A Nonparametric Estimator of Species Overlap

Jack C. Yue[1], Murray K. Clayton[2], and Feng-Chang Lin[1]

[1]Department of Statistics, National Chengchi University, Taipei, Taiwan 11623, R.O.C. and [2]Department of Statistics, University of Wisconsin-Madison, Madison, WI. 53706, U.S.A.

SUMMARY

For two communities, species overlap has been defined by Smith et al. (1996) as the probability that a randomly selected species is present in both communities, given that it is present in at least one community. Species overlap can thus be used to describe the similarity of two communities. In contrast to the parametric estimator of Smith et al., we propose a Nonparametric Maximum Likelihood Estimator (NPMLE). We prove that the NPMLE is consistent and asymptotically normally distributed, and show that computation of the NPMLE and its standard error is straightforward. We also compare the NPMLE and the estimator of Smith et al. for a variety of situations.

Key Words: Bootstrap; Jaccard's index; Maximum likelihood estimator; Similarity index; Species diversity.

1. Introduction

In ecology, the comparison of two or more populations and the evaluation of a population's change over time are often of interest. The number of species and proportions of species, and functions of these, are often used to measure species diversity. Shannon's index (or the Shannon-Wiener index) and Simpson's index are two well-known measures used to describe community structure. The comparisons of populations are based on the indices calculated separately for each population.

Jaccard's index is another way to compare populations, specifically for describing their similarity. It is defined as the ratio of the number of common species to the number of distinct species in two populations, i.e. Jaccard's index is given by $\theta_J = c/(s_1 + s_2 - c)$, where s_i is the number of species in population i, $i = 1, 2$ and c is the number of common (i.e., "shared") species. The Jaccard index does take the species common to both populations into account and is easy to compute. However, all species have equal weight and species proportion information is not used in the Jaccard index. Because all species are equally weighted, it is possible that the similarity of two communities would be underestimated by the Jaccard index.

Smith et al. (1996) proposed a new species overlap measure, defined as the probability that, given a randomly selected species is present in at least one of the two communities, it is present in both communities. This measure takes into account the number of species and puts larger weight on those species which appear more frequently in the sample. When Smith et al. estimated this measure for data of Abele (1979), their estimate of community species overlap was 50% larger than Jaccard's index.

The species overlap measure proposed by Smith et al. seems more appropriate than Jaccard's measure of species overlap, since more information is used in the new measure. Moreover, it has an interpretation that is intuitive, and expressible as a probability. In this paper, we propose a Nonparametric Maximum Likelihood Estimator (NPMLE) of species overlap, given that the species proportions are fixed. The NPMLE is easy to compute (similar to the Jaccard index) and its standard error can be computed via bootstrapping or via an asymptotic expression.

We first introduce the NPMLE in Section 2, followed by some theoretical results in Section 3. We use two examples to compare the estimate of Jaccard's index, the estimate proposed by Smith et al., and the NPMLE in Section 4. Further comparisons among these estimates, based on computer simulation, are in Section 5. In Section 6 we make some concluding remarks, paying special attention to the context in which our evaluations and comparisons are made.

2. Notation

Smith et al. (1996) first proposed describing the similarity of two communities using the probability that a randomly selected species is present in both communities, given that it is present in at least one community. Their approach is based on a parametric assumption, which they called the delta-beta-binomial model. Let N_{i1} and N_{i2} be the numbers of individuals of species i in the sample from population 1 and 2, let $B(\cdot, \cdot)$ be the beta function, and let α and β be parameters. Also, let p be the probability of observing species that are in population 1 but not in population 2, and let q be the probability of observing species that are in population 2 but not in population 1. Smith et al. assume that the

probability of observing j individuals of species i in population 1, conditional on $N_{i1} + N_{i2} = n_i$, is

$$d(j; n_i, \alpha, \beta, p, q) \;=\; \delta(j)q + \delta(n_i - j)p + (1 - p - q)b(j; n_i, \alpha, \beta), \qquad (1)$$

where

$$\delta(j) = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } j \neq 0 \end{cases}$$

and

$$b(j; n_i, \alpha, \beta) \;=\; \binom{n_i}{j} \frac{B(j + \alpha, n_i - j + \beta)}{B(\alpha, \beta)}. \qquad (2)$$

Based on this model, $1 - p - q$ is the corresponding overlap index. An estimate of the species overlap can be obtained by numerical maximum likelihood estimation, and Smith et al. approximate its standard error by inverting the observed information matrix.

To introduce an NPMLE of species overlap, consider an experiment where a species is selected at random. For this experiment, define two events: $A = \{\text{Species is observed in population 1}\}$ and $B = \{\text{Species is observed in population 2}\}$. Then $A \cap B$ is the event that the species selected is in both populations 1 and 2. Let θ_n denote the probability that a randomly selected species is present in both populations, given that it is present in at least one population. Then using the notation defined above, $\theta_n = P(A \cap B)/P(A \cup B)$. We will not evaluate $P(A \cap B)$ and $P(A \cup B)$ directly. Instead, since $P(A) + P(B) = P(A \cap B) + P(A \cup B)$, we can express θ_n as

$$\theta_n \;=\; \frac{P(A \cap B)}{P(A \cup B)} = \frac{a\,b}{a + b - a\,b}, \qquad (3)$$

where $a = P(A \cap B|A)$ and $b = P(A \cap B|B)$, i.e. a and b are the probabilities of observing shared species in populations 1 and 2, respectively. Note that,

similar to Smith et al. θ_n is defined conditionally on $A \cup B$. In other words, we do not explicitly define a mechanism for sampling species directly; given any such mechanism that follows the usual rules of probability, our definition is logically consistent.

When any species in population 1 is equally likely to be selected, i.e. population 1 is uniform, then $P(A \cap B|A) = c/s_1$. Similarly, $P(A \cap B|B) = c/s_2$ under the uniform distribution assumption for population 2. Then it is straightforward to show that $\theta_n = \theta_J$, and thus the Jaccard index can be treated as a special case of θ_n if sampling of species is uniform within each population. However, as pointed out by Smith et al., the Jaccard index is likely to underestimate the true percentage of overlap since the species proportions are not included in the overlap measure. In order to reduce the underbias of the Jaccard index, in this study, we assume that the probability of observing a certain species in a population is equal to its species proportion. Then a is the sum of species proportions for all shared species in population 1, and b is the sum of species proportions for all shared species in population 2.

To gain some intuition for our model, note that, based on our definition, the probability that a randomly sampled species is from population 1 is given by $P(A)/P(A \cup B) = b/(a + b - ab)$ if $a, b \neq 0$. Then, for example, suppose that $a = 1$, i.e. all species in population 1 are shared species, and so population 1 can be treated as a sub-population of population 2. Then $P(A)/P(A \cup B) = P(A \cap B)/P(B) = b$ and $P(B)/P(A \cup B) = 1$. These results make intuitive sense under the assumption $a = 1$. Likewise, if $a = 0$, i.e. there are no shared species and $b = 0$ as well, then $A \cap B = \emptyset$ and $P(A \cap B)/P(A \cup B) = 0$.

We can use Maximum Likelihood to find estimates of a and b, and then find the MLE of θ_n, i.e. $\hat{\theta}_n = \hat{a}\hat{b}/(\hat{a} + \hat{b} - \hat{a}\hat{b})$ where $\hat{a}$ and $\hat{b}$ are the MLE's

of a and b, respectively. For example, suppose that 6 of the observed species, which account for 20 of 100 individuals in the sample of population 1, are also observed in the sample of population 2. Then $\hat{a}$ equals $20/100 = 0.20$. The standard errors of $\hat{a}, \hat{b}$, and $\hat{\theta}_n$ can be obtained from a bootstrap simulation. The computation of the NPMLE will be discussed in detail in Section 4. First, we show some theoretical results in the following section.

3. Theoretical Results

Let p_i be the proportion of species i $(i = 1, \ldots, s_1)$ in population 1, and q_j the proportion of species j $(j = 1, \ldots, s_2)$ in population 2. We assume that the p_i's and q_j's are fixed. Let ξ_X and ξ_Y be the observed species counts for populations 1 and 2, respectively. Finally, let n_1 and n_2 be the numbers of observations from population 1 and 2, and let x_i and y_i be the numbers of occurrences of the ith species in populations 1 and 2, respectively.

We now proceed to introduce an NPMLE of a. Note that a can be expressed as $a = \sum_i p_i I\{i \in C\}$ where C is the index set of the shared species. A natural choice of the estimate for p_i is $\hat{p}_i = x_i/n_1$, and a natural estimate of $I\{i \in C\}$ is $I\{x_i > 0\}I\{y_i > 0\}$, i.e. the ith species is in both populations if we observe it from the sample at least once in each population. Then the NPMLE of a can be expressed as

$$\hat{a} = \sum_i x_i/n_1 \, I\{x_i > 0\} \, I\{y_i > 0\} = \sum_i x_i/n_1 \cdot I\{y_i > 0\}.$$

The last equality holds because $x_i = 0$ implies $\hat{p}_i = 0$. Finally, without loss of generality we assume that species 1 to c are species common to both populations. Then $a = \sum_{i=1}^{c} p_i$, $\hat{a} = \sum_{i=1}^{c} x_i/n_1 \cdot I\{y_i > 0\}$, $b = \sum_{i=1}^{c} q_i$, and $\hat{b} = \sum_{i=1}^{c} y_i/n_2 \cdot I\{x_i > 0\}$.

Intuitively, $\hat{a}$ is close to a when the number of observations taken from population 2 is sufficiently large. In fact,

$$\begin{aligned} E(\hat{a}) &= E[E(\sum_{i=1}^{c} \frac{x_i}{n_1} \cdot I\{y_i > 0\}|\xi_X)|\xi_Y] = \sum_{1\le i\le c} p_i(1-(1-q_i)^{n_2}) \\ &\to \sum_{1\le i\le c} p_i = a, \quad \text{as } n_2 \to \infty. \end{aligned}$$

We can show a similar property for $\hat{b} = \sum_{i=1}^{c} y_i/n_2 \cdot I\{x_i > 0\}$.

It is straightforward to show that the variances of $\hat{a}$ and $\hat{b}$ are

$$\begin{aligned} Var(\hat{a}) &= \sum_{i=1}^{c} p_i^2[1-(1-q_i)^{n_2}](1-q_i)^{n_2} + \sum_{i=1}^{c} \frac{p_i(1-p_i)}{n_1}[1-(1-q_i)^{n_2}] \\ &\quad +2\sum_{1\le i<j\le c} p_i p_j[(1-q_i-q_j)^{n_2} - (1-q_i)^{n_2}(1-q_j)^{n_2}] \\ &\quad -2\sum_{1\le i<j\le c} \frac{p_i p_j}{n_1}[1-(1-q_i)^{n_2} - (1-q_j)^{n_2} + (1-q_i-q_j)^{n_2}] \quad (4) \end{aligned}$$

and

$$\begin{aligned} Var(\hat{b}) &= \sum_{i=1}^{c} q_i^2[1-(1-p_i)^{n_1}](1-p_i)^{n_1} + \sum_{i=1}^{c} \frac{q_i(1-q_i)}{n_2}[1-(1-p_i)^{n_1}] \\ &\quad +2\sum_{1\le i<j\le c} q_i q_j[(1-p_i-p_j)^{n_1} - (1-p_i)^{n_1}(1-p_j)^{n_1}] \\ &\quad -2\sum_{1\le i<j\le c} \frac{q_i q_j}{n_2}[1-(1-p_i)^{n_1} - (1-p_j)^{n_1} + (1-p_i-p_j)^{n_1}] \quad (5) \end{aligned}$$

respectively, which can be computed via

$$Var(X) = Var[E(X \mid Y)] + E[Var(X \mid Y)].$$

Therefore, $Var(\hat{a}) \to a(1-a)/n_1$ as $n_2 \to \infty$ and $Var(\hat{b}) \to b(1-b)/n_2$ as $n_1 \to \infty$.

Similarly, the covariance of $\hat{a}$ and $\hat{b}$ is given by

$$\begin{aligned} Cov(\hat{a},\hat{b}) &= \sum_{i=1}^{c} p_i q_i[(1-q_i)^{n_2} + (1-p_i)^{n_1} - (1-p_i)^{n_1}(1-q_i)^{n_2}] \\ &\quad + \sum_{1\le i\ne j\le c} p_i q_j[p_j(1-p_j)^{n_1-1} + q_i(1-q_i)^{n_2-1} \\ &\qquad - (1-p_j)^{n_1-1}(1-q_i)^{n_2-1}(p_j+q_i-p_j q_i)] \quad (6) \end{aligned}$$

and $Cov(\hat{a}, \hat{b}) \to 0$ as $\min\{n_1, n_2\} \to \infty$.

From direct calculation, we can show that the moment generating function of $n_1\hat{a} = \sum_{i=1}^{c} x_i I\{y_i > 0\}$ satisfies

$$E[e^{\sum_{i=1}^{c} x_i t} I\{y_1 > 0, \ldots, y_c > 0\}] \leq E[e^{\sum_{i=1}^{c} x_i I\{y_i > 0\} t}] \leq E[e^{\sum_{i=1}^{c} x_i t}].$$

Note that the left term in the preceding inequality

$$\begin{aligned} E[e^{\sum_{i=1}^{c} x_i t} I\{y_1 > 0, \ldots, y_c > 0\}] &\geq [(1-a) + ae^t]^{n_1}[1 - \sum_{i=1}^{c}(1-q_i)^{n_2}] \\ &\to \left[(1-a) + ae^t\right]^{n_1} \end{aligned}$$

as $n_2 \to \infty$, and the right term also converges to the same limit. In other words, $n_1\hat{a}$ converges to a binomial random variable if n_2 is sufficiently large. Thus, $\hat{a}$ converges to a in probability and $\hat{a}$ is asymptotically normally distributed if $\min\{n_1, n_2\} \to \infty$. $\hat{b}$ behaves similarly, as does the joint distribution of $(\hat{a}, \hat{b})$. Since $(\hat{a}, \hat{b})$ are asymptotically jointly normally distributed, applying Cramer's delta theorem, $\hat{\theta}_n$ is also asymptotically normally distributed, and $\hat{\theta}_n$ converges to θ_n in probability.

4. Examples

In this section, two examples are used to compare the performance of different estimates of species overlap: one example was originally introduced in Abele (1979), and the other is from Chao (1995). Let $\hat{\theta}_J$ be the estimate of the Jaccard index, let $\hat{\theta}_b$ the estimate of θ_n proposed by Smith et al., and let $\hat{\theta}_n$ the NPMLE of θ_n.

Example 1. Abele describes the species abundance distribution of decapod crustacean (crab) communities at two locations in Panama (data shown in Smith et al.). Table 1 shows the estimate of Jaccard's index, the estimate

of Smith et al., and the NPMLE. Note that the estimate of Jaccard's index is usually calculated by plugging in the numbers of observed species, yielding $\hat{\theta}_J = 31/74 \approx 0.419$. The tabulated estimate of the species overlap proposed by Smith et al. is from their paper in 1996. The standard errors of the estimate for Jaccard's index, the estimator of Smith et al., and the NPMLE are from 1,000 bootstrap simulations. (Note that Smith et al. calculate the standard error of their estimator to be .100. This results from a different sampling model, as discussed in Section 6.)

As mentioned previously, the estimate by Smith et al. is about 50% larger than that of Jaccard's index. The NPMLE is also about 50% larger, but is slightly smaller than (and within 2 standard errors of) that of Smith et al. Also, the standard error of the NPMLE is the smallest among these estimates, and is about one-half of that for Jaccard's index. The standard error of the estimate of Smith et al. is the largest.

The variances of $\hat{a}$ and $\hat{b}$ can be estimated from (4) and (5), yielding 2.3015×10^{-4} and 3.3892×10^{-5}, respectively. Similarly, the covariance of $\hat{a}$ and $\hat{b}$ estimated from (6) equals 7.5614×10^{-6}. Applying the delta method, the standard error of $\hat{\theta}_n$ is approximately 0.01426, which is very close to the standard error obtained via bootstrapping (0.0150).

Table 1 Species Overlap Estimates

	Decapod crustaceans			Wild birds		
	$\hat{\theta}_J$	$\hat{\theta}_b$	$\hat{\theta}_n$	$\hat{\theta}_J$	$\hat{\theta}_b$	$\hat{\theta}_n$
Estimate	0.419	0.668	0.646	0.603	0.848	0.954
s.e.	0.028	0.046	0.015	0.016	0.024	0.008

Example 2. Chao (1995) and Chao et al. (2000) describe the species abun-

dance of wild bird communities at two heavily polluted river estuaries (the Ke-Yar River and the Chung-Kang River) of north-western Taiwan. Bird counts were collected by the Wild Bird Society of Hsin-Chu on a weekly basis for one year. Species overlap is of interest here because the two estuaries are environmentally similar. Table 2 shows the numbers of individuals observed for different species (with ranks representing different species) of birds in these two estuaries. The standard errors of all three estimates are from 1,000 bootstrap simulations. These and the estimates $\hat{\theta}_J$, $\hat{\theta}_n$, and $\hat{\theta}_b$ are listed in Table 1.

The NPMLE is 58% larger than that of Jaccard's index, and the estimate by Smith et al. is about 40% larger, about 12% smaller than the NPMLE. Similar to the previous example, the standard error of the NPMLE is the smallest and is one-half the size of the estimate of Jaccard's index. The standard error of the estimator of Smith et al. again is the largest, about three times that for the NPMLE, similar to the result in the previous example. However, based on the standard errors of $\hat{\theta}_n$ and $\hat{\theta}_b$, and since $\hat{\theta}_n$ is asymptotically unbiased, $\hat{\theta}_b$ appears to be significantly underbiased.

From (4), (5), and (6), we have $Var(\hat{a}) \doteq 7.4943 \times 10^{-5}$, $Var(\hat{b}) \doteq 3.0090 \times 10^{-7}$, and $Cov(\hat{a}, \hat{b}) \doteq 2.0103 \times 10^{-7}$. The standard error of $\hat{\theta}_n$ via the delta method thus equals 0.00855 and again is very close to that obtained via bootstrapping (0.0083).

5. Simulation

In this section, we use simulation to compare the performance of an estimate of Jaccard's index, the estimate proposed by Smith et al., and the NPMLE. Two types of species proportion distribution are considered: balanced and unbal-

anced. In a balanced population, every species is equally likely to be observed, while in an unbalanced population some species are dominant. In particular, we assume that the species proportions follow a geometric distribution, i.e. $p_i \propto \alpha^i$ for $1 \leq i \leq s_1$, and similarly for q_j. Computations and simulations in this report were based on a Pentium II– IBM compatible PC. The simulations were based on S-Plus statistical software, version 4.5.

We model three different types of species overlap between the two populations:

Type 1: The shared species are dominant in both populations;

Type 2: The shared species have low abundance in both populations;

Type 3: The shared species are dominant in one population, but have low abundance in the other.

When every species has the same species proportion, i.e. the balanced population case, these 3 types of overlap are the same. But in the unbalanced population case, if all other conditions are the same θ_n has its maximum value in the Type 1 case since common species are dominant in both populations. Note that, because the true population structure and type of overlap is known for these simulations, it is possible to calculate the true values of θ_J and θ_n, which we present in Table 3.

In addition to comparing the accuracy of the estimates to the real species overlap, we also use SD (standard deviation) to measure the precision of the estimates. We define

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\hat{\theta}}_i)^2}{n-1}}$$

where $\hat{\theta}_i$ is the estimate of species overlap in the ith simulation run, and $\bar{\hat{\theta}}_i$ is

the average of the estimates in n simulations. The results shown in this section are all based on 500 simulation runs.

Table 3(a) shows the simulation results for the balanced population cases (thus $\theta_J = \theta_n$), where both populations have 20 species, and 5 and 15 species are in common, respectively. In both cases, $\hat{\theta}_J$ appears to have the fastest convergence rate and is quite accurate even when the number of observations is small. The NPMLE appears to have the slowest convergence rate but is generally good compared to $\hat{\theta}_b$ since the overestimation by $\hat{\theta}_b$ is large when $n = 50$. When the number of observations is large, all three estimates perform well.

Tables 3(b) to 3(d) show the simulation results for unbalanced population cases with $s_1 = s_2 = 20$, and when the species proportions in each population are geometric with $\alpha = 0.8$. Table 3(b) shows the results of the Type 1 case, where the common species are dominant. The NPMLE and the estimate of Jaccard's index are both very accurate when the number of observations is large. The SD's of the NPMLE decrease in proportion to the inverse of the square root of the sample size, while the SD's of the estimate of Jaccard's index decrease much faster. The estimates of Smith et al. are significantly different from $\hat{\theta}_n$ and $\hat{\theta}_J$ even when the number of observations is 1000, and $\hat{\theta}_b$ is actually closer to θ_J, instead of θ_n. The SD's of $\hat{\theta}_b$ decrease faster than the inverse of the square root of the sample size when the number of common species is 5, but do not decrease monotonically as the sample size increases when the number of species is 15.

Similar patterns appear in the Type 2 and Type 3 cases as well. The NPMLE is the most accurate estimate of θ_n in both Type 2 (Table 3(c)) and Type 3 (Table 3(d)) cases. But unlike the Type 1 case where the species overlap is the largest, it takes more observations to have the NPMLE close to θ_n. The

SD's of the NPMLE are also the smallest among these three estimates. The estimate of Smith et al. is closer to θ_J in the Type 2 case, but does not seem to converge to θ_n or θ_J in the Type 3 case. Note that the SD's of $\hat{\theta}_b$ in the Type 3 cases do not decrease monotonically as the sample size increases, suggesting that θ_b may be slow to converge, if it converges at all.

6. Conclusion

Our paper deals with the notion of species overlap as defined by Smith et al. We find this definition appealing: it makes natural reference to a probabilistic interpretation that is intuitive and straightforward. In this paper we have compared a nonparametric maximum likelihood estimate of this quantity with the estimator of Smith et al. It is important to delineate the context in which this comparison has taken place.

We have focused on models for which the species proportions are fixed, similar to the notion of "fixed" populations defined by Engen (1978). This is different from the setting in Smith et al., which is based on a superpopulation model. Superpopulation models may be viewed in some sense as similar to the "random" populations of Engen, or, put more broadly, the comparison may be viewed as analogous to the difference between fixed and random effects in ANOVA. Although we have focused on fixed populations, we believe that fixed and superpopulation models are equally valid, and their specific use depends on the problem at hand. Engen notes a number of reasons for considering populations as fixed, and numerous authors use such models in the ecological setting. (See, for example, Engen, 1974, and the references cited therein.)

So, for example, consider an ecologist who wishes to study and compare two

populations that are relatively fixed in their composition at the time of sampling. The ecologist samples these by sampling individuals from them. The ecologist has at hand two possible estimators: the NPMLE and the estimator of Smith et al. How would the ecologist expect those two estimators to behave? Our paper provides some insights into this question. It is important to emphasize that, regardless of the genesis of the estimators, the comparison of the estimators remains valid. (To make a broad analogy, it is appropriate to ask what the frequentist properties are for a given Bayesian estimator, even if that estimator was not conceived with frequentist issues in mind.)

In this context, then, our proposed NPMLE generally provides a good estimate to the new index if the sample size is not too small. Also, $\hat{\theta}_n$ has good theoretical and empirical properties, and it is easy to compute, similar to the estimate of Jaccard's index. The variances of $\hat{\theta}_n$ derived from the delta method and bootstrapping are very close in the two examples shown. The decreasing rate of the SD's in $\hat{\theta}_n$ is approximately proportional to $1/\sqrt{(\text{Sample Size})}$ from our simulation.

The parametric estimate by Smith et al. was designed to estimate their new similarity index. However, their estimate seems to be strongly influenced (or dictated) by its parametric (species structure) assumption. If the parametric assumption can describe well the species structure, e.g., the balanced population case, $\hat{\theta}_b$ performs well when the sample size is fairly large. In particular, in the balanced population case and when the sample size is large, all species would have about the same number of occurrences. As a result, $d(j; n_i, \alpha, \beta, p, q)$ approximately equals p, q, and $1-p-q$ if species i appears only in population 1, 2, and both populations, respectively. Applying maximum likelihood estimation, we have, approximately, $\hat{p} = \hat{q} = (s-c)/(2s-c)$ and $\hat{\theta}_b = c/(2s-c) = \theta_J = \theta_n$

when $s_1 = s_2 = s$. This gives a heuristic sense of why $\hat{\theta}_b$ converges faster than $\hat{\theta}_n$ in the balanced population case of our simulation.

When the underlying species structure is not equiprobable, $\hat{\theta}_b$ may not be a good estimate of θ_n (or θ_J). For example, in the Type 1 case, the number of occurrences for the common species in our model follows a geometric pattern and $b(n_i; j, \alpha, \beta)$ would be small when n_i is large. Since fewer observations have a large weight, using the MLE would produce an inappropriate $\hat{\theta}_b$. Also, when the number of observations is very large, $d(j; n_i, \alpha, \beta, p, q)$ would have a form similar to that in the balanced population case. For example, suppose that $s_1 = s_2 = 2, c = 1$, and $\alpha = 0.8$ in the Type 1 unbalanced population case, and suppose the sample size is large. Again, arguing heuristically, we would expect that $\hat{\theta}_b = 1/3 (= \theta_J)$, instead of $\theta_n = 5/13$. This might explain why $\hat{\theta}_b$ is actually closer to the Jaccard index instead of the index proposed by Smith et al. Therefore, if the new index of overlap is to be used, we would recommend the NPMLE as its estimate.

Although it is not a specific focus of our study, we note that the plug-in estimate of Jaccard's index also performs well when the sample size is not too small. In particular, $\hat{\theta}_J$ appears to have the fastest convergence rate in the balanced population case, and is also quite comparable in unbalanced population cases when the sample size is big. Except in the Type 2 unbalanced population cases, the decreasing rate of SD's in $\hat{\theta}_J$ appears to be faster than $1/\sqrt{(\text{Sample Size})}$. Note that McCormick et al. (1992) proved the asymptotic normality of $\hat{\theta}_J$ in the balanced population case, but they require that the sample size and the number of species increase at the same rate.

In this paper, we modified the conditional probability definition of species overlap proposed by Smith et al. and generalized the setting to include the

Jaccard index as a special case. All three estimates of species overlap measures considered perform well in various situations, and we think that the NPMLE has properties that make it appealing. In particular, the NPMLE has good theoretical properties and is easy to compute. As a result, it provides a good estimate to the similarity index proposed by Smith et al. under the setting of fixed populations. However, although the overlap measure defined via the conditional probability setting can reduce the underbias of the Jaccard index in describing the species overlap, neither Smith et al. nor we provide a concrete plan for sampling species from two populations. In the future, we will continue working on searching a species overlap measure which not only reflect the true species overlap and has a probability interpretation, but also comes with a firm sampling plan.

ACKNOWLEDGEMENTS

The authors are grateful for the S-Plus program from Professor W. Smith, Taiwan wild bird data from Professor A. Chao, and insightful comments from the associate editor and an anonymous reviewer which helped us to clarify the context of our work.

REFERENCES

Abele, L. G. (1979). The Community Structure of Coral-associated Decapod Crustaceans in a Variable Environment. In *Ecological Processes in Coastal Marine Systems: Marine Science* **10**, 265-287, Florida State University. New York: Plenum Press.

Chao, A. (1995). How Many Classes? (in Chinese). *Communications in Mathematics* **19**, 1-7.

Chao, A., Hwang, W., Chen, Y., and Kuo, C. (2000). Estimating the Number of Shared Species in Two Communities. *Statistica Sinica* **10**, 227-246.

Engen, S. (1974). On Species Frequency Models. *Biometrika* **61**, 263-270.

Engen, S. (1978). *Stochastic Abundance Model*. London: Chapman and Hall.

McCormick, W. P., Lyons, N. I., and Hutcheson, K. (1992). Distributional Properties of Jaccard's Index of Similarity. *Communications in Statistics: Theory and Methods* **21**, 51-68.

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.