Authors: Jack C. Yue[1] and Murray K. Clayton[2]

1. Department of Statistics, National Chengchi University, Taipei 11623, Taiwan, R.O.C.

2. Department of Statistics, University of Wisconsin-Madison, Madison, WI. 53706, U.S.A.

Title of the paper: A Multiple-Community Overlap Measure based on a Probabilistic Approach

Running Title: Probabilistic Multiple-Community Overlap

Number of Words: 4,384

Correspondence to:

Dr. Jack C. Yue

Department of Statistics

National Chengchi University, Taipei

Taiwan, ROC 11605

Tel: 886-2-2938-7695, Fax: 886-2-2939-8024

e-mail: csyue@nccu.edu.tw

Summary

Species diversity indices are designed to measure the species diversity of a community and compare the species distribution structure of two communities. The Shannon and Simpson indices are for describing one community, and the Jaccard and Morisita indices are for comparing two communities. Only a few indices allow the simultaneous comparison of three or more communities, such as Lande (1996) and Chao et al. (2008). In this study, we propose a multiple-community similarity index based on a probabilistic approach, and compare it with other multiple-community indices. Empirical examples are considered as a demonstration of the proposed similarity indices.

Key Words: Similarity index; Species diversity; Jaccard Index; Simpson index; Shannon Index; Maximum likelihood estimator; Delta method

1. Introduction

Similarity indices have been used in ecology, biology, and biogeography to compare, in some form, the species distributions between two populations, or the change of a distribution over time. The growing needs of analyzing large data sets have given rise to new focus on these tools. For example, some Internet search engines determine the similarity between web sites, and between web sites and search keyword strings, to construct and sort query responses.

A relatively simple way to judge the similarity of two populations is to compute a diversity index for each population, such as Shannon's or Simpson's index, and compare these diversity indices. However, such a comparison may miss both large and small structural differences. Another way to decide whether two populations are similar is to compute a similarity index specifically designed to compare the structures of two populations. Similarity indices for two populations include the Jaccard index, Morisita index, and indices proposed by Smith et al. (1996), and Yue and Clayton (2005). These indices only compare populations in pairs, however, and there is growing interest in being able to compare multiple populations simultaneously.

There is relatively little literature on multiple-population comparisons. Lande (1996) is perhaps the first paper to propose methods for measuring the similarity of three populations and more; he uses a notion analogical to the analysis of variance, i.e., separating total species diversity into between- and within- species diversity. Diserud and Ødegraad (2007) used Lande’s idea and proposed a multiple-community version of Sørensen index. Baselga et al. (2007) modified this multiple-site similarity index to have a better discrimination power when the shared species are evenly distributed. Instead of looking at functions of numbers of species, Chao et al. (2008) proposed another type of multiple-population similarity index, using a two-stage probabilistic approach to construct multiple-assemblage similarity.

In this study, we generalize the probabilistically-based two-population similarity indices outlined in Yue et al. (2001) and Yue and Clayton (2005), to construct a similarity index among three or more populations. We will talk about the motivation and development of the proposed multiple-community similarity index in the next section, including its estimate and variance. Empirical and simulation studies of the proposed index are given in Section 3. The conclusion and discussion are in Section 4. We note that the terms "population" and "community" will be used interchangeably in this paper. Also, the term "species" may be used in its biological sense, or may be used to designate other unique identifiers of members of a community, such as words on a web page, coins in a collection, etc.

2. Proposed Multiple-Community Index

In this section, we use the notion of species overlap to motivate the proposed similarity index for multiple communities. We shall briefly review the indices by Chao et al. (2008), and by Diserud and Ødegraad (2007). Their similarity indices are similar to that proposed by Lande (1996), and are the ratio of between species diversity to within species diversity. The numerator includes species overlap between two and more than two communities, while the denominator is species diversity of one community. In other words, these indices measure a mixture of all levels of overlap and are not restricted to two communities. In this study, we will develop overlap measures of a specific level, instead of mixtures of different levels.

The "species overlap" refers to a measure of the probability of discovering a shared species (Smith et al., 1996, Yue et al., 2001). Specifically, let A_i = {Species is observed in population i}. Then, in a two-community setting, the species overlap can be defined as $\theta_2 = \dfrac{P(A_1 \cap A_2)}{P(A_1 \cup A_2)}$. Extending the notion into a multiple-community setting, the species overlap is $\theta = \dfrac{P(\bigcap_{1\le j\le k} A_j)}{P(\bigcup_{1\le j\le k} A_j)}$, where k is the number of communities. The species overlap θ is to the conditional probability of discovering a species common in all k communities, given that the species is present in at least one of the k communities.

From a practical view, θ has two drawbacks. First, to compute $P(\bigcap_{1\le j\le k} A_j)$ and $P(\bigcup_{1\le j\le k} A_j)$, all possible combinations of $P(\bigcap_{j\in I} A_j)$ and $P(\bigcup_{j\in I} A_j)$ are needed, where I is a subset of {1, 2, …, k}. This is impractical as the number of communities k increases. Second, unless there are quite a few common shared species, the value of θ is likely to be very small and cannot help to distinguish which group of communities are the most similar.

Alternatively, we propose a set of conditional indices for measuring multiple-community overlap. Let θ_r be the r^{th} order species overlap, r = 2, 3, …, k. For example, θ_2 is the conditional probability that a species is present in two communities, given that it appears in at least one community. Likewise, θ_r is the conditional probability that a species is present in r communities, given that it appears in r−1 communities. The overall similarity for the k communities can still be achieved by $\theta = \prod_{r=2}^{k} \theta_r$. However, while both the Jaccard and Morisita indices can be considered natural extensions of the two-community similarity means, the values of θ_r's can provide extra information that cannot be seen from θ alone. Also, if $\theta_{r_0} = 0$, then $\theta = 0$ and we do not need to compute the values of θ_r for $r > r_0$. In the following, we will discuss the implementation of this measure of similarity under different scenarios, including different normalization factors. These will lead naturally to the generalization of three frequently used two-community similarity indices: the Jaccard index, the index by Smith et al. (1996), and the Morisita index.

We shall cover the case of sampling species first and find a feasible definition for θ_r. Adopting the approach in Yue et al. (2001), θ_r is the ratio of two probabilities, where the probability of observing shared species in r communities is in the numerator and that in r–1 communities is in the denominator. If every species is equally likely to be observed, then the probability of observing a species which is common to r communities is s_r / s, where s_r is the number of shared species in r communities and s is the number of distinct species in all k communities. In other words, a possible way to define θ_r is by $\frac{s_r / s}{s_{r-1} / s} = s_r / s_{r-1}$. Note that there are r out of k possible combinations to consider the shared species in r communities. Therefore, we can define a Jaccard-type (or Sørensen-type) version of θ_2, namely $\theta_2(J)$, as the ratio of the average number of shared species of any two communities to the average number of species of any one community. A special case of $\theta_2(J)$ is the Jaccard and Sørensen indices with k = 2. The overall multiple community Jaccard index can be expressed as $\theta(J) = \prod_{r=2}^{k} \theta_r(J)$. Also, the multiple-site similarity index proposed by Diserud and Ødegraad (2007) has s_1 on the denominator and the sum of shared species for two and more communities $s_r\ (r \ge 2)$ on the numerator in the form of $\sum_{r \ge 2} (-1)^r s_r$.

The similarity index by Smith et al. (1996) can be treated as a generalization of the Jaccard index, assuming that the species are not equally likely to be observed. In addition, the species can be classified into the shared species group or non-shared species group. Yue et al. (2001) showed that this index can be expressed as

$\dfrac{(\sum_{i\in I_C} p_{i1})(\sum_{i\in I_C} p_{i2})}{(\sum_{i\in I_C} p_{i1})+(\sum_{i\in I_C} p_{i2})-(\sum_{i\in I_C} p_{i1})(\sum_{i\in I_C} p_{i2})}$, where p_{ij} is the proportion of species i abundance in the total abundance of individuals in community j ($j = 1, 2$) and I_C is the index set of shared species for communities 1 and 2. If a species is drawn from each community and its occurrence in one community is independent from its occurrence in another community, the denominator is the probability of observing the share species group for at least one community and the numerator is the probability of observing the shared species group for both communities. To simplify the computation, we adapt the format of the Morisita index, i.e., $\dfrac{2(\sum_{i\in I_C} p_{i1})(\sum_{i\in I_C} p_{i2})}{(\sum_{i\in I_C} p_{i1})+(\sum_{i\in I_C} p_{i2})}$ and define

$$\theta_2(S)=\frac{\sum_{1\le j<l\le k}\left[2(\sum_{i\in I_C} p_{ij})(\sum_{i\in I_C} p_{il})\right]}{\sum_{1\le j<l\le k}\left[\sum_{i\in I_C}(p_{ij}+p_{il})\right]} \qquad (1).$$

It can be shown easily that the value of $\theta_2(S)$ is between 0 and 1, and the symbol “S” represents that the similarity index is of the type Smith et al. (1996).

Note that $\theta_2(S)$ can be easily extended to the 3rd order index as $\theta_3(S)=\dfrac{\sum_{1\le j<l<m\le k}\left[(\sum_{i\in I_C} p_{ij})(\sum_{i\in I_C} p_{il})(\sum_{i\in I_C} p_{im})\right]\Big/\binom{k}{3}}{\sum_{1\le j<l\le k}\left[(\sum_{i\in I_C} p_{ij})(\sum_{i\in I_C} p_{il})\right]\Big/\binom{k}{2}}$, Note that $\theta_3(S)$ is like the conditional probability of discovering a species in three communities, given that the species appears in two communities. Higher order indices can be defined accordingly. The advantages of using $\theta_r(S)$ is that the original overlap measure approximately satisfies $\theta(S)=\prod_{r=2}^{k}\theta_r(S)$, and once $\theta_{r_0}(S)=0$, it is not necessary to

compute the values of $\theta_r(S)$ for $r > r_0$ and $\theta = 0$.

The conditional index $\theta_r(S)$ is a ratio of the r^{th} order overlap to the $(r-1)^{\text{th}}$ order overlap, for all the possible combinations of k communities. In other words, both the numerator and denominator of the index are the weighted sum of possible combinations. If we use a "leave-one-out method", then we can verify whether there are any communities significantly different from the others, taken as a group. Also, the standard errors of the index θ_r can be computed using the delta method or bootstrap simulation. (Yue et al., 2001, Yue and Clayton, 2005).

The proposed similarity index can also be generalized to the Morisita-type, from the perspective of sampling single observations. Assume that one observation is sampled randomly from any community. We shall use θ_2 as a demonstration, where θ_2 is the probability of observing a shared species from both communities, given that the shared species is observed in at least one community. Then we can define a 2nd order index, in which the numerator is the probability of observing a shared species, i.e., $\sum_i p_{ij} p_{il}$, and the denominator is $\sum_i \left(p_{ij}(1-p_{il}) + (1-p_{ij})p_{il} + p_{ij}p_{il}\right) = \sum_i \left((p_{ij}+p_{il}) - p_{ij}p_{il}\right)$, where j and l ($j \neq l$) represent two different communities. If move the $p_{ij}p_{il}$ from the denominator to the numerator, then we can define the 2nd order Morisita-type similarity index as

$$\theta_2(M) = \frac{\sum_{1\le j<l\le k}\left[2(\sum_i p_{ij}p_{il})\right]}{\sum_{1\le j<l\le k}\left[\sum_i (p_{ij}+p_{il})\right]} = \sum_{1\le j<l\le k}\left[(\sum_i p_{ij}p_{il})\right] \Big/ \binom{k}{2} \qquad (2).$$

The Morisita-type index can be extended to the r^{th} order, defined as the probability of observing a shared species in r communities, given that it is also a shared species in $r-1$ communities. In other words,

$$\theta_r(M) = \frac{Ave.\,of \sum_i (p_{ij_1} \cdots p_{ij_r})}{Ave.\,of \sum_i (p_{ij_1} \cdots p_{ij_{r-1}})}, \tag{3}$$

where $2 \le r \le k$ and $0 \le \theta_r(M) \le 1$. In addition to the fact that $\theta(M) = \prod_{r=2}^{k} \theta_r(M)$, the proposed index also has an advantage in computation. Since it is assumed that one observation is taken from a community, the probability calculation only involves first order terms in the species proportions.

Note that the maximum likelihood estimator can be used for the proposed similarity measure, similar to that in Yue et al. (2001), or Yue and Clayton (2005). For example, we can plug in $\hat{p}_{ij} = \frac{X_{ij}}{n_j}$ as the estimate of p_{ij}, where X_{ij} is the number of occurrences of species i for n_j observations taken from the j^{th} community. As long as the numbers of observations in the communities satisfy $Min\{n_1, n_2, \ldots, n_k\} \to \infty$, we can show that $\hat{\theta}_r \to \theta_r$ in probability according to Slutsky's lemma. The asymptotic variance of $\hat{\theta}_r$ can be derived via Cramer's delta method.

Of course, the variance of $\hat{\theta}_r$ can also be derived from bootstrap simulation, as in Yue et al. (2001) and Yue and Clayton (2005). Based on previous studies, the variances from the delta method and bootstrap simulations are close, provided that there are many observations from each community. These results are not shown here, but instead, we will use computer simulations to explore the large sample properties of the variance for $\hat{\theta}_r(M)$.

3. Examples and Simulation

In this section, we will use examples and computer simulation to demonstrate the proposed similarity indices. First, we use examples to compare the proposed indices with a multiple community index proposed by Chao and colleagues.

Example 1. Chao et al. (2008) also proposed a multiple-community similarity index based on a probabilistic approach, and thus we first compare our proposed index with theirs. The index proposed by Chao et al. is a Morisita-type index and is a ratio of between- and within- species diversity, similar to Lande (1996). It is based on a two-stage probabilistic approach. The first stage is to choose r communities with replacement from among the total of k communities (so $2 \le r \le k$) and the second stage is to randomly select an individual from each of the communities chosen in stage 1. The index is denoted as C_{rk}, and defined to be the ratio of two probabilities: the denominator is the probability that the r individuals are of the same species, given that the Stage I observations are all from the same community; the numerator is the probability of the same event given that at least two communities are represented.

$$C_{rk} = \frac{\frac{1}{k^r - k}\sum_{i=1}^{S}\left[(p_{i1} + \cdots + p_{ik})^r - (p_{i1}^r + \cdots + p_{ik}^r)\right]}{\frac{1}{k}\sum_{i=1}^{S}\left[p_{i1}^r + \cdots + p_{ik}^r\right]}. \tag{4}$$

Chao et al. constructed Table 1 to demonstrate the need for introducing a multiple-community similarity index; their table shows two groups of communities with different species structures. In their table it is obvious that there is a shared species for 3 communities in Group 1, but not in Group 2. Intuitively, these two groups should have different similarity values, especially for any measure of similarity that compares the three communities simultaneously.

We slightly modify the setting of Table 1, in order to explore whether the shared species is a dominant (or rare) species and how that could influence the similarity value. Let the proportion of the shared species be α in each community, meaning that the other species is of proportion 1−α. We shall consider three different values of α: 0.1, 0.5, and 0.9, indicating that the share species is rare, not so rare, and dominant, respectively. We shall compare the values of the similarity indices by Chao et al. and the proposed approach.

[Insert Table 1 here.]

Table 2 lists similarity indices for 2-community and 3-community cases, for three kinds of similarity indices: $\theta_r(J)$, $\theta_r(M)$, and C_{r3} is the multiple-community index proposed by Chao et al., where 3 denotes the number of communities in a group. These two groups have quite different similarity index values for the case r = 3, which indicates that the proposed index and the index proposed by Chao et al. distinguish between these two groups. However, there are some differences between the two indices that bear discussion.

[Insert Table 2 here.]

First, as mentioned by Chao et al., C_{rk} is more sensitive to the proportions of dominant shared species. This is the case since the value of C_{rk} when r = 2 for Group 1 is closer to 1 when α = 0.9. We found that $\theta_r(M)$ also possesses a similar behavior as C_{rk} under same situation, and may be more sensitive than C_{rk} in that case. Also, our proposed index does detect whether there are no shared species in the r communities, as in Group 2 where the similarity value is 0, but C_{rk} does not produce the same result. It looks like that the similarity measured by C_{33} is not restricted to be of 3rd order and it is a mixture of the second and third order overlap, since there is no shared species in 3 communities and the similarity value should be 0.

There is another difference which is subtle. If we look at the 3rd order similarity, the values of the proposed index $\theta_r(M)$ are always smaller in Group 2, but the value of C_{rk} in Group 2 has a fairly large similarity value (bigger than in Group 1) for the case α = 0.1 and 0.9. If larger value of C_{rk} indicates a more similar structure within a group, then using C_{rk} would imply that the communities in Group 2 are more similar than those in Group 1, which contradicts our intuition.

Example 2. The second example is from Chao et al. (2008), and consists of data that were collected from four forests in Costa Rica. The forest data can be assessed at the Biometrics website http://www.biometrics.tibs.org. There are three records for each species, seedlings, saplings, and trees. We compute the values of $\theta_r(J)$ and $\theta_r(M)$, and compare the differences for these three records (Table 3). Both indices show that there are no shared species in the four communities for trees, but the values of $\theta_r(J)$ and $\theta_r(M)$ differ quite a lot otherwise. This indicates that the shared species in r–1 communities are also likely to be shared species in r communities, but the shared species are not the dominant species.

[Insert Table 3 here.]

From Table 3, we can see that the multiple-community similarity indices $\theta_r(J)$ and $\theta_r(M)$ carry different information. As expected, the Jaccard-type index $\theta_r(J)$ is based on the number of shared species, and the Morisita-type index $\theta_r(M)$ provide further information with respect to species proportions. With these two indices showing different orders of similarity, we can have a better idea of how the communities in a group are structured. In other words, it is possible to judge not only the number of shared species but also their species proportions. For example, the sapling data show very different patterns for $\theta_r(J)$ and $\theta_r(M)$. The values of $\theta_r(J)$ are close to 1, but the values of $\theta_r(M)$ are small. Together, these suggest that, although the communities have many species in common, the species proportions are not similar, and this holds at each of the levels (r) of comparison.

Example 3. The third example is from the Breeding Bird Survey (BBS) -- the related information can be found at the website http://www.pwrc.usgs.gov/BBS/. The BBS is a cooperative effort between the U.S. and Canada to monitor the status and trends of North American bird populations. Table 4 lists the values of $\theta_r(J)$ and $\theta_r(M)$ for different 4 routes of BBS data.

[Insert Table 4 here.]

For r=2 and 3, the values of $\theta_r(J)$ are large for the cases of Routes 44 and 58, and moderately large for Route 1. This means that the shared species in r–1 communities are likely to be shared species in r communities (r = 2 and 3). On the other hand, the values of $\theta_r(M)$ are small for r=2 and 3, which means that the shared species are not the dominant species. This is similar to the results in Table 3, indicating that the communities have many species in common but differ in the species proportions.

In addition to these real examples, we also use simulation to check the proposed indices. In particular, we should show that the NPMLE can provide a reliable estimate to the proposed indices.

Example 4. Suppose there are k = 3, 5, and 10 identical populations, with 10 species in each population, and all 10 species are shared species. Let the species proportions follow a geometric distribution, with $p_{ij} \propto \alpha^i, i = 1, 2, \ldots$ and $0 < \alpha < 1$ for every community. We shall check the performance of an NPMLE-based estimate. As expected, the NPMLE of the similarity index converges rapidly as the number of observations increases, and its average is close to the theoretical value (these results are not shown here). We will focus on the variance of the NPMLE.

Because the results are similar for $0 < \alpha < 1$, we will use the result of the case $\alpha = 0.5$ as a demonstration. Figure 1 shows the variance of NPMLE times the sample size vs. the sample size for different numbers of populations (1,000 simulation runs). Interestingly, the variances of the NPMLE show some patterns and it seems that the variance multiplied by the sample size is approximately a constant. This is very similar to the results in Yue and Clayton (2005) and Yue et al. (2001), where the asymptotic variance of the NPMLE is inversely proportional to the sample size, as shown in the previous section.

[Insert Figure 1 here.]

For the case of $\alpha = 1$, i.e., the species proportions follow a uniform distribution (even population), the asymptotic variance of the NPMLE shows a similar pattern but the variance converges much faster. Figure 2 shows the similarity index values vs. the number of populations for the even population case (1,000 simulation runs). The variance of the NPMLE, multiplied by the square of the sample size (instead of the sample size) is close to a constant. If the graphs are drawn using the reciprocals in Figure 2, these values are close to the number of all possible combinations for choosing 2 among *k*, which are 3, 10, and 45 for *k* = 3, 5, 10, respectively. Yue et al. (2001) also found that the variance of the NPMLE converges much faster in the even population case.

[Insert Figure 2 here.]

4. DISCUSSION

In this study, we propose a probabilistic approach to measure multiple-community similarity. In particular, the similarity index can be computed recursively and is adapted from the approach used in Yue et al. (2001) and Yue and Clayton (2005). The proposed similarity index can be separated into components that measure various orders of similarity, and thus can provide more thorough information regarding shared species. If we are sampling species, the proposed similarity index can be treated as a generalization of the Jaccard index and the index by Smith et al. (1996). If one observation is taken from each community, then the proposed approach can be generalized to indices similar to the Morisita index.

The proposed index is different from the one by Chao et al. (2008), which is also a multiple-community similarity index and also based on a probabilistic approach. In brief, the index by Chao et al. involves sampling populations while the proposed index involves sampling species. The proposed index can be viewed as measuring the conditional (or marginal) information regarding shared species at different levels, while the index of Chao et al. measures the unconditional information regarding shared species and is a mixture of all levels of similarity. These two similarity indices are different in nature.

We used an example to compare their difference. We found that both indices can measure the similarity among three and more than three populations. They both are sensitive to the dominant shared species, and the one by Chao et al. is more sensitive.

On the other hand, the proposed index can distinguish among different levels of shared structure. It is difficult to conclude that one index is better than the other.

In practice, we recommend using together the two similarity indices defined via the proposed approach, the Jaccard-type index $\theta_r(J)$ and the Morisita-type index $\theta_r(M)$. The first index provides the number of shared species and the other suggests whether the shared species are dominant species.

Also, since the proposed index is a conditional-type similarity index, it can be used to identify populations which have the most different species structure. For example, suppose there are *k* populations and one of them has a completely different species structure. We can apply the omit-one procedure to detect this population. First, we compute the values of proposed similarity index for all possible combinations of *k*–1 populations. Next, we can evaluate if the omit-one index values can be separated into two classes. The group with the larger omit-one index value will have only one member.

Still, there are limitations in applying the proposed similarity index. For example, suppose there are two communities and both have identical species structure. For simplicity, assume there are only two species, with proportion α and 1−α, respectively. By the definition of the proposed index $\theta_r(M)$, $\theta_r(M)$ is an increasing (or decreasing) function of α if $0.5 \le \alpha \le 1$ (or $0 \le \alpha \le 0.5$). But intuitively, two identical populations shall have similarity value as 1. The reason for producing such result is that the sampling is involved. One way to modify $\theta_r(M)$ is to multiply by a normalizer when computing the probability of discovering shared species. For example, a possible modification of $\theta_r(M)$ is

$$\theta_2^*(M) = \sum_{1 \le j < l \le k} \left[(\sum_i p_{ij} p_{il}) / (p_{\bullet j} p_{\bullet l}) \right] \Big/ \binom{k}{2},$$

where $p_{\bullet j} p_{\bullet l} = \sum_i p_{ij} p_{\tau(i)l}$ is the normalizer and is the maximum among all possible one-to-one function $\tau : \{1, \ldots, S\} \to \{1, \ldots, S\}$. It can be shown that the value of $\theta_2^*(M)$ is always one, given two or more identical communities. A similar modification can be applied to higher orders of the Morisita-type index $\theta_r(M)$.

REFERENCES

Baselga, A., Jimenez-Valverde, A., and Niccolini, G. (2007). A Multiple-site Similarity Measure Independent of Richness. *Biological Letters* 3: 642-645.

Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T. (2006). Abundance-Based Similarity Indices and Their Estimation when There are Unseen Species in Samples. *Biometrics* 62:361-371.

Chao, A., Jost, L., Chiang, S. C., Jiang, Y.-H., and Chazdon, R. L. (2008). A Two-Stage Probabilistic Approach to Multiple-Community Similarity Indices. *Biometrics* 64:1178-1186.

Diserud, O. H. and Ødegaard, F. (2007). A Multiple-site Similarity Measure. *Biology Letters* 3: 20–22.

Emigh, T. H. (1983). On the Number of Observed Classes from a Multinomial Distribution. *Biometrics* 39:485-491.

Lande, R. (1996). Statistics and Partitioning of Species Diversity, and Similarity among Multiple Communities. *OIKOS* 76:5-13.

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.

Yue, J. C., Clayton, M. K., and Lin, F. (2001). A Nonparametric Estimator of Species Overlap. *Biometrics* 57:743-749.

Yue, J. C. and Clayton, M. K. (2005). An Overlap Measure based on Species Proportions. *Comm. Statist. Theory Methods* 34:2123-2131.

Table 1. Two groups of communities

Species	Group 1			Group 2		
	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
1	X*a*	X*a*	X*a*	X*a*	—	X*a*
2	X	—	—	X	X*a*	—
3	—	X	—	—	X	X
4	—	—	X	—	—	—

Note: "X" and "Xa" indicate that the species are present, and "—" is absent. "X" and "X*a*" are the species with proportion 1−α and α, respectively.

Table 2. Multiple-community Indices for Two Groups of Communities

		α = 0.1		α = 0.5		α = 0.9		$\theta_r(J)$
	r	C_{r3}	$\theta_r(M)$	C_{r3}	$\theta_r(M)$	C_{r3}	$\theta_r(M)$	
Group 1	2	0.0122	0.01	0.5	0.25	0.9878	0.81	0.5
	3	0.0014	0.1	0.5	0.5	0.9986	0.9	1
Group 2	2	0.3699	0.0303	0.5	0.25	0.3699	0.0303	0.5
	3	0.2654	0	0.375	0	0.2654	0	0

Table 3. Similarity indices of tropical rain forest data

	r	Seedling	Sapling	Tree
$\theta_r(M)$	2	0.0376	0.0236	0.0356
	3	0.0195	0.0272	0.0077
	4	0.0038	0.0069	0
$\theta_r(J)$	2	0.4968	0.7988	0.3193
	3	0.5516	0.8556	0.2368
	4	0.5614	0.9351	0

Table 4. Similarity indices of breeding bird survey data

	r	Route 44	Route 58	Route 15	Route 1
$\theta_r(M)$	2	0.0799	0.0664	0.0303	0.0366
	3	0.1638	0.0926	0.0545	0.0722
	4	0.0456	0.0049	0.0161	0.0230
$\theta_r(J)$	2	0.7813	0.7453	0.3211	0.5097
	3	0.8436	0.8228	0.4684	0.6852
	4	0.2220	0.2154	0.1486	0.1923

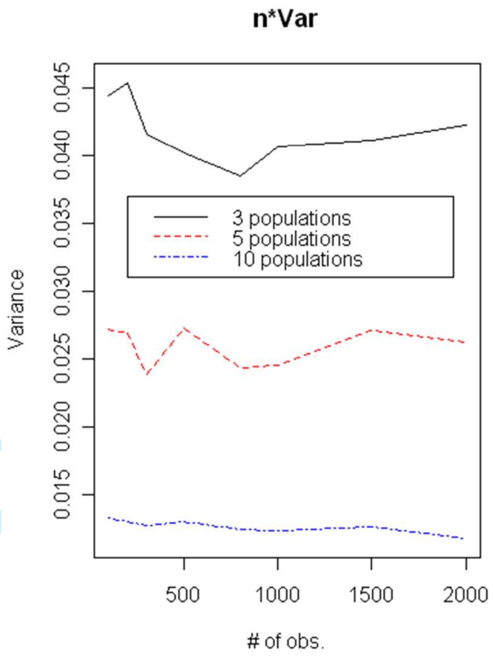

Figure 1. Variance of Similarity index vs. number of populations (geom. dist.)

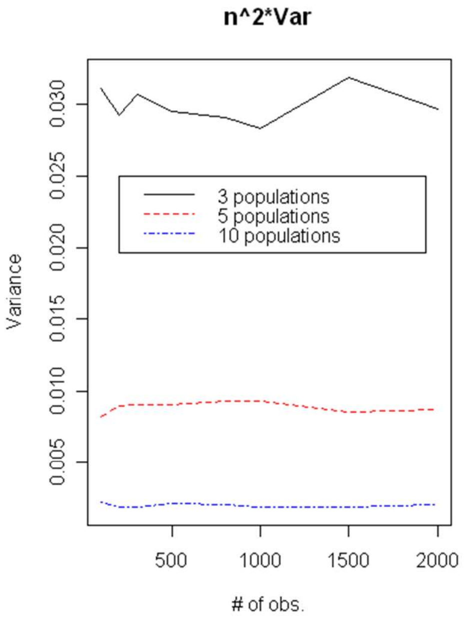

Figure 2. Variance of Similarity index vs. number of populations (uniform dist.)