A new statistical approach for assessing similarity of species composition with incidence and abundance data

Anne Chao,1 Robin L. Chazdon,2* Robert K. Colwell2 and Tsung-Jen Shen1

1Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan
2Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA
*Correspondence: E-mail: chazdon@uconn.edu

Abstract

The classic Jaccard and Sørensen indices of compositional similarity (and other indices that depend upon the same variables) are notoriously sensitive to sample size, especially for assemblages with numerous rare species. Further, because these indices are based solely on presence–absence data, accurate estimators for them are unattainable. We provide a probabilistic derivation for the classic, incidence-based forms of these indices and extend this approach to formulate new Jaccard-type or Sørensen-type indices based on species abundance data. We then propose estimators for these indices that include the effect of unseen shared species, based on either (replicated) incidence- or abundance-based sample data. In sampling simulations, these new estimators prove to be considerably less biased than classic indices when a substantial proportion of species are missing from samples. Based on species-rich empirical datasets, we show how incorporating the effect of unseen shared species not only increases accuracy but also can change the interpretation of results.

Keywords

Abundance data, beta diversity, biodiversity, complementarity, incidence data, shared species, similarity estimators, similarity index, species overlap, succession.

oldest and most widely used similarity indices for assessing compositional similarity of assemblages (sometimes called 'species overlap') and hence, its complement, dissimilarity. Both measures are based on the presence/absence of species in paired assemblages and are simple to compute (Magurran 2004). Many other similarity indices exist that are based on the same information: the number of species shared by two samples and the number of species unique to each of them (Legendre & Legendre 1998), and new indices continue to appear (e.g. Lennon et al. 2001). A modified version of the Sørensen index was developed by Bray & Curtis (1957), based on abundance data (also known as the Sørensen abundance index; Magurran 2004), and a large number of other abundance-based indices have been developed (Legendre & Legendre 1998), including the widely applied Morisita–Horn index (Magurran 2004).

Despite their wide application in ecological studies, the classic Jaccard and Sørensen indices, when computed for sample data, perform poorly as measures of similarity between diverse assemblages that include a substantial fraction of rare species (Wolda 1981; Colwell & Coddington 1994; Plotkin & Muller-Landau 2002), because the sample data are (usually wrongly) assumed to be true and complete representations of assemblage composition. Indeed, with very few exceptions (e.g. Grasse & Smith 1976; MacKenzie et al. 2004), nearly all existing approaches to measuring similarity make this assumption. In general, as we will show with simulations, these measures are likely to severely underestimate true similarity between two (genuinely similar) assemblages that contain numerous rare species. Because many species are missed by the samples, the rare species that appear in one sample are likely to be different than the rare species that show up in the other sample, even if all are actually present in both assemblages. Similar problems arise from comparing two samples of substantially different size: simply because it contains fewer individuals or sampling units, the smaller sample may lack species that appear in the larger sample. In short, the underestimation of similarity occurs because of the failure to account for unseen shared species.

In principle, overestimation of similarity can also occur when comparing undersampled, high-dominance communities in which the common species are widespread and rare ones tend to be locally endemic. In this case, two samples might yield the same few common species, but fail to reveal rare species that would differentiate the assemblages in larger samples (Colwell & Coddington 1994; Ruokolainen & Tuomisto 2002 discuss a possible example). In nearly all cases we have examined quantitatively, however, rarity (either in nature or because of small sample size) increases the chance that a species will be spuriously absent from one sample but not from the other, thus negatively biasing similarity indices. [Fisher (1999, Fig. 8) comes to the same conclusion for several datasets, based on rarefaction tests.] Moreover, for the new indices we present here, it can be shown theoretically that sampling bias, when present, is always negative. [The authors demonstrate the expected negative bias mathematically (A. Chao, R. L. Chazdon, R. K. Colwell & T.-J. Shen, unpublished data); it can be proved for any abundance models given in Magurran (2004) and Plotkin & Muller-Landau (2002).]

Recently, interest has intensified in the development and evaluation of indices to measure beta diversity, or turnover rate, of species assemblages (Duijvenvoorden 1995; Lennon et al. 2001; Arita &Rodriguez 2002, 2004; Condit et al. 2002; Plotkin & Muller-Landau 2002; Koleff et al. 2003; Rodriguez & Arita 2004), underscoring the need for robust statistical estimators for inferring compositional similarity from sample data. Increasing species turnover (decreasing similarity) with increasing distance between sites may reflect spatial patterns of dispersal or may be driven by increasing environmental heterogeneity at greater scales (Harte et al. 1999; Hubbell 2001; Balvanera et al. 2002; Chave & Leigh 2002; Condit et al. 2002; Duijvenvoorden et al. 2002; Ruokolainen & Tuomisto 2002; Rodriguez & Arita 2004; Valencia et al. 2004). Unfortunately, most indices of beta diversity rely on the same information as the classic Jaccard and Sørensen indices and share the limitations discussed above.

With this problem in mind, Plotkin & 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. Condit et al. (2002) adopt an approach to measuring beta diversity using Leigh et al.’s (1993) ‘codominance’ index $F$, the probability that two individuals chosen randomly from each of two assemblages are the same species. Although this measure is based on abundance data, $F$, itself, is not a statistically valid index of similarity. For two identical assemblages with many species, $F$ tends to 0. Moreover, it is possible for any two identical assemblages to have any value of $F$ from 0 to 1, depending on how many species are present and patterns of relative abundance. It is possible, however, to normalize $F$ to produce a valid similarity index. Chave & Leigh (2002) point out that the Morisita–Horn index is a normalized version of $F$.

We begin by developing a new, probabilistic approach for the classic Jaccard and Sørensen incidence-based indices. We then extend this approach to formulate Jaccard-type and Sørensen-type indices that consider species abundances. In contrast to Plotkin & Muller-Landau (2002), we adopt a non-parametric approach that does not require any assumptions about species abundance distributions. We then propose a method to estimate both incidence-based and abundance-based Jaccard and Sørensen indices from sample data, incorporating the effect of unseen shared species.
We then carry out sampling simulations with empirical data sets to assess the relative performance of the classic Jaccard and Sørensen indices; their new, abundance-based Jaccard and Sørensen counterparts; and the corresponding Jaccard and Sørensen estimators. We show that incorporating the effect of unseen species substantially reduces the sample-size bias of these estimators and improves their suitability for inferring similarity (or its complement, dissimilarity) between hyper-diverse assemblages for which a large proportion of species are missing from samples. Finally, we illustrate an application of the new abundance-based Jaccard index and the Jaccard abundance-based estimator, using data from a successional study of tree, sapling and seedling abundance of canopy species. Based on data sets for rich, tropical insect and plant assemblages, we show how incorporating the effect of unseen shared species not only increases accuracy, but also can change the interpretation of results.

DEVELOPING THE NEW INDICES AND ESTIMATORS

The classic Sørensen and Jaccard similarity indices

The classic Sørensen and Jaccard indices depend on three simple incidence counts: the number of species shared by two assemblages and the number of species unique to each of them. It has become traditional to refer to these counts as \( A, B \) and \( C \), respectively (Table 1). The classic Jaccard and Sørensen indices for incidence counts are then

\[
J_{\text{clas}} = \frac{A}{A + B + C} \quad (1)
\]

and

\[
I_{\text{clas}} = \frac{2A}{2A + B + C} \quad (2)
\]

We use \( L \) for the Sørensen index to avoid confusion with \( S \) for species.) There is a close, monotonic relation between the two indices: \( I_{\text{clas}} = 2J_{\text{clas}}/(J_{\text{clas}} + 1) \) and \( J_{\text{clas}} = 1/(2/I_{\text{clas}} - 1) \).

Assume that there are \( S_1 \) species in Assemblage 1 and \( S_2 \) species in Assemblage 2. Let the number of shared species be \( S_{12} \). Then, the incidence counts \( A, B, C \) in Table 1 correspond to the \( A = S_{12}, B = S_1 - S_{12} \) and \( C = S_2 - S_{12} \). Substituting these expressions in eqns 1 and 2, we have an alternate way to write the classic indices that will be required for the next steps in developing the new indices:

\[
J_{\text{clas}} = \frac{A}{A + B + C} = \frac{S_{12}}{S_1 + S_2 - S_{12}} \quad (3)
\]

and

\[
I_{\text{clas}} = \frac{2A}{2A + B + C} = \frac{2S_{12}}{S_1 + S_2}. \quad (4)
\]

Table 1 Species classification counts used in the classic indices

<table>
<thead>
<tr>
<th>Assemblage 2</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>( A )</td>
<td>( B )</td>
</tr>
<tr>
<td>Absent</td>
<td>( C )</td>
<td></td>
</tr>
</tbody>
</table>

A probabilistic approach to the classic Jaccard and Sørensen indices

The classic Jaccard and Sørensen indices consider only the presence or absence (incidence) of species. Two pairs of assemblages, one pair sharing abundant species but not rare ones and the other pair sharing rare species, but not common ones, will yield the same index value. From the point of view of overall assemblage similarity, taking similarity of assemblage composition to the level of individuals often makes more sense (Magurran 2004). Our next objective is to extend the incidence indices to take account of the relative abundance of species, a prerequisite for developing index estimators for sampling data that take account of unseen rare species.

We must first provide a probabilistic derivation of the classic Jaccard and Sørensen incidence indices. Suppose we randomly select a \( S \) species from Assemblage 1 and a \( S \) 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 shown graphically in Fig. 1 and specified in Table 2.

Although the probabilities in Table 2 are not counts, they can be thought of as 'normalized counts,' because they sum to unity. Substituting these probabilities into eqns 1 and 2, we have

\[
J_{\text{clas}} = \frac{A}{A + B + C} = \frac{S_{12}}{S_1 + S_2 - S_{12}} \quad (3)
\]

which is exactly eqn 3. Likewise, we have

\[
I_{\text{clas}} = \frac{2A}{2A + B + C} = \frac{2S_{12}}{S_1 + S_2} \quad (4)
\]

which is the same as eqn 4.
Figure 1 A graphical representation of the meaning of shared species for two assemblages. Assemblage 1 (a1) is grey, Assemblage 2 (a2) is white. The grey dot represents a species selected at random from Assemblage 1 and the white dot represents a species selected at random from Assemblage 2. Case 1 is the only case in which both species are shared species (but not necessarily the same species). In Case 2, the species chosen at random from Assemblage 1 is a shared species, but the species chosen from Assemblage 2 is not shared with Assemblage 1. The reverse is true for Case 3. In Case 4, neither of the chosen species is a shared species. These patterns are described mathematically in Table 2.

Table 2 Probabilistic derivation of species counts for the classic indices

<table>
<thead>
<tr>
<th>Select any species from Assemblage 2</th>
<th>Shared</th>
<th>Non-shared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select any species from Assemblage 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shared</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case 1)</td>
<td>( A = \frac{\pi_1 \pi_2}{\pi_1 + \pi_2} )</td>
<td>( B = \frac{\pi_1}{\pi_1 + \pi_2} (1 - \frac{\pi_2}{\pi_1}) )</td>
</tr>
<tr>
<td>(Case 3)</td>
<td>( C = \left(1 - \frac{\pi_1}{\pi_2} \right) \frac{\pi_2}{\pi_1} )  (1 - \frac{\pi_1}{\pi_2} )(1 - \frac{\pi_2}{\pi_1}) )</td>
<td></td>
</tr>
</tbody>
</table>

It might appear that we have made no progress, but this probabilistic approach lays the groundwork for developing abundance-based indices, which in turn allow for the estimation of indices that take into account the effect of unseen shared species. Note that, using this approach, we can also calculate the chance that both randomly chosen species are non-shared species (Case 4 as shown in Fig. 1 and Table 2). However, the basic concept for the Jaccard and Sørensen indices is based only on information for the other three cells (Cases 1–3).

Extending the probabilistic approach to abundance-based indices

Let the probabilities of species discovery (which depend primarily on relative abundance, assuming random mixing and equivalent detectability) in Assemblages 1 and 2 be denoted, respectively, by \( (p_1, p_2, \ldots, p_n) \) and \( (\pi_1, \pi_2, \ldots, \pi_n) \), where \( p_i > 0, \pi_i > 0 \) and \( \sum_{i=1}^n p_i = \sum_{i=1}^n \pi_i = 1 \). 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 from the previous section, we randomly select one individual from Assemblage 1 and one individual from Assemblage 2. For each individual of the pair, note whether it belongs to a shared species or not.

We now derive the general formulas for the abundance-based versions of the Jaccard and Sørensen indices. 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 of individuals belonging to the shared species, \( U = p_1 + p_2 + \cdots + p_{s_{12}} \). Likewise in Assemblage 2, let \( V \) denote the total relative abundances of individuals belonging to shared species, \( V = \pi_1 + \pi_2 + \cdots + \pi_{s_{12}} \). Table 3 shows the probabilities that two individuals, one from each assemblage, represent each of the usual four categories.

Based on eqns 1 and 2 for the three probabilities \( (A, B \) and \( C \) in Table 3), we obtain the following abundance-based indices in terms of \( U \) and \( V \):

\[
J_{\text{abd}} = \frac{A}{A + B + C} = \frac{UV}{U + V - UV} \quad (5)
\]

Table 3 Probabilities for individual-based species counts

<table>
<thead>
<tr>
<th>Select any individual from Assemblage 2</th>
<th>Shared</th>
<th>Non-shared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select any individual from Assemblage 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shared</td>
<td>( A = UV )</td>
<td>( B = U(1 - V) )</td>
</tr>
<tr>
<td>Non-shared</td>
<td>( C = (1 - U)V )</td>
<td>( D = (1 - U)(1 - V) )</td>
</tr>
</tbody>
</table>
and

\[ I_{\text{abd}} = \frac{2A}{2A + B + C} = \frac{2UV}{U + V} \]  \hspace{1cm} (6)

As \( U \) and \( V \) represent the total abundances of the shared species in Assemblages 1 and 2, respectively, we see that both indices reach 1 for identical assemblages and tend to 0 for disjoint assemblages. In the latter case, for example, \( I_{\text{abd}} = 2/[1/(1/U) + (1/V)] \) tends to 0 as both \( U \) and \( V \) approach 0.

**Estimation of the abundance-based indices from sample data**

Up to now, we have considered only the species and individuals observed in two assemblages. Both the classic Jaccard and Sørensen and the new, abundance-based versions assume full and complete knowledge of the two assemblages being contrasted. In practice, we need to estimate similarity indices from sample data, the task that we turn to now. Our approach is non-parametric in the sense that we do not need to postulate any particular species abundance distribution to derive the estimators, which are therefore valid under many statistical abundance models (e.g. log-normal, broken stick, gamma, etc.). The derivation does assume that the number of species is finite so that species discovery probabilities are bounded below. [The authors show that the estimators are valid under many of the statistical abundance models (A. Chao, R. L. Chazdon, R. K. Colwell & T.-J. Shen, unpublished data) (e.g. log-normal, exponential, gamma, negative binomial, Zipf–Mandelbrot, broken-stick models, etc.) that appear in Magurran (2004, Table 2.1) or in Plotkin & Muller-Landau (2002, Table 1).]

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. 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. (Note that if a species is missing from a sample, \( X_i \) or \( Y_j \) will equal zero.) Thus, the pair of frequencies for the \( S_{12} \) species truly shared by the two assemblages 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 available are actually observed in both samples, and their frequencies are the first \( D_{12} \) pairs. Thus, an additional \( S_{12} - D_{12} \) species are shared by the two assemblages, but absent from one or both of the samples. The greater the frequencies of rare, shared species observed in one of the two samples, the more probable it is that additional shared species are present in both assemblages, but are absent from one or both samples. We refer to these as unseen shared species.

To incorporate the effect of unseen shared species on the probabilities of Table 3, we use the frequencies of observed rare, shared species to estimate an appropriate adjustment term for \( U \) and \( V \) to account for unseen shared species. We first define the indicator function \( I(\text{expression}) \) such that \( I = 1 \) if ‘expression’ is true and \( I = 0 \) if ‘expression’ is false. Let \( f_{+1} = \sum_{i=1}^{D_{12}} I(X_i = 1, Y_i \geq 1) \) be the observed number of shared species that are singletons \((X_i = 1)\) in Sample 1 (these species must be present in Sample 2, but may have any abundance). Now, let \( f_{+2} \) be the observed number of shared species that are doubletons \((X_i = 2)\) in Sample 1. Similarly, we define \( f_{-1} \) and \( f_{-2} \) to be the observed number of shared species that are, respectively, singletons \((Y_i = 1)\) and doubletons \((Y_i = 2)\) in Sample 2.

Then the proposed estimator for \( U \) is

\[
\hat{U} = \sum_{i=1}^{D_{12}} \frac{X_i}{n} + \frac{(m - 1) f_{+1} + D_{12}}{m} \sum_{i=1}^{D_{12}} \frac{X_i}{n} I(Y_i = 1) \]  \hspace{1cm} (7)

Notice that the first term in the right-hand side of eqn 7 denotes the observed total of frequencies associated with the observed shared species; the second term accounts for the estimated effect of unseen shared species. Similarly, we have

\[
\hat{V} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m} + \frac{(n - 1) f_{+1} + D_{12}}{n} \sum_{i=1}^{D_{12}} \frac{Y_i}{m} I(X_i = 1) \]  \hspace{1cm} (8)

When \( f_{+2} = 0 \) or \( f_{+2} = 0 \), replace \( f_{+2} \) and \( f_{-2} \) in the denominators by \( f_{+2} + 1 \) or \( f_{+2} + 1 \), respectively. If the value of \( \hat{U} \) or \( \hat{V} \) is greater than 1 (which rarely happens), then it is replaced by 1. Our proposed abundance-based Jaccard and Sørensen estimators are

\[
\hat{J}_{\text{abd}} = \frac{\hat{U} \hat{V}}{\hat{U} + \hat{V} - \hat{U} \hat{V}} \]  \hspace{1cm} (9)

and

\[
\hat{L}_{\text{abd}} = \frac{2\hat{U} \hat{V}}{\hat{U} + \hat{V}} \]  \hspace{1cm} (10)

The variances for these two estimators can be derived by a bootstrap method. (The complete derivation of eqns 7 and 8 and details on the bootstrap procedure for computing variance estimators for eqns 9 and 10 are available upon request from the first author.)

**Estimation of similarity indices from incidence frequencies**

Because information about the frequencies and identities of rare species provides the critical information for adjusting similarity indices to account for the effect of unseen shared species, a simple pair of lists of the species present in two assemblages (incidence data) cannot be used, even in principle, to adjust similarity indices for the effect of unseen species. On the other hand, the estimation-based approach...
can be extended to replicated incidence (presence–absence) data.

Suppose we take a set of \( w \) replicated incidence samples from Assemblage \( X \) and a set of \( z \) replicated incidence samples from Assemblage \( Y \). For both sets of samples combined, there are \( \lambda \) species. The number of samples in which a species is found in Assemblage \( X \) or \( Y \) is the frequency for that species in that sample set. The frequencies for species \( i \) are thus defined as

\[
X_i = \sum_{j=1}^{w} x_{ij} \quad \text{and} \quad Y_i = \sum_{j=1}^{z} y_{ij},
\]

where \( x_{ij} \) and \( y_{ij} \) represent the presence (1) or absence (0) of species \( i \) in sample \( j \).

Note that \( X_i \) or \( Y_i \) will be zero for some species, unless all species are shared and observed.

Under the assumption that replicate incidence samples are statistically homogeneous (within each assemblage), the chance of a species being present in a particular sample is proportional to its relative abundance in the assemblage, and the frequency vectors \( X_i \) or \( Y_i \) are thus statistical proxies for the relative abundance of species in Assemblages \( X \) and \( Y \) (e.g. Chao 2004; Colwell et al. 2004). Thus, with minor changes, eqns 7 and 8 can be used to compute adjusted probabilities that a randomly chosen incidence (species detection) from each of the two assemblages will both represent shared species (though not necessarily the same shared species).

For replicated incidence data, \( f_{1+} \) is the number of observed shared species that occur in exactly one sample (\( X_i = 1 \)) in \( X \) and \( f_{2+} \) is the number of observed shared species that occur in exactly two samples (\( X_i = 2 \)) in \( X \); \( f_{+1} \) and \( f_{+2} \) are the corresponding numbers for sample matrix \( Y \). Define the sum of the incidence frequencies for the matrices as

\[
n = \sum_{i=1}^{s} X_i \quad \text{and} \quad m = \sum_{i=1}^{s} Y_i.
\]

Then the proposed estimators are

\[
\hat{U}_{\text{inc}} = \sum_{i=1}^{s} \frac{X_i}{n} + \frac{(\lambda - 1) f_{1+}}{\lambda} \sum_{i=1}^{s} \frac{X_i}{n} I(Y_i = 1)
\]

and

\[
\hat{V}_{\text{inc}} = \sum_{i=1}^{s} \frac{Y_i}{m} + \frac{(w - 1) f_{1+}}{w} \sum_{i=1}^{s} \frac{Y_i}{m} I(X_i = 1)
\]

(The same modifications described for eqns 7 and 8 may be applied here if \( f_{+2} = 0 \) or \( f_{+1} = 0 \).) Thus, our proposed incidence-based Jaccard and Sørensen estimators are

\[
\hat{J}_{\text{inc}} = \frac{\hat{U}_{\text{inc}} \hat{V}_{\text{inc}}}{\hat{U}_{\text{inc}} + \hat{V}_{\text{inc}} - \hat{U}_{\text{inc}} \hat{V}_{\text{inc}}}
\]

and

\[
\hat{S}_{\text{inc}} = \frac{2\hat{U}_{\text{inc}} \hat{V}_{\text{inc}}}{\hat{U}_{\text{inc}} + \hat{V}_{\text{inc}}}
\]

**PERFORMANCE TESTS: CLASSIC VS. NEW INDICES**

**Indices tested**

We carried out performance tests for: (1) the classic Jaccard and Sørensen indices (eqns 1 and 2); (2) the new, abundance-based Jaccard and Sørensen indices (eqns 5 and 6); (3) the estimators for the abundance-based indices (eqns 9 and 10); and (4) the replicated-incidence estimators for the abundance-based indices (eqns 13 and 14).

**Data sets used in the tests**

We conducted the performance tests on a large, species-rich data set for tropical rainforest ants (Longino et al. 2002), collected using several replicated, mass-collecting techniques at La Selva Biological Station in Costa Rica. Here, we present representative results for three collection methods: Berlese extraction of soil samples (217 samples, 4318 individuals, 117 species, of which 19 were singletons), Malaise trap samples for flying and crawling insects (62 samples, 1600 individuals, 103 species, of which 35 were singletons), and Fogging samples from canopy fogging (459 samples, 26302 individuals, 165 species of which 19 were singletons). [Relative abundance diagrams appear in Longino et al. (2002).] As Longino et al. (2002) point out, these three methods intentionally sample different, but overlapping segments of the local ant fauna. Whereas the raw species sum for the three methods would be \( 117 + 103 + 165 = 385 \) species, the actual number of species captured by the three methods together was only 276 species. Parallel tests for other high-richness data sets, including the rainforest tree data discussed later in this paper, yielded concordant results (A. Chao, R. L. Chazdon, R. K. Colwell & T.-J. Shen, unpublished data).

**The tests**

Although the classic Jaccard and Sørensen indices and our new indices all measure ‘similarity,’ they are intended to measure different aspects of this construct: the classic indices ostensibly measure similarity in species composition while ignoring relative abundance (although they are strongly affected by it, when sampling is involved), whereas our new indices [and many others (Legendre & Legendre 1998; Magurran 2004)] explicitly consider relative abundance. Thus, for any particular data set, differences in the absolute magnitude of incidence- vs. abundance-based Jaccard or Sørensen values (or indeed, differences between most other indices of similarity) are meaningless, in themselves.
Nevertheless, indices of compositional similarity can be compared in terms of their performance in tests of sensitivity to undersampling. Using the ant data, we illustrate three tests: (1) Test 1: equal-sized samples from a single data set (within-assemblage rarefaction); (2) Test 2: unequal-sized samples from a single data set; and (3) Test 3: equal-proportion samples from two data sets (between-assemblage rarefaction). For purposes of these tests, we treated the ant data from each collecting method (Berlese, Malaise, or Fogging) as a separate, complete assemblage, referred to here as a sampling pool. Samples of specified sizes (in terms of numbers of individuals) were then selected, at random, with replacement, from these pools. Of course, not all species present in a sampling pool are represented in smaller samples. However, because sampling was done with replacement, not all species are present even when the number of individuals selected is the same as the number of individuals in the pool.

RESULTS

Test 1: Equal-sized samples from a single data set

All similarity indices yield a true value of 1 when a complete sampling pool (assemblage) is compared with itself. What happens when a similarity index is computed for two random samples of a single sampling pool? If an index is unbiased by sample size, it should yield a value of 1 when applied to samples of any size. First, we randomly sampled individuals (with replacement) from the pooled ant data for a single collecting method to produce pairs of samples having the same number of individuals as the pools themselves (full samples). Next, we randomly selected smaller samples, each totalling one-half the number of individuals in the original sampling pool, then computed similarity indices for this sample pair. We then repeated this procedure for a pair of samples each 1/4 the size of the original pool, then a pair 1/8 the size of the pool, and so on, successively halving sample size, down to 1/64 the original number of individuals. (Note that this is quite a severe test of undersampling bias, even for these very large pools.) This entire process was repeated 1000 times and means taken, for each test of each index, and for each of the three ant collecting methods.

Figure 2 shows representative results of this test for the classic Jaccard and Sørensen indices (first column of panels, Test 1: Berlese rarefaction). Clearly both of these indices were quite sensitive to undersampling. Figure 3 (first column of panels) shows the corresponding results for the new indices for this test. The new abundance-based Jaccard and Sørensen indices, without adjustment for unseen shared species (J_{abd} and L_{abd}), were also sensitive to sample size. In
Figure 3 Random sampling tests the new overlap indices. The graphs show the effect on each index of considering random samples composed of 1/1 (Full), 1/2, 1/4, ..., 1/64 of the abundances or incidence-equivalents in the sampling pools, sampled with replacement. (The labels on the lower left graph are the same for all graphs.) Columns are described in the caption for Fig. 2. Jaccard indices: $J_{\text{ab}}$ is the new abundance-based Jaccard index, not adjusted for unseen species, computed by eqn 5. $\hat{J}_{\text{ab}}$ is the corresponding abundance-based estimator that takes unseen species into account, computed by eqn 9. The estimator based on replicated incidence data, $\hat{J}_{\text{inc}}$, is computed by eqn 13. Sørensen indices: $L_{\text{ab}}$ is the new abundance-based Sørensen index, not adjusted for unseen species, computed by eqn 6. $\hat{L}_{\text{ab}}$ is the corresponding abundance-based estimator that takes unseen species into account, computed by eqn 10. The estimator based on replicated incidence data, $\hat{L}_{\text{inc}}$, is computed by eqn 14. The true value of each index for the sampling pools considered are shown by horizontal dotted lines in the columns for Test 3 (Malaise–Fog and Malaise–Berlese rarefaction). The true index value for Test 1 and Test 2 is 1.0, the top of the graphs. To allow a valid comparison of the incidence-based estimators ($\hat{J}_{\text{inc}}$ and $\hat{L}_{\text{inc}}$) with the corresponding abundance-based estimators ($\hat{J}_{\text{ab}}$ and $\hat{L}_{\text{ab}}$, respectively), the $X$-axis for each incidence-based estimator was re-scaled so that the minimum number of incidences matches the minimum abundance of the corresponding abundance-based estimator, thus equalizing the amount of statistical information.
contrast, the Jaccard and Sørensen estimators, which include
the estimated effect of unseen shared species, proved to be
less sensitive to undersampling, remaining substantially
closer to 1 even for small samples (Fig. 3). This was true
for both the abundance-based estimators \( \hat{J}_{\text{abd}} \) and \( \hat{L}_{\text{abd}} \) and
the estimators based on replicated incidence data \( \hat{J}_{\text{inc}} \) and \( \hat{L}_{\text{inc}} \).

**Test 2: Unequal-sized samples from a single data set**

A similarity index should ideally be robust to sample size not
only for equal-sized samples, but also for samples of
unequal size. To test for this property we computed
similarity indices for samples of successively smaller size, vs.
‘full’ samples, equal in number of individuals to the number
in the corresponding sampling pool. As with the first test, an
ideal index should remain at 1, regardless of the discrepancy
in sample sizes. Figures 2 and 3 (second column, Test 2:
Berlese unequal) show such a test for the Berlese sample ant
data, using samples created by the same scheme outlined for
the first method. Even more than in the first test, the classic
Jaccard and Sørensen indices (Fig. 2) were strongly affected
by the size of the sample, leading to a severe negative bias when one sample was markedly smaller than the
full sample. In contrast, the new Jaccard and Sørensen estimators (Fig. 3, second column) were strikingly resistant
to undersampling, including both abundance-based estima-
tors \( \hat{J}_{\text{abd}} \) and \( \hat{L}_{\text{abd}} \) and the estimators based on replicated incidence data \( \hat{J}_{\text{inc}} \) and \( \hat{L}_{\text{inc}} \).

**Equal-proportion samples from two data sets**

It is all very well for a similarity index to be robust to sample size in
comparing paired samples from the same pool, but an
index is of little use if it does not retain that robustness in
comparing different data sets, while successfully detecting
compositional differences between them. We performed the
same sample size comparison procedures described for the
first set of tests, but instead of comparing sample pairs from
the same sampling pool, we compared successively smaller
sample pairs from the Malaise and Fogging [high similarity
(Longino & et al. 2002)], and from the Malaise and Berlese
(low similarity) data sets. The results for the classic Jaccard
and Sørensen indices appear in the third and fourth columns
of Fig. 2. An ideal index would yield and maintain the true
value computed for the full pools (the dotted horizontal line
in each panel) in the face of rarefaction. The classic Jaccard
and Sørensen indices proved quite sensitive to undersam-
pling in this test (Fig. 2). The new abundance-based Jaccard
and Sørensen indices, uncorrected for unseen species \( \hat{J}_{\text{abd}} \) and \( \hat{L}_{\text{abd}} \) in third and fourth columns of Fig. 3), also suffer
from undersampling bias, but the bias is quite substantially
reduced for their abundance-based counterparts corrected
for unseen species \( \hat{J}_{\text{abd}} \) and \( \hat{L}_{\text{abd}} \) in third and fourth columns of Fig. 3) as well as for the corresponding
estimators based on replicated incidence data \( \hat{J}_{\text{inc}} \) and \( \hat{L}_{\text{inc}} \)
in third and fourth columns of Fig. 3).

**APPLICATION**

As an example of the application of the new indices, we
apply the classic Jaccard index (eqn 1), the new abundance-
based Jaccard index (eqn 5) and its estimator (eqn 9) to
data from two mature and four second-growth rainforest
sites in Costa Rica. We examine compositional similarity
between species of trees \( \geq 25 \) cm diameter at breast height
(DBH; canopy individuals), canopy tree saplings (1–5 cm
DBH) and canopy tree seedlings (\( > 20 \) cm height, but
\(< 1 \) cm DBH) within four second-growth forests of
different age since pasture abandonment and in two old-
growth forests in the same study area. During early stages of
succession, when the forest canopy is first beginning to
close, fast-growing, shade-intolerant colonizing tree species
are present as canopy trees and are also found as smaller
individuals in the understory, as seedlings and saplings.
As time progresses and the understory becomes more shaded,
these shade-intolerant tree species are eliminated from the
seedling and sapling pool and shade-tolerant species readily
colonize these small size classes. These shade-tolerant
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). Thus, we would predict that, as secondary
forests mature, compositional similarity between tree species

<table>
<thead>
<tr>
<th>Site</th>
<th>Age</th>
<th>( \lambda_{\text{obs}} ) seedlings</th>
<th>( \lambda_{\text{obs}} ) saplings</th>
<th>( \lambda_{\text{obs}} ) canopy trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSUR old-growth</td>
<td>&gt; 200</td>
<td>47</td>
<td>101</td>
<td>37</td>
</tr>
<tr>
<td>LEP old-growth</td>
<td>&gt; 200</td>
<td>69</td>
<td>102</td>
<td>43</td>
</tr>
<tr>
<td>TIR</td>
<td>15</td>
<td>45</td>
<td>68</td>
<td>12</td>
</tr>
<tr>
<td>LEP</td>
<td>23</td>
<td>47</td>
<td>67</td>
<td>24</td>
</tr>
<tr>
<td>CR</td>
<td>28</td>
<td>57</td>
<td>91</td>
<td>33</td>
</tr>
</tbody>
</table>

All trees and saplings were marked and measured for diameter
within a 1 ha plot in each forest. Seedlings were sampled in 144
1 \( \times \) 5 m quadrats within the 1 ha plot, for a total area sampled
of 0.072 ha. In these analyses, we included only canopy tree species;
shrubs, treelets and midstory trees were excluded. Note that young
sites show a low number of canopy tree species per ha (individuals
\( \geq 25 \) cm DBH) and fewer sapling species compared with old-
growth forests, but differences in seedling species richness were
less pronounced.
and seedlings or saplings 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 classic Jaccard index (eqn 1) showed low compositional similarity between trees and seedlings for the four second-growth forests compared with the old-growth forests, with similarity decreasing slightly with age among the four second-growth forests (Fig. 4). Similarity between trees and saplings, in contrast, showed gradual increases from the youngest forest to the older second-growth forest, continuing the trend to old-growth forests (Fig. 4).

The abundance-based Jaccard index (eqn 5) showed a strikingly different pattern across the six forest stands. Compositional similarity between seedling and tree assemblages and between sapling and tree assemblages was initially high in the youngest stand, as we had predicted. 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 that reached a minimum in the 23-year-old LEP stand (Fig. 4). 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, the abundance-based Jaccard index began to increase, reflecting recruitment of shade-tolerant species in all three-size classes (Fig. 4). The similarity index continued to increase and stabilized at 0.4–0.5 in the two old-growth stands. With the exception of one old-growth stand, similarity indices were higher for seedlings vs. trees than for saplings vs. trees. At the scale of 1 ha plots, compositional similarity between canopy trees and seedling and sapling size classes in old-growth forests was comparable to that observed within a 15-year-old second-growth forest, but greater than that observed in second-growth forests of intermediate age. By design, the abundance-based Jaccard index responds sensitively to changes in total relative abundances of shared species during forest succession.

The abundance-based Jaccard estimator (eqn 9), which incorporates the effects of unseen shared species, showed similar general trends across stands when compared with the abundance-based Jaccard index (Fig. 4). The 28-year-old second-growth stand, however, had nearly comparable estimates of similarity compared with the two old-growth stands, suggesting that the estimator is responding to rare or infrequent species that are shared between the size classes (Fig. 4). The estimator for sapling vs. tree similarity was higher than for seedling vs. trees in the TIR second-growth site, indicating that this stand has more rare species of shared saplings than seedlings.

**CONCLUSIONS**

Because similarity is a qualitative human construct, it has no precise mathematical definition. Nevertheless, measuring ‘similarity’ relies on quantitative indices devised for the purpose, and in practice, we may expect that similarity indices fulfill reasonable criteria for their mathematical behaviour (Legendre & Legendre 1998). Given indices that make sense mathematically, it is their statistical performance under the realities of field sampling that we have concerned ourselves with here, particularly for species-rich taxa for which complete inventories are impractical or even impossible.
Using sampling simulations applied to representative field data sets, we confirmed that two of the most widely used classic indices, Jaccard and Sørensen, are negatively biased under conditions of undersampling, often quite substantially (Fig. 2). Our objective was to develop new, probability-based indices that reduce undersampling bias by estimating and compensating for the effects of unseen, shared species. We based a new similarity index on the probability that two randomly chosen individuals, one from each of two samples, both belong to any of the species shared by the two samples [not necessarily to the same shared species, the basis of \( F \) (Chave & Leigh 2002; Condit et al. 2002) and the Morisita–Horn index]. This approach opened the way to the crucial step, adjusting this probability to account for the chance that larger samples would reveal a larger proportion of shared species. As anticipated, the new indices consistently reduced undersampling bias in the performance tests, in most circumstances quite substantially. Inevitably some bias remains, especially under severe undersampling and for highly dissimilar samples. Under such conditions, relatively little information exists to guide bias reduction.

Ecologists distinguish two aspects of the compositional similarity of species assemblages: similarity of species lists (incidence) and similarity of species’ relative abundances. Classic abundance-based indices (e.g. Morisita–Horn or Bray–Curtis) match abundances, species-by-species. Our new indices take an intermediate path, by assessing the probability that individuals belong to shared vs. unshared species, without regard to which species they belong to. Unfortunately for many studies, unreplicated, pure incidence data (pairs of species lists) provide no information that can be used to estimate the number of unseen, shared species. In principle, it may be possible to derive estimators that use abundance data to correct pure incidence similarity indices for unseen species, but it is currently statistically difficult for biologically realistic data. However, we recommend the new indices for any application in which not only species matching but similarity of relative abundance is of interest. Moreover, these new indices are better suited than the corresponding classic indices for assessing compositional similarity between samples that differ in size, are known or suspected to be undersampled, or are likely to contain numerous rare species.

ACKNOWLEDGEMENTS

We thank three anonymous referees for their comments and suggestions. 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 Jorge Leiva for sharing vegetation data for tree species in mature forests. The new estimators presented in this paper are included in version 7.5 of estimates (Colwell 2004) and the program SPADe (Chao & Shen 2003), to be released upon publication of this paper. The complete derivation of eqns 7 and 8 and the variance estimators for eqns 9 and 10 are available upon request from the first author. The complete ant data sets are available from RKC.

REFERENCES


A new statistical approach for assessing similarity


Editor, Nicholas Gotelli

Manuscript received 30 June 2004

First decision made 6 August 2004

Manuscript accepted 20 October 2004

©2004 Blackwell Publishing Ltd/CNRS