (Sample 1) is taken from Assemblage 1 and a random sample of  $n_2$  individuals (Sample 2) is taken from Assemblage 2. Individuals are sampled with replacement. Denote the *sample abundances* or sample frequencies by  $(X_{11}, X_{21}, \dots, X_{S1})$  and  $(X_{12}, X_{22}, \dots, X_{S2})$ , respectively. The basic model is that each of the two sets of frequencies follows a multinomial distribution with cell total being sample size and cell probabilities being  $\{p_{11}, p_{21}, \dots, p_{S1}\}$  (for Sample 1) or  $\{p_{12}, p_{22}, \dots, p_{S2}\}$  (for Sample 2). Under this model, it is implicitly assumed that sampled individuals are independent. Assume  $D_1$  and  $D_2$  species are, respectively, observed in Samples 1 and 2 and there are  $D_{12}$  observed shared species. Morisita's original estimate based on pairs of sample abundance data (Krebs, 1999) is

$$\hat{C}_{22} = \frac{2 \sum_{i=1}^S \frac{X_{i1} X_{i2}}{n_1 n_2}}{\sum_{i=1}^S \frac{X_{i1}(X_{i1} - 1)}{n_1(n_1 - 1)} + \sum_{i=1}^S \frac{X_{i2}(X_{i2} - 1)}{n_2(n_2 - 1)}}. \tag{5}$$

There are actually only  $D_{12}$  nonzero terms in the numerator of (5) and  $D_1$  and  $D_2$  nonzero terms in the first and second sums in the denominator. Any missing species in a sample (i.e., sample frequency = 0) does not contribute to any sum in (5). Hence the species index  $i$  in (5) can be summed from

1 to  $S$ , although  $S$  is unknown to us. Similar summation will be used throughout the article.

Although this estimator may exceed 1 in two highly similar assemblages, it has been found to be very stable and nearly independent of sample size (e.g., Wolda, 1981). Horn (1966) modified it to form the following Morisita–Horn estimator:

$$\tilde{C}_{22} = \frac{2 \sum_{i=1}^S \frac{X_{i1}}{n_1} \frac{X_{i2}}{n_2}}{\sum_{i=1}^S \left(\frac{X_{i1}}{n_1}\right)^2 + \sum_{i=1}^S \left(\frac{X_{i2}}{n_2}\right)^2},$$

which is always between 0 and 1. Our extensive simulation has shown that the Morisita–Horn estimator systematically underestimates similarity (see also Ricklefs and Lau, 1980), whereas the Morisita original estimator is nearly unbiased, though slightly overestimates it. Thus, pairwise results in our data analysis were calculated from equation (5). Both estimators are dominated by large sample frequencies and thus are likely to be resistant to undersampling, because the influential abundant species are always present in samples.

For estimating the NESS, we follow Grassle and Smith (1976) by expressing  $NESS_{22}(m)$  as a function of  $(1 - p_{i1})^m$ ,  $(1 - p_{i2})^m$ ,  $(1 - p_{i3})^m$ , and  $(1 - p_{i4})^m$ . An unbiased estimator for each term is, respectively,  $(n_1 - X_{i1})^{(m)}/n_1^{(m)}$ ,  $(n_1 - X_{i2})^{(2m)}/n_1^{(2m)}$ ,  $(n_2 - X_{i2})^{(m)}/n_2^{(m)}$ , and  $(n_2 - X_{i3})^{(2m)}/n_2^{(2m)}$ , where  $x^{(k)} = x(x - 1) \dots (x - k + 1)$ . Then a nearly unbiased estimator for the NESS index can be obtained.

Based on the multinomial model for the observed species frequencies, we have derived asymptotic variance formulas for estimators of the Morisita and NESS indices. However, the theoretical formulas become very complicated when there are more than two assemblages. Because our simulation studies have suggested that the resulting asymptotic variance is very close to the variance obtained by a bootstrap method (Efron, 1979; Efron and Tibshirani, 1993), we adopt the use of the latter that can be easily implemented for any number of assemblages.

Suppose we want to obtain a bootstrap variance for any estimator of  $C_{22}$  or  $NESS_{22}(m)$  based on two sets of sample frequencies  $(X_{11}, X_{21}, \dots, X_{S1})$  and  $(X_{12}, X_{22}, \dots, X_{S2})$ . The bootstrap procedures under the multinomial model and the assumption of independence between individuals are the following:

- (a) First generate two sets of bootstrap frequencies. One set of frequencies is generated from the multinomial distribution with cell total  $n_1$  (size of Sample 1) and cell probabilities  $(X_{11}/n_1, \dots, X_{S1}/n_1)$  (i.e., sample proportions of Sample 1). The other set is similarly generated from Sample 2.
- (b) Calculate the estimate (called a bootstrap estimate) of  $C_{22}$  or  $NESS_{22}(m)$  based on the two sets of generated frequencies.
- (c) Replicate the procedure in (a)  $B$  times and obtain  $B$  bootstrap estimates. The bootstrap variance estimator of the estimator is the sample variance of these  $B$  estimates. (Efron [1979] suggested that  $B$  should be at least 200.)

### 3. Extension to More Than Two Assemblages

#### 3.1 Three Assemblages

Because three assemblages are involved in our application, we first discuss this case and then generalize it. Assume that there are  $S$  species in the combined assemblage and let the species be indexed by  $1, 2, \dots, S$ . As in the case of two assemblages, we define three sets of probabilities of species discovery  $\{(p_{1r}, p_{2r}, \dots, p_{Sr}); r = 1, 2, 3\}$ , where  $p_{ir} \geq 0$ .

For comparing three assemblages, we need a profile of two similarity measures to characterize similarity. These two measures are constructed by varying the number of selections in Stage I. One measure corresponds to making two selections in Stage I whereas the other corresponds to making three selections. Consider first making two selections in Stage I; each selection could be Assemblage 1, Assemblage 2, or Assemblage 3 with equal probability. There are nine outcomes with probability  $1/9$  for each. Then in Stage II, one individual is randomly chosen from each of the two selected assemblages. As derived in (2a) and (2b), we have two conditional probabilities for the event  $Z$  that the two chosen individuals belong to the same species:

$$P(Z | \text{Stage I results in the same assemblages}) = \frac{(1/9) \left( \sum p_{i1}^2 + \sum p_{i2}^2 + \sum p_{i3}^2 \right)}{3(1/9)}, \quad (6a)$$

$$P(Z | \text{Stage I results in different assemblages}) = \frac{(1/9) \sum_{i=1}^S (2p_{i1}p_{i2} + 2p_{i1}p_{i3} + 2p_{i2}p_{i3})}{6(1/9)}. \quad (6b)$$

Note the second probability follows that there are six outcomes: (1, 2), (2, 1), (1, 3), (3, 1), (2, 3), and (3, 2). Therefore, we can formulate the following similarity measure as the ratio of the above two conditional probabilities:

$$C_{23} = \frac{\sum_{i=1}^S (p_{i1}p_{i2} + p_{i1}p_{i3} + p_{i2}p_{i3})}{\sum_{i=1}^S (p_{i1}^2 + p_{i2}^2 + p_{i3}^2)}. \quad (7)$$

Instead of two selections in Stage I, we now make three selections, so there are a total of 27 outcomes with probability  $1/27$  for each. Then in Stage II, we select one individual from each of the three selected assemblages. Consider the event  $Z$  that the *three* individuals belong to the same species. We have

$$P(Z | \text{Stage I results in the same assemblages}) = \frac{(1/27) \left( \sum p_{i1}^3 + \sum p_{i2}^3 + \sum p_{i3}^3 \right)}{3(1/27)}. \quad (8a)$$

$P(Z | \text{the three selections in Stage I include at least one assemblage that is different from the others})$

$$= \frac{(1/27) \sum_{i=1}^S \left( 3p_{i1}^2p_{i2} + 3p_{i1}p_{i2}^2 + 3p_{i1}^2p_{i3} + 3p_{i1}p_{i3}^2 + 3p_{i2}^2p_{i3} + 3p_{i2}p_{i3}^2 + 6p_{i1}p_{i2}p_{i3} \right)}{24(1/27)}. \quad (8b)$$

In evaluating the second conditional probability, conditioned on the event that at least one assemblage is different from the others, we have 24 possible outcomes: (1, 2, 1), (2, 1, 1), (1, 1, 2), (1, 2, 2), (2, 1, 2), (2, 2, 1), (1, 1, 3), (1, 3, 1), (3, 1, 1), (1, 3, 3), (3, 1, 3), (3, 3, 1), (2, 2, 3), (3, 2, 2), (2, 3, 2), (2, 3, 3), (3, 2, 3), (3, 3, 2), (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1). The ratio of (8b) to (8a) produces the following index:

$$C_{33} = \frac{\frac{1}{24} \sum_{i=1}^S \left( 3p_{i1}^2 p_{i2} + 3p_{i1} p_{i2}^2 + 3p_{i1}^2 p_{i3} + 3p_{i1} p_{i3}^2 + 3p_{i2}^2 p_{i3} + 3p_{i2} p_{i3}^2 + 6p_{i1} p_{i2} p_{i3} \right)}{\frac{1}{3} \sum_{i=1}^S [p_{i1}^3 + p_{i2}^3 + p_{i3}^3]} \quad (9)$$

The two measures  $C_{23}$  and  $C_{33}$  jointly characterize the similarity among three assemblages. The index  $C_{23}$  gives an overall measure of pairwise similarity. Note that the proportion of individuals observed in all assemblages is  $\sum p_{i1} p_{i2} p_{i3}$ . Thus, the index  $C_{33}$  takes the information shared by all assemblages into account.

Assume that a random sample of  $n_j$  individuals (Sample  $j$ ) is taken from Assemblage  $j$  for  $j = 1, 2$  and  $3$ . Denote the three sets of sample frequencies by  $\{(X_{1j}, X_{2j}, \dots, X_{Sj}); j = 1, 2, 3\}$ . A nearly unbiased estimator for  $C_{23}$  is

$$\hat{C}_{23} = \frac{\sum_{i=1}^S \left[ \frac{X_{i1} X_{i2}}{n_1 n_2} + \frac{X_{i1} X_{i3}}{n_1 n_3} + \frac{X_{i2} X_{i3}}{n_2 n_3} \right]}{\sum_{i=1}^S \left[ \frac{X_{i1}^{(2)}}{n_1^{(2)}} + \frac{X_{i2}^{(2)}}{n_2^{(2)}} + \frac{X_{i3}^{(2)}}{n_3^{(2)}} \right]}, \quad (10)$$

where  $x^{(k)} = x(x - 1) \dots (x - k + 1)$ . Similarly we have the following estimator for  $C_{33}$ :

$$\hat{C}_{33} = \frac{\frac{1}{24} \sum_{i=1}^S \left[ 3 \frac{X_{i1}^{(2)} X_{i2}}{n_1^{(2)} n_2} + 3 \frac{X_{i1} X_{i2}^{(2)}}{n_1 n_2^{(2)}} + 3 \frac{X_{i1}^{(2)} X_{i3}}{n_1^{(2)} n_3} + \dots + 6 \frac{X_{i1} X_{i2} X_{i3}}{n_1 n_2 n_3} \right]}{\frac{1}{2} \sum_{i=1}^S \left[ \frac{X_{i1}^{(3)}}{n_1^{(3)}} + \frac{X_{i2}^{(3)}}{n_2^{(3)}} + \frac{X_{i3}^{(3)}}{n_3^{(3)}} \right]} \quad (11)$$

In Stage II, if we select  $m$  individuals with replacement from each assemblage, then the two indices  $C_{23}$  and  $C_{33}$  are extended to the general class of  $NESS_{23}(m)$  and  $NESS_{33}(m)$ . That is,  $NESS_{23}(m)$  and  $NESS_{33}(m)$  are formulated by replacing  $p_{i1}$  by  $\mu_{i1}(m)$  and  $p_{i2}$  by  $\mu_{i2}(m)$  in (7) and (9). We suggest the use of a similar estimation procedure and bootstrap variance estimator as in the case of two assemblages.

### 3.2 General $N$ Assemblages

Assume that there are  $S$  species in the combined assemblage and define  $N$  sets of probabilities of species discovery  $\{(p_{1r}, p_{2r}, \dots, p_{Sr}); r = 1, \dots, N\}$ , where  $p_{ir} \geq 0$ . For comparing  $N(N > 2)$  assemblages, there is an obvious two-stage probabilistic generalization:

- (a) In Stage I, we sequentially select  $q$  assemblages from these  $N$  assemblages with equal probability. There are a total of  $N^q$  outcomes with probability  $(1/N^q)$  for each.

A profile of  $N - 1$  indices is obtained by varying  $q$  from 2 to  $N$ .

- (b) In Stage II, we randomly select an individual from each of the assemblages that were selected in Stage I.

Now consider the probability of event  $Z$  that the  $q$  individuals belong to the same species given Stage I results in the same assemblages. Arguments parallel to those used in obtaining (2a), (6a), and (8a) lead to

$$P(Z | \text{the } q \text{ selections of Stage I all result in the same assemblage}) = (1/N) \sum [p_{i1}^q + p_{i2}^q + \dots + p_{iN}^q]. \quad (12a)$$

To compute the conditional probability of  $Z$  given that at least one assemblage is different from the others, we note that there are  $N^q - N$  possible outcomes. Assume that in Stage I Assemblage  $j$  is selected  $r_j$  times,  $j = 1, 2, \dots, N$ , where  $0 \leq r_j < q$ ,  $\sum r_j = q$ . Then,

$$P(Z | \text{the } q \text{ selections in Stage I include at least one assemblage that is different from the others}) = \frac{1}{(N^q - N)} \sum_{i=1}^S \sum_{\substack{0 \leq r_j < q, j=1, 2, \dots, N \\ r_1 + r_2 + \dots + r_N = q}} \frac{q!}{r_1! r_2! \dots r_N!} p_{i1}^{r_1} p_{i2}^{r_2} \dots p_{iN}^{r_N}, \quad (12b)$$

$$= \frac{1}{(N^q - N)} \sum_{i=1}^S [(p_{i1} + p_{i2} + \dots + p_{iN})^q - (p_{i1}^q + p_{i2}^q + \dots + p_{iN}^q)]. \quad (12c)$$

Hence from (12a) and (12c), we extend the Morisita similarity index to  $N$  assemblages. The general formula is given by

$$C_{qN} = \frac{\frac{1}{(N^q - N)} \sum_{i=1}^S [(p_{i1} + p_{i2} + \dots + p_{iN})^q - (p_{i1}^q + p_{i2}^q + \dots + p_{iN}^q)]}{\frac{1}{N} \sum (p_{i1}^q + p_{i2}^q + \dots + p_{iN}^q)}, \quad (13)$$

which is always between 0 and 1. We have a profile of  $\{C_{qN}; q = 2, 3, \dots, N\}$  to describe similarity across  $N$  assemblages. The relative sensitivity  $C_{qN}$  to common or rare species is controlled by the value of  $q$ ; higher values of  $q$  strongly emphasize the dominant species.

Based on  $N$  sets of sample frequencies with sample size  $n_k$  from Assemblage  $k$ ,  $k = 1, 2, \dots, N$ , a nearly unbiased estimator of  $C_{qN}$  can be constructed by estimating the term  $p_{i1}^{r_1} p_{i2}^{r_2}, \dots, p_{iN}^{r_N}$  in equation (12b) by  $X_{i1}^{(r_1)} X_{i2}^{(r_2)}, \dots, X_{iN}^{(r_N)} / (n_1^{(r_1)} n_2^{(r_2)}, \dots, n_N^{(r_N)})$ . An approximate variance can be obtained by a similar bootstrap method as described in the case of the two assemblages. Replace  $p_{ij}$  by  $\mu_{ij}(m) = 1 - (1 - p_{ij})^m$  in (13), we then obtain the corresponding profile of  $\{NESS_{qN}(m); q = 2, 3, \dots, N\}$  for  $N$  assemblages.

### 4. Application to Forest Data

We apply the proposed indices to compare the similarities among three plant assemblages as described in the Introduction: the seedling size class, the sapling size class, and the tree size class within a tropical rain forest. The original data were collected in 2000 from four secondary and two primary

**Table 2**

*The observed species richness and sample size (in parentheses) in each forest*

Forest	Age	Trees	Saplings	Seedlings
LSUR second growth	15	12 (88)	68 (1917)	45 (421)
LSUR old growth	>200	37 (119)	101 (508)	47 (300)
LEP second growth	23	24 (169)	67 (1199)	47 (551)
LEP old growth	>200	43 (111)	102 (729)	69 (557)

forests in Costa Rica; see Chazdon et al. (1998, 2005, 2007) for detailed description of some of the forests. The actual forest areas range from about 5–20 ha. For comparison of primary and secondary forests, we used data from four forests: Lindero Sur (LSUR) primary forest, LSUR secondary forest, Lindero El Peje (LEP) primary forest, and LEP secondary forest. Each assemblage contains all individuals in that forest. The age of each forest in 2000 is shown in Table 2.

Complete census of all individuals in a forest is rarely possible for forest research. Simple random sampling in which all individuals have the same probabilities of being selected is not feasible either. Our purpose is to infer similarity among the assemblages based on data from plots. Our data were collected from a 1-ha plot in each forest (200 × 50 m). All canopy trees and canopy tree saplings were marked and measured for diameter within a 1-ha plot in each forest. Canopy tree seedlings were sampled in 144 (1 × 5 m) quadrats within the 1-ha plot, for a total area sampled of 0.072 ha. The observed species richness along with sample size for each size class is given in Table 2.

The 1-ha plot in each forest was selected to be representative of each forest area, which had homogenous land use. The plot was chosen to encompass the topographic variation within each area. All were on the same soil type (upland, residual volcanic soils) and were at least 50 m from the trail (to avoid edge effects). Thus, we treat our data as a representative multinomial sample from each assemblage for illustrative purposes, although the recorded individuals may be dependent. Our analyses focus on comparing three size classes within a secondary forest and within a primary forest. Species of seedlings and saplings that do not reach the canopy (subcanopy or understory species) were excluded from the data. All numerical results are shown in Table 3.

In the upper half of Table 3, we present the traditional pairwise comparisons between any two size classes. We calculated  $NESS_{22}(m)$  for  $m = 1$  to  $m = 10$ , but only estimates for two values ( $m = 5$  and 10) along with the Morisita index  $C_{22}$  are presented in Table 3. Other values of  $m$  produce similar conclusions. The standard error (SE) for each estimate was obtained from a bootstrap method using 200 generated replications.

In the LSUR plot, the estimated Morisita index  $C_{22}$  shows that the primary and secondary forests have comparable similarity between saplings and trees; the secondary forest has a much higher similarity between saplings and seedlings than primary forest; whereas the primary forest has a much higher similarity between trees and seedlings than secondary forest. As described in the Introduction, the Morisita index is mainly sensitive to dominant species, thus the above finding emphasizes the relationships among the more abundant species in the assemblages. On the other hand, both

**Table 3**

*Various similarity indices/estimates for pairwise and simultaneous comparison (bootstrap SE in parentheses based on 200 replications)*

Index	LSUR plot		LEP plot	
	Secondary forest	Primary forest	Secondary forest	Primary forest
Pairwise comparison				
Morisita $C_{22}$				
Trees vs. saplings	0.20 (0.04)	0.18 (0.03)	0.16 (0.03)	0.48 (0.07)
Trees vs. seedlings	0.20 (0.05)	0.92 (0.03)	0.29 (0.05)	0.79 (0.04)
Saplings vs. seedlings	0.83 (0.02)	0.32 (0.04)	0.81 (0.03)	0.47 (0.04)
NESS <sub>22</sub> (5)				
Trees vs. saplings	0.30 (0.05)	0.36 (0.04)	0.20 (0.02)	0.65 (0.05)
Trees vs. seedlings	0.32 (0.07)	0.82 (0.03)	0.32 (0.03)	0.70 (0.03)
Saplings vs. seedlings	0.85 (0.02)	0.53 (0.04)	0.85 (0.02)	0.56 (0.04)
NESS <sub>22</sub> (10)				
Trees vs. saplings	0.38 (0.06)	0.46 (0.05)	0.23 (0.02)	0.74 (0.04)
Trees vs. seedlings	0.40 (0.07)	0.66 (0.04)	0.33 (0.02)	0.63 (0.03)
Saplings vs. seedlings	0.88 (0.02)	0.68 (0.04)	0.87 (0.02)	0.64 (0.03)
Simultaneous comparison				
Morisita $C_{23}$	0.34 (0.04)	0.55 (0.02)	0.37 (0.03)	0.61 (0.02)
Morisita $C_{33}$	0.19 (0.04)	0.42 (0.03)	0.24 (0.04)	0.50 (0.04)
NESS <sub>23</sub> (5)	0.48 (0.04)	0.60 (0.03)	0.44 (0.02)	0.64 (0.02)
NESS <sub>33</sub> (5)	0.34 (0.04)	0.49 (0.03)	0.33 (0.03)	0.54 (0.03)
NESS <sub>23</sub> (10)	0.55 (0.04)	0.61 (0.03)	0.48 (0.01)	0.67 (0.02)
NESS <sub>33</sub> (10)	0.44 (0.05)	0.52 (0.04)	0.39 (0.02)	0.58 (0.03)

NESS<sub>22</sub>(5) and NESS<sub>22</sub>(10) consider the less abundant species, and their estimates indicate that the primary forest has a higher similarity between trees and saplings as well as between trees and seedlings, but a lower similarity between saplings and seedlings. The same pattern holds for the LEP plot for all three indices.

However, the three pairwise indices are not sufficient to characterize the similarity among all three assemblages in each forest type. In the lower half of Table 3, we show estimates of three indices for simultaneously comparing trees, saplings, and seedlings. The three indices include the three-assemblage Morisita index that consists of the profile of two indices ( $C_{23}$ ,  $C_{33}$ ) and two NESS( $m$ ) indices,  $m = 5$  and 10; each consists of two indices (NESS<sub>23</sub>( $m$ ), NESS<sub>33</sub>( $m$ )). Based on a bootstrap method for 200 replications, all SEs in the lower half of Table 3 are in the range of 2–5%.

Intuitively, we can regard  $C_{23}$  as an overall measure of the three values of  $C_{22}$  computed from three pair comparisons, whereas  $C_{33}$  is the more global measure taking into account the proportion of individuals that belong to species shared among all three assemblages. For example, in the LSUR secondary plot, the estimate of  $C_{22}$  for three pair comparisons is 0.20 (trees vs. saplings), 0.20 (trees vs. seedlings), and 0.83 (saplings vs. seedlings) with an average of 0.41. The estimated  $C_{23}$  is 0.34, but the global measure  $C_{33}$  is a much lower 0.19. This means that in the secondary forest, relatively few of the individuals belong to the species shared by all three assemblages. A similar interpretation applies to the profile of (NESS<sub>23</sub>( $m$ ), NESS<sub>33</sub>( $m$ )). For example, in the LSUR secondary plot, the average of three pairwise estimates of NESS<sub>22</sub>(5) is 0.49, which closely matches the estimated NESS<sub>23</sub>(5) of 0.48, but the global NESS<sub>33</sub>(5) is only 0.34.

Table 3 shows that for both plots, no matter which index is used for simultaneous comparisons, the primary forest has consistently higher similarity among the three size classes than the secondary forest. The bootstrap SE can be used to perform a statistical test whether the difference is significant or not. For example, in the LSUR primary plot, the estimated  $C_{23}$  is 0.55 (SE = 0.02). The corresponding estimate of  $C_{23}$  for the LSUR secondary forest is 0.34 (SE = 0.04). The observed difference is  $0.55 - 0.34 = 0.21$  with an estimated SE =  $[(0.02)^2 + (0.04)^2]^{0.5} = 0.045$  assuming the two estimates of  $C_{23}$  in two different forests are independent. The standardized test statistic becomes  $0.21/0.045 = 4.67$ , which is highly significant at the 5% level under a normal test. Here the asymptotic normality for the observed difference is approximately valid because the sample sizes (in Table 2) are sufficiently large. Thus, we can conclude that the size classes in the LSUR primary forest have a significantly higher similarity  $C_{23}$  than those of the secondary forest. A similar finding is valid for  $C_{33}$  as well. In this example, the global  $C_{33}$  is more sensitive to the difference between primary and secondary forest than the pairwise  $C_{23}$ ; the higher  $C_{33}$  in primary forest means that relatively more individuals (for dominant species) are shared by all size classes in primary forest compared to secondary forest. A similar conclusion applies to the LEP plot. The profile of (NESS<sub>23</sub>( $m$ ), NESS<sub>33</sub>( $m$ )) tends to increase with  $m$ , but the difference between primary and secondary forests diminishes as  $m$  is increased. That is, when less dominant species are also included, the NESS indices indi-

cate that the difference between primary forest and secondary forest similarity becomes less pronounced. This implies that species shared among size classes in secondary forests tend to be rare or infrequent, whereas they tend to be more common in primary forests.

## 5. Concluding Remarks and Discussion

For simultaneously comparing species composition of  $N > 2$  communities, we have proposed a general two-stage probabilistic framework to construct a profile of similarity indices  $\{C_{qN}; q = 2, \dots, N\}$  given in equation (13). In Stage I, we select  $q$  out of the  $N$  communities with equal probability. In Stage II, we select an individual from each of the communities that are selected in Stage I. Our proposed similarity in equation (13) is the ratio of two conditional probabilities:  $C_{qN} = P(Z | A)/P(Z | A^c)$ , where  $Z$  denotes the event that the  $q$  selected individuals in Stage II belong to the same species,  $A$  denotes that the  $q$  selected assemblages in Stage I are the same, and  $A^c$  denotes the complement of  $A$ . This approach extends the Morisita index to multiple communities. Instead of selecting an individual, if we select  $m$  individuals with replacement in Stage II, we then similarly extend the NESS index to multiple communities.

Based on species frequencies data, we have proposed nearly unbiased estimators of the generalized Morisita and NESS indices as well as their bootstrap variances. The basic model assumptions for the validity of the proposed inference procedures are: (1) sampled individuals are independent; and (2) sample frequencies follow a multinomial distribution with cell probabilities being equal to the species abundances (namely, data are representative of the true community). Although the above assumptions may not be fully satisfied for plot sampling taken from rain forests, the data analysis in Section 4 provides an illustrative example demonstrating methodological innovation. When individuals are dependent, the bootstrap variance estimator may underestimate the true variance. As pointed out by an associate editor, a proper sampling procedure for applications to forest data would be to randomly select many small plots, and regard each plot (rather than each individual plot) as a sampling unit. However, such sampling procedures are generally not feasible in forest studies.

The generalized Morisita and NESS indices, like the original Morisita and NESS indices, are comparisons of species compositions without regard to the total numbers of individuals in each community; that is, each community is given equal weight in Stage I. For the case of unequal weights, our two-stage approach can be similarly constructed. However, a proper weight should be proportional to the total number of individuals, which is generally unknown. The inference for unequal probability case requires further research.

Other approaches to the problem of multiple-community similarity have used homogeneity measures based on the ratio of alpha diversity indices over gamma diversity indices (e.g., Lande, 1996), or the ratio of the numbers equivalents (Hill numbers of order  $q$ ) of these indices (MacArthur, 1965; Olszewski, 2004; Jost, 2006). For equally weighted communities this latter ratio is the reciprocal of  ${}^qD_\beta$ , the true beta diversity of order  $q$  (Jost, 2007) and this ratio can be transformed onto the unit interval to obtain similarity indices that range from zero to one (Jost, 2006, 2007). Several

transformations are possible, and each results in a family of indices illuminating a different aspect of similarity. Jost (2006) gave one family of similarity indices for multiple communities, a set of “shared diversity” measures:  $S_{qN} = (1/q D_\beta - 1/N)/(1 - 1/N)$ . Our index  $C_{qN}$  extends Jost’s second family of measures, the set of “overlap” measures, to the case of  $N$  communities:  $C_{qN} = [(1/q D_\beta)^{q-1} - (1/N)^{q-1}]/[1 - (1/N)^{q-1}]$ . Because this is a monotonic transformation of beta diversity of order  $q$ , conclusions based on this index will always be consistent with conclusions based on diversity indices of order  $q$  (Jost, 2007). For example,  $C_{2N}$  is consistent with all Simpson measures of diversity, and with the Hurlbert–Smith–Grassle index  $S(m)$  with  $m = 2$  (Smith and Grassle, 1977).

When  $N = 2$ ,  $C_{qN}$  is a “true overlap measure” in the sense of Wolda (1981; proof in Jost, 2006). This means that when there are  $S$  equally common species in each community, and  $D$  species are shared between communities, the measure gives the proportion of overlap,  $D/S$ . It seems reasonable to generalize Wolda’s definition to the case of  $N$  communities as follows: If  $N$  communities each have  $S$  equally common species, and if exactly  $D$  species are shared by all of them, and the remaining species are not shared by any communities, then a true overlap measure should give  $D/S$  for all orders of  $q$ . In this sense, the new measure  $C_{qN}$  is a true overlap measure for  $N$  communities. We thus have an intuitive interpretation of  $C_{qN}$  as follows. If we have  $N$  communities, then this set is equivalent in terms of diversity and similarity to a set of  $N$  communities with the same alpha and gamma as the original set but consisting of all equally large communities with all equally common species. The measure  $C_{qN}$  can be interpreted as the percent of species overlap (as defined in our extension of Wolda’s definition) in this set of equivalent communities.

Our two-stage probabilistic interpretation applies only to the case of an integer  $q \geq 2$ . If we allow the order  $q$  to be any nonnegative real numbers, then the index  $C_{qN}$  includes other well-known overlap measures besides the Morisita index. For example,  $C_{q2}$  becomes the Horn (1966) index of overlap as  $q$  tends to 1, and  $C_{02}$  is the classic Sorensen index. Our  $C_{0N}$  may be considered the multiple-community generalization of the Sorensen index. For orders  $q \geq 2$ , rare species have negligible effect on the estimates, and thus accurate estimation is feasible. We remark that the estimation for  $q < 2$  becomes statistically difficult and requires further research because the effect of rare species may be large. We also remark that with slight modification in model formulation and arguments, our procedure can be extended to deal with replicated incidence data.

## 6. Supplementary Materials

The forest data analyzed in Section 4 are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

## ACKNOWLEDGEMENTS

This work was supported by the Taiwan National Science Council Contract NSC94-2118-M007-006 and 95-2118-M007-003 to A. Chao and Y.-H. Jiang. Vegetation studies in secondary and primary forests of Costa Rica were supported by grant NSF DEB 0424767 to R. Chazdon. We thank an

associate editor for carefully reading an earlier version and providing insightful comments and suggestions.

## REFERENCES

- Arita, H. T. and Rodríguez, P. (2002). Geographic range, turnover rate and the scaling of species diversity. *Ecography* **25**, 541–550.
- 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.
- Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T.-J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62**, 361–371.
- 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., Brenes, A. R., and Alvarado, B. V. (2005). Effects of climate and stand age on annual tree dynamics in tropical second-growth rain forests. *Ecology* **86**, 1808–1815.
- Chazdon, R. L., Letcher, S. G., van Breugel, M., Martinez-Ramos, M., Bongers, F., and Finegan, B. (2007). Rates of change in tree communities of secondary Neotropical forests following major disturbance. *Philosophical Transactions of the Royal Society of London, Series B* **362**, 273–289.
- Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London, Series B* **345**, 101–118.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.
- Gower, J. C. (1985). Measures of similarity, dissimilarity and distance. In *Encyclopedia of Statistical Sciences*, Vol. 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.
- Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology* **54**, 427–431.
- Horn, H. S. (1966). Measurement of “overlap” in comparative ecological studies. *American Naturalist* **100**, 419–424.
- Jost, L. (2006). Entropy and diversity. *Oikos* **113**, 363–375.
- Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology* **88**, 2427–2439.
- Krebs, C. J. (1999). *Ecological Methodology*, 2nd edition. New York: Harper & Row.

- Lande, R. (1996). Statistics and partitioning of species diversity and similarity among multiple communities. *Oikos* **76**, 5–13.
- Ludwig, J. A. and Reynolds, J. F. (1988). *Statistical Ecology: A Primer on Methods and Computing*. New York: Wiley.
- MacArthur, R. (1965). Patterns of species diversity. *Biological Review* **40**, 510–533.
- Magurran, A. E. (2004). *Measuring Biological Diversity*. Oxford, U.K.: Blackwell.
- Olszewski, T. (2004). A unified mathematical framework for the measurement of richness and evenness within and among communities. *Oikos* **104**, 377–387.
- Plotkin, J. B. and Muller-Landau, H. C. (2002). Sampling the species composition of a landscape. *Ecology* **83**, 3344–3356.
- Ricklefs, R. E. and Lau, M. (1980). Bias and dispersion of overlap indices: Results of some Monte Carlo simulations. *Ecology* **61**, 1019–1024.
- Simpson, E. H. (1949). Measurement of diversity. *Nature* **163**, 688.
- Smith, W. and Grassle, J. F. (1977). Sampling properties of family diversity measures. *Biometrics* **33**, 283–292.
- 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 non-parametric estimator of species overlap. *Biometrics* **57**, 743–749.

Received May 2007. Revised December 2007.

Accepted December 2007.