A SIMILARITY MEASURE BASED ON SPECIES PROPORTIONS[1]

Jack C. Yue and Murray K. Clayton

Department of Statistics and Department of Statistics

National Chengchi University and University of Wisconsin-Madison

Taipei, Taiwan, R.O.C. 11641 and Madison, WI. 53706, U.S.A.

csyue@nccu.edu.tw

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

ABSTRACT

There are several indices for measuring the similarity of two populations, including the ratio of the number of shared species to the number of distinct species (Jaccard's index) and the conditional probability of observing a shared species (Smith et al., (1)). However, these indices only take into account the number of species and species proportions of shared species. In this paper, we propose a new similarity index which includes the species proportions of both the shared and non-shared species in each population, and also propose a Nonparametric Maximum Likelihood Estimator (NPMLE) for this index. Bootstrap and delta methods are used to evaluate the standard errors of the NPMLE. Based on a loss function, we also compare a class of nonparametric estimators for the proposed index in various situations.

1. INTRODUCTION

The comparison, especially the similarity, of two populations or the evaluation of a population's change over time has been an widely studied topic in areas such as ecology, biology, and biogeography. For example, Abele (2) studies the similarity of decapod crustaceans

[1]This research was supported in part by a grant from National Science Council in Taiwan, NSC 90-2118-M-004-005.

(crabs) living in coral communities at two locations in Panama. Chao (3, 4) estimates the number of shared species of wild bird in two heavily polluted estuaries of northern Taiwan. In addition to these applications, there is an increasing demand in the design of good search engines, to determine how closely two Internet web sites or home pages are related.

Among all similarity indices, Jaccard's index is perhaps the most frequently used index to compare two communities. It is defined as the ratio of the number of shared species to the number of distinct species in two communities, i.e. Jaccard's index can be expressed as $\theta_J = s_0/(s_1 + s_2 - s_0)$, where s_i is the number of species in community i, $i = 1, 2$ and s_0 is the number of shared species for the two communities. Although the Jaccard index does consider the number of shared species and is easy to compute, the species proportions are not used and the similarity of two communities is likely to be under-estimated in real life examples (such as the two examples mentioned above).

Smith et al. (1) 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 index, called the community Jaccard index by Smith et al., can be defined as θ_n (Yue et al., (5)) where

$$\theta_n = \frac{(\sum_{i=1}^{s_0} p_i)(\sum_{i=1}^{s_0} q_i)}{\sum_{i=1}^{s_0} p_i + \sum_{i=1}^{s_0} q_i - (\sum_{i=1}^{s_0} p_i)(\sum_{i=1}^{s_0} q_i)},$$

and where p_i and q_i are the species proportions of species i in communities 1 and 2, and, without loss of generality, species 1 to s_0 are the shared species in both communities. This index takes into account the number of shared species and their proportions and has an appealing probabilistic interpretation. However, for this index the species proportions of all species are not considered fully in assessing the similarity of two communities and similar to the Jaccard index, the degree of similarity could be mis-judged. For example, let $s_1 = s_2 = 3$, $s_0 = 2$, $p_1 = .01 = q_2$, and $p_2 = .98 = q_1$. Since the dominant species in one population have low abundance in the other population, intuitively the similarity should not be large. But θ_n tends to give large values (for example, $\theta_n \approx .96$ in this example) when, in each community,

the probability of discovering shared species is large and the shared species proportions differ considerably.

In this paper, we propose a new similarity index which is a function of species proportions from both the shared and non-shared species. In addition, the shared species proportions in each community are compared one-to-one, and as a result, more weight is placed on those shared species which have similar species proportions in both communities.

2. THEORETICAL RESULTS

A natural and intuitive similarity index to consider would be $\sum_{i=1}^{s_0} p_i q_i$, which represents the probability of observing the same species in each population given that one observation is selected randomly from each population. (For a detailed discussion of measures of similarity and dissimilarity, see Gower, (6).) The range of this index is between 0 and 1. However, this index does not give the value 1 to cases where two populations have an identical species structure. For example, if $s_1 = s_2 = s_0$ and $p_i = q_i = 1/s_0$ for $i = 1, \ldots, s_0$, then the index is $1/s_0$. So, we divide $\sum_{i=1}^{s_0} p_i q_i$ by a normalizing constant. By noting the inequality $x^2 + y^2 - xy \geq xy$, we define the new similarity index to be

$$\theta = \frac{\sum_{i=1}^{s_0} p_i q_i}{\sum_{i=1}^{s_1} p_i^2 + \sum_{i=1}^{s_2} q_i^2 - \sum_{i=1}^{s_0} p_i q_i} = \frac{\sum_i p_i q_i}{\sum_i (p_i - q_i)^2 + \sum_i p_i q_i}. \tag{2.1}$$

Then $\theta = 1$ if $s_1 = s_2 = s_0$ and $p_i = q_i = 1/s_0$ for $i = 1, \ldots, s_0$, and θ is close to 0 if p_i and q_i are very different, or the probability of observing a shared species is small in at least one of the two communities (e.g. $\sum_{i=1}^{s_0} p_i$ or $\sum_{i=1}^{s_0} q_i$ is small).

When the species proportions of both populations are uniformly distributed, the Jaccard index, the community Jaccard index, and the proposed index have identical values. In other words, if $p_i = 1/s_1$ and $q_i = 1/s_2$, then

$$\theta = \frac{\sum_{i=1}^{s_0} \frac{1}{s_1}\frac{1}{s_2}}{\sum_{i=1}^{s_1}(\frac{1}{s_1})^2 + \sum_{i=1}^{s_2}(\frac{1}{s_2})^2 - \sum_{i=1}^{s_0} \frac{1}{s_1}\frac{1}{s_2}} = \frac{s_0}{s_1 + s_2 - s_0} = \theta_J$$

and

$$\theta_n = \frac{\left(\sum_{i=1}^{s_0} \frac{1}{s_1}\right)\left(\sum_{i=1}^{s_0} \frac{1}{s_2}\right)}{\sum_{i=1}^{s_0} 1/s_1 + \sum_{i=1}^{s_0} 1/s_2 - (\sum_{i=1}^{s_0} \frac{1}{s_1})(\sum_{i=1}^{s_0} \frac{1}{s_2})} = \frac{s_0}{s_1 + s_2 - s_0} = \theta_J.$$

Note that the proposed index is closely related to Morisita's index (Krebs, (7)), defined as

$$\theta_{MH} = \frac{2\sum_{i=1}^{s_0} p_i q_i}{\sum_{i=1}^{s_1} p_i^2 + \sum_{i=1}^{s_2} q_i^2}.$$

But these two indices have different upper bounds for the summations in the denominator and have a slightly different meaning. For example, let $p_i = 1/s_1$ for $i = 1, 2, \ldots, s_1$ and $q_j = 1/s_2$ for $j = 1, 2, \ldots, s_2$. Also, without loss of generality, let $s_1 \geq s_2$. Then

$$\theta = \frac{s_2}{s_1} \leq \theta_{MH} = \frac{2 \times s_2/s_1}{1 + (s_2/s_1)^2}.$$

Suppose that $s_1 = 2s_2$, i.e., the number of species in population 2 is half of that in population 1. In this case, $\theta = 0.5$ and $\theta_{MH} = 0.8$. Because half of the species in population 1 cannot be found in population 2, a similarity value of $\theta = 0.5$ seems more appropriate.

Let $a = \sum_i p_i^2$, $b = \sum_i q_i^2$, and $d = \sum_i p_i q_i$. Then the proposed index can be written as the form:

$$\theta = \frac{d}{a + b - d}. \tag{2.2}$$

As mentioned in Engen (8, 9), since θ is not a linear function of p_i's and q_j's, unbiased estimators likely do not exist. (Or if they exist, their forms are likely to be very complicated.) One possible solution is to plug in the maximum likelihood estimators for the p_i's and q_j's, i.e.,

$$\hat{a} = \sum_{i=1}^{s_1} \left(\frac{x_i}{n_1}\right)^2, \quad \hat{b} = \sum_{i=1}^{s_2} \left(\frac{y_i}{n_2}\right)^2, \quad \text{and} \quad \hat{d} = \sum_{i=1}^{s_0} \left(\frac{x_i}{n_1}\right)\left(\frac{y_i}{n_2}\right), \tag{2.3}$$

where x_i and y_i are the numbers of occurrences for the ith species from n_1 and n_2 observations taken in communities 1 and 2, respectively. Hence $\hat{\theta} = \hat{d}/(\hat{a} + \hat{b} - \hat{d})$, where $\hat{\theta}$, $\hat{a}, \hat{b}$, and $\hat{d}$ can be referred to as Nonparametric Maximum Likelihood Estimators (NPMLE).

It is obvious that $E(\hat{a}) \to a$ and $E(\hat{b}) \to b$, as $\min\{n_1, n_2\} \to \infty$, since

$$\begin{aligned} E(\hat{a}) &= E\left[\sum_{i=1}^{s_1} \frac{x_i(x_i-1)}{n_1^2}\right] + E\left[\sum_{i=1}^{s_1} \frac{x_i}{n_1^2}\right] \\ &= \frac{n_1-1}{n_1}\sum_{i=1}^{s_1} p_i^2 + \frac{1}{n_1} = \frac{n_1-1}{n_1}a + \frac{1}{n_1}, \end{aligned}$$

and $E(\hat{b}) = (n_2-1)b/n_2 + 1/n_2$. We can also show, using a similar argument, that $\hat{a} \to a$ and $\hat{b} \to b$ in probability, given that $\min\{n_1, n_2\} \to \infty$. Similarly, we can show that $E(\hat{d}) \to d$ and $\hat{d} \to d$ in probability as $\min\{n_1, n_2\} \to \infty$. Thus, we have that $\hat{\theta} \to \theta$ in probability and so $E(\hat{\theta}) \to \theta$ as $\min\{n_1, n_2\} \to \infty$.

Applying Cramer's delta theorem, it is obvious that $\hat{\theta}$ is asymptotically normally distributed. The asymptotic variance can be derived via the delta method, and approximately

$$\begin{aligned} Var(\hat{\theta}) \approx\ & \frac{\hat{d}^2}{(\hat{a}+\hat{b}-\hat{d})^4}\left[Var(\hat{a}) + Var(\hat{b})\right] + \frac{(\hat{a}+\hat{b})^2}{(\hat{a}+\hat{b}-\hat{d})^4} Var(\hat{d}) \\ & -2\frac{(\hat{a}+\hat{b})\hat{d}}{(\hat{a}+\hat{b}-\hat{d})^4}\left[Cov(\hat{a},\hat{d}) + Cov(\hat{b},\hat{d})\right], \end{aligned} \tag{2.4}$$

where

$$\begin{cases} Var(\hat{a}) & \approx \frac{4}{n_1}\left[\sum_{i=1}^{s_1} p_i^3 - a^2\right] \\ Var(\hat{b}) & \approx \frac{4}{n_2}\left[\sum_{i=1}^{s_1} q_i^3 - b^2\right] \\ Var(\hat{d}) & \approx \frac{1}{n_1}\sum_i p_i q_i^2 + \frac{1}{n_2}\sum_i p_i^2 q_i - (\frac{1}{n_1}+\frac{1}{n_2})d^2 \\ Cov(\hat{a},\hat{d}) & \approx \frac{2}{n_1}\left[\sum_i p_i^2 q_i - ad\right] \\ Cov(\hat{b},\hat{d}) & \approx \frac{2}{n_2}\left[\sum_i p_i q_i^2 - bd\right]. \end{cases}$$

3. EXAMPLES AND SIMULATION

In this section, we use two examples and simulation to evaluate the performance of the NPMLE. The first example involves decapod crustacean (crab) communities at two locations in Panama (data shown in Smith et al., (1)), studied by Abele (2). The second example is based on wild bird communities at two locations of north-western Taiwan (data shown in Yue et al., (5)), studied by Chao (3). The probabilities of observing shared species in each population (i.e., $\sum_{i=1}^{s_0} p_i$ and $\sum_{i=1}^{s_0} q_i$), Simpson's index (i.e., $\sum p_i^2$ and $\sum q_i^2$), and Shannon's

index (entropy, defines as $-\sum p_i \log(p_i)$ and $-\sum q_i \log(q_i)$) for each community are listed on Table 1.1 and the species proportions (in squared root scale and descending order) of the two communities are plotted in Figures 1.1 and 1.2. We can use this information to judge how similar the two communities are. For the crab data, both the probabilities of shared species and the Simpson indices indicate that these two communities are very different. (The Shannon indices show some differences.) This can also be seen from Figure 1.1, where dominant species in one community often have low abundance in the other community. We would expect that the value of an appropriate similarity index for these two communities would not be too large. On the other hand, all three indices for the two communities and the species proportions graph (Figure 1.2) in the case of the wild bird data are very similar, and the value of an appropriate similarity index should be fairly large.

However, the Jaccard index and the index proposed by Smith et al. do not coincide with the information in Table 1.1, Figure 1.1, and Figure 1.2. The estimates of three indices and their standard errors (calculated via bootstrapping) are given in Table 1.2. The Jaccard index gives a smaller (instead of large) similarity value in the wild bird case, while the Jaccard community index of Smith et al. gives a larger (instead of small) similarity value in the crab case.

As a check of the variance computed via bootstrapping of the NPMLE for θ, we also compute the approximate variances (via delta method) of the new index from (2.4) and compare them to those from bootstrapping. The standard errors computed via the delta method for the two examples are .0234 and .0030 for the crab and wild bird cases, respectively, which are very close to those computed via bootstrapping (.0233 and .0029).

For the simulation study, we consider two types of species proportion distribution: balanced and unbalanced. In a balanced population, every species is equally likely to be observed, while in an unbalanced population some species are dominant. In particular, for the latter we assume that the species proportions follow a geometric distribution, i.e. $p_i \propto \alpha^i$ for

$s_1 \geq i \geq 1$, 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 2000.

Figure 1.1 Crab Data:$\sqrt{\text{Species Prop.}}$ Figure 1.2 Bird Data:$\sqrt{\text{Species Prop.}}$

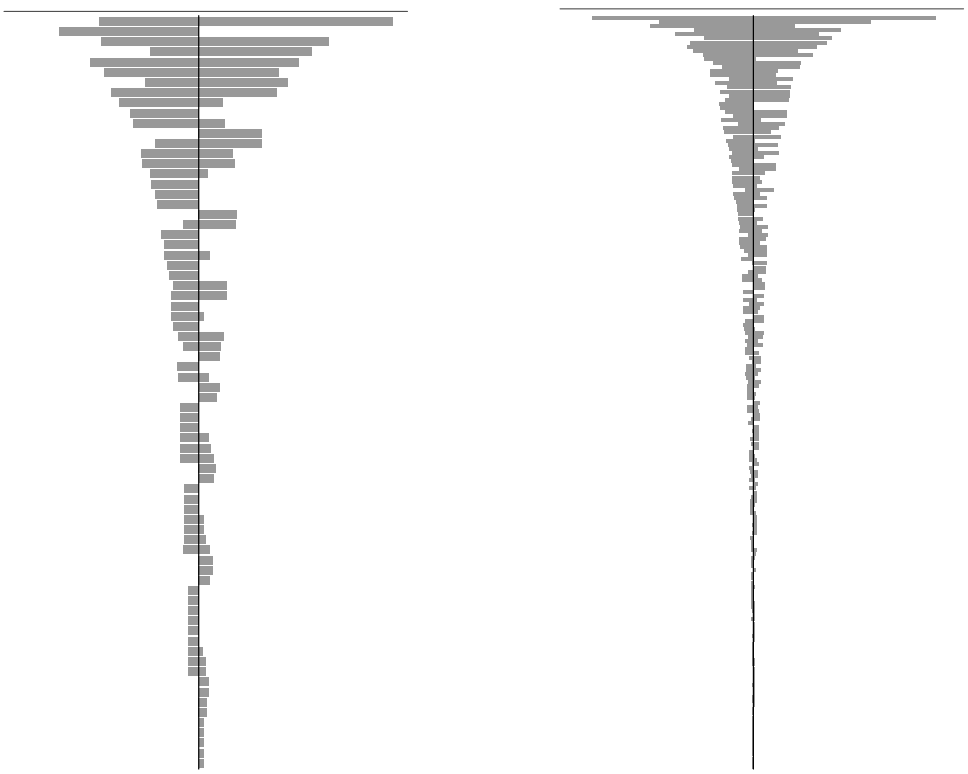

Table 1.1 Species structures of two examples

	Decapod crustaceans			Wild birds		
	Shared prob.	Simpson index	Shannon index	Shared prob.	Simpson index	Shannon index
community 1	.6784	.0693	3.0948	.9664	.0910	3.1917
community 2	.9308	.1518	2.3943	.9917	.1329	2.8058

Table 1.2 Species Overlap Estimates

	Decapod crustaceans			Wild birds		
	Jaccard	Smith	New	Jaccard	Smith	New
Index	(θ_J)	(θ_n)	(θ)	(θ_J)	(θ_n)	(θ)
Estimate	.419	.646	.359	.603	.954	.840
s.e.	.028	.015	.023	.016	.008	.003

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.

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 1,000 simulation runs.

Table 2 shows the simulation results for balanced and unbalanced cases. We assume that both populations have 20 species, and 5 or 15 species are in common, respectively. The true values of θ and the values of the Jaccard index and the index proposed by Smith et al. are also listed in Table 2 for reference. In all cases, the difference between $\hat{\theta}$ and θ is within 2 times the standard deviation when the sample size reaches 300, and the differences between $\hat{\theta}$ and θ are smaller as the sample size increases. In other words, there are good agreements between the NPMLE's and population values in both the balanced and unbalanced population cases,

provided that there are sufficiently many observations. Even in the type 3 case, where the species structures of the shared species differ a lot, the NPMLE still performs fairly well. Also, the standard deviation of $\hat{\theta}$ decreases monotonically as the sample size increases.

Table 2 Estimates ($\hat{\theta}$) for varying sample sizes, the number of shared species, and species structures, given that $s_1 = s_2 = 20$ and $\alpha = 0.8$. Values in parentheses represent SD.

	$s_0 = 5$ ($\theta_J = .1429$)			
n	Balanced	Type 1	Type 2	Type 3
100	.1178(.0345)	.7082(.0639)	.000489(.000639)	.0150(.0112)
200	.1272(.0263)	.7540(.0425)	.000535(.000430)	.0158(.0076)
300	.1333(.0231)	.7704(.0333)	.000551(.000345)	.0152(.0065)
400	.1357(.0191)	.7773(.0284)	.000544(.000290)	.0157(.0055)
500	.1369(.0173)	.7841(.0250)	.000551(.000257)	.0156(.0048)
1000	.1402(.0128)	.7954(.0178)	.000560(.000172)	.0160(.0036)
θ	.1429	.8062	.000553	.0159
θ_n	.1429	.5153	.0121	.0237

	$s_0 = 15$ ($\theta_J = .6$)			
n	Balanced	Type 1	Type 2	Type 3
100	.4663(.0715)	.8666(.0568)	.0531(.0168)	.1805(.0446)
200	.5208(.0511)	.9264(.0323)	.0551(.0127)	.1879(.0324)
300	.5460(.0443)	.9488(.0222)	.0559(.0102)	.1922(.0278)
400	.5580(.0373)	.9595(.0184)	.0558(.0088)	.1924(.0242)
500	.5662(.0348)	.9675(.0139)	.0558(.0080)	.1925(.0221)
1000	.5837(.0245)	.9831(.0071)	.0562(.0056)	.1937(.0155)
θ	.6	.9977	.0567	.1883
θ_n	.6	.9533	.1904	.3173

4. A CLASS OF NONPARAMETRIC ESTIMATORS

The NPMLE of the similarity index defined in (2.1) is a result of directly plugging in the MLE's of a, b, and d in (2.3). Like most MLE's, $\hat{a}$ and $\hat{b}$ are biased (and yet they are consistent). Although we have shown that the NPMLE $\hat{\theta}$ is consistent, it would be interesting to see if it would reduce the bias or the variance if we plugged in unbiased estimators of a and b, i.e., $\tilde{a} = \sum x_i(x_i-1)/n_1(n_1-1)$ and $\tilde{b} = \sum y_i(y_i-1)/n_2(n_2-1)$, (see, for example, Engen, (8)). In this section, we use simulation to study the performance of plugging in different estimators of a and b. In particular, we consider

$$\hat{a}_c = \sum \frac{x_i(x_i-c)}{n_1(n_1-c)} \quad \text{and} \quad \hat{b}_c = \sum \frac{y_i(y_i-c)}{n_2(n_2-c)}, \tag{4.1}$$

for $c = 0$ and 1, i.e. NPMLE and "unbiased" estimators, and define

$$\hat{\theta}_c = \frac{\hat{d}_c}{\hat{a}_c + \hat{b}_c - \hat{d}_c}. \tag{4.2}$$

Also, in order to evaluate the performance of these estimators, we consider the loss function

$$L(\hat{\theta}_c, \theta) = Var(\hat{\theta}_c) + E(\hat{\theta}_c - \theta)^2,$$

which includes both the variance and (bias)^2 for the estimator $\hat{\theta}_c$. The variance and (bias)^2 of each estimator are defined as

$$Var(\hat{\theta}) = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\hat{\theta}})^2,$$

and

$$E(\hat{\theta}_i - \theta)^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2,$$

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.

Similar to the previous section, we consider three types of species structure for both the balanced and unbalanced situations. The number of species in each population is 20, and the number of shared species is 5 or 15. Table 3 shows the simulation results, with

each number representing the result of 1,000 replications. Note that in our simulation, the numbers of observations taken from each population are 200, 300, 400, 500, 600, 800, 1000, 1500, and 2000. Since the simulation results for different sample sizes show identical patterns, in Table 3, we only use the case where 1000 observations are taken from each population as a demonstration.

Briefly, the results suggest using $\hat{\theta}_0$ if the shared species have low abundance in either of the populations. If there is no information about the species abundance, we suggest using $\hat{\theta}_1$ since its loss is generally smaller than or similar to that of $\hat{\theta}_0$.

Table 3 Simulated loss ratio of $\hat{\theta}_1$ to $\hat{\theta}_0$

	Balanced	Type 1	Type 2	Type 3
$s_0 = 5$	1.0064	0.5231	1.0525	1.0532
$s_0 = 15$	0.5186	0.2124	1.0677	1.1681

It happens that the simulated loss of $\hat{\theta}_1$ comes mainly from the variance, since the bias is almost negligible. On the other hand, the bias of $\hat{\theta}_0$ is not always negligible (such as in Type 1 and balanced cases) but $\hat{\theta}_0$ has a smaller variance in general.

5. CONCLUSION

In this paper, we proposed a new overlap index for measuring the similarity of two populations. It is founded on a probabilistic interpretation, scaled to represent a measure between 0 and 1. Unlike previously described measures, this index takes fuller account of all the species proportion information. However, when the populations are uniform, it provides a measure of overlap identical to the measures of Jaccard or Smith et al.

The NPMLE of the proposed index makes the proposed index even more appealing, since the NPMLE has theoretically and empirically sound properties and is easy to compute. Theoretically, it has good asymptotic properties, including consistency and asymptotic normality. In addition, in our examples, we found that the standard errors of the NPMLE computed via

formula in (2.4) and bootstrapping are very close, and in practice we can use either method. The NPMLE has good accuracy in estimating the true overlap as well. From our simulations, we found that the NPMLE of the proposed index also performs well, whether the distribution is balanced or unbalanced. In particular, in the Type 2 unbalanced case where the shared species have low abundance in both populations, the NPMLE is well within the error range when the sample size is moderately large, say 300.

An open question which we are currently exploring is the definition and estimation of overlap when several populations are at hand. We believe that the measure introduced in this paper can be extended in a natural way to the consideration of more than two populations, and we are currently working on developing those ideas.

ACKNOWLEDGMENTS

The authors are grateful to Professor A. Chao for the Taiwan wild bird data and insightful comments from an anonymous reviewer which helped us to clarify the context of our work.

BIBLIOGRAPHY

(1) Smith, W., Solow, A. R., and Preston, P. E. An Estimator of Species Overlap Using a Modified Beta-binomial Model. Biometrics, **1996**, *52*, 1472-1477.

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

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

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

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

(6) Gower, J. C. 1985. Measures of Similarity, Dissimilarity, and Distance. Pages 399-405 in S. Kotz and N. L. Johnson, editors. Encyclopedia of Statistical Sciences **1985**, *5*, Wiley, New York.

(7) Krebs, C. J. Ecological methodology. **1999**, Menlo Park, California.

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

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