PN-ABK-245
75280

WORKING PAPER SERIES

FISHERIES STOCK ASSESSMENT

TITLE XII
Collaborative Research Support Program

Fisheries Stock Assessment CRSP Management Office
International Programs, College of Agriculture
The University of Maryland, College Park, Maryland 20742

In cooperation with the United States Agency for International Development (Grant No. DAN-4146-G-SS-5071-00) the Fisheries Stock Assessment CRSP involves the following participating institutions:

The University of Maryland—Center for Environmental and Estuarine Studies
The University of Rhode Island—International Center for Marine Resource Development
The University of Washington—Center for Quantitative Sciences
The University of Costa Rica—Centro de Investigación en Ciencias del Mar y Limnología
The University of the Philippines—Marine Science Institute (Diliman)—College of Fisheries (Visayas)

In collaboration with The University of Delaware; The University of Maryland—College of Business and Management; The University of Miami; and The International Center for Living Aquatic Resources Management (ICLARM).

WORKING PAPER SERIES

Working Pap r No. 58

COMPARISONS BETWEEN MULTISPECIES ASSEMBLAGES USING SIMILARITY IND

by
Saul B. Saila and James E. McK
Graduate School of Oceanograp
The University of Rhode Isla

August 1989

The Fisheries Stock Assessment CRSP (sponsored by part by USAID Grant No. DAN-4146-G-SS-5071-00) is intended to support collaborative research between U.S. and developing countries' universities and institutions on fisheries stock assessment and management strategies.

This Working Paper has been produced by The University of Rhode Island research component in collaboration with The University of the Philippines and in association with the International Center for Living Aquatic Resources Management. Additional copies are also available from:

International Center for Marine Resource Development
The University of Rhode Island
126 Woodward Hall
Kingston, Rhode Island 02881

Comparisons Between Multispecies Fish Assemblages Using Similarity Indices

Saul B. Saila
and
James E. McKenna
Graduate School of Oceanography
The University of Rhode Island
Narragansett, RI 02882-1197

General—

Various measures of similarity have been used in making pairwise comparisons between ecological communities, and they have also been frequently used as elements of a matrix of input coefficients for clustering algorithms. Therefore, we consider it useful to consider similarity indices and their potential utility in the context of applications to empirical studies related to tropical multispecies fisheries.

The basic definition of a similarity index as used herein is restricted to a single number, which is a function of the pairwise comparison of values for each attribute for two samples. This use of the word similarity is not construed to exclude indices of dissimilarity, since one is the logical complement of the other. That is, similarity indices relate how "close" two samples are to each other, and dissimilarity indices refer to how "far apart" they are. For the purposes of this report, only those indices which are intermediate similarity scores defined on the pairwise comparisons of the values for each assemblage attribute in the two samples will be considered. This excludes any type of index which only summarizes the attribute values for each sample into a single index and then makes comparisons between these indices. Thus, this definition excludes so-called diversity or community structure indices as defined by Pielkham and Pearson (1976) and excludes the diversity indices reviewed by Washington (1984). The usage of the α diversity index in multispecies fisheries applications is considered in another report.

As shown by Johnston (1976), a general similarity index can be defined on two N-dimensional vectors, say $\vec{X}_l$. and $\vec{X}_{l'}$ where l and l' are two selections of location by time (ij) assemblage samples. The subscripts for the vectors are changed from ij to l in order to avoid a confusing array of subscripts. We then have:

$$S_{ll'} = F(s_{ll'k}), \tag{1}$$

where F is some function of the N pairwise comparisons, $s_{ll'k}$, and:

$$s_{ll'k} = g(x_{lk}, x_{l'k}) \tag{2}$$

where g is some function of the attribute values observed in the two samples.

From Equation (1) and Equation (2), any particular similarity index ($S_{ll'}$) can be defined by specifying the functions g and F.

In summary, a similarity index is characterized as the result of a two-step process defined on a pair of vectors. In the first step, an attribute similarity score is obtained for each attribute by comparing the attribute values observed in the pair of vectors. The result is a vector of attribute similarity scores. These are combined in the second step to arrive at a similarity index. The operation in the first step was characterized as a function (g) defined on pairs of attribute values. The second operation was characterized as a function (F) defined on the vector of attribute similarity scores from the first step. In most instances, F is a simple score or weighted score of attribute similarity scores.

A number of similarity indices ($S_{ll'}$) have been based on a transformation of the original data to presence-absence form by:

$$\delta_{lk} = \begin{cases} 1 \text{ if } x_{lk} > 0 \\ 0 \text{ if } x_{lk} = 0 \end{cases}$$

The data matrix in this case, or in the case where binary data are collected, will consist of vectors of 0's and 1's. Another form of transformation is to reduce individual counts and continuous data to more than two categories by defining cells. These can be of

equal, geometric, or logarithmic intervals. The data are then classified into these cells. This transformation is of the form:

$$C_{l,k} = \left| \begin{array}{l} 1 \text{ if } 0 \leq x_{lk} < C_1 \\ 2 \text{ if } C_2 \leq x_{lk} < C_2 \\ \vdots \\ m \text{ if } C_{m-1} \leq x_{lk} < C_m \end{array} \right.$$

The choice of a particular measure or index of similarity may influence the interpretation of the original data (Czarnecki 1979) in a significant manner. Different indices appear to be useful in indicating different changes in community composition. An example includes the fact that one kind of index may be useful for detecting changes in numerical abundance while another may be better at indicating changes in proportional abundance. In addition to the choice of similarity index, there is another and probably more important problem. This concerns the determination of what value of a given similarity index is important for biological interpretation.

Many users of similarity indices have not taken sampling variability into account. Some suggestions for estimating the variance of a similarity index have been made, but it remained for Smith, Genter, and Cairns (1986) to present two alternate methods for estimating confidence intervals for comparing two multispecies groups (assemblages) that provide an accurate estimate of variability. The methods described above are appropriate for most measures of similarity computed from species counts, species biomass, or species density. This work has been extended by Nemec and Brinkhurst (1988) to the comparison of dendograms. Further description and methods for comparing dendograms will appear in another report.

The purpose of this report is to briefly explain and illustrate the methods for estimating confidence intervals for generalized similarity indices as described by Smith et al. (1986), to illustrate these methods with an example applied to a tropical multispecies

fishery, and to provide documentation and a program for computing a measure of similarity with confidence intervals for similarity comparisons.

Specific Similarity Indices—

Johnston (1976) indicated that more than 50 similarity indices have been proposed or used. He investigated the characteristics of 25 similarity indices used in studies of ecological communities. Other similarity measures, such as PS_1, which is a measure of proportional similarity, has been recommended by Pielou (1978) and Kohn and Riggs (1982). Their reasons for this recommendation appear to be related to the fact that PS_1 is relatively independent of sample size. Washington (1984) reviewed diversity, biotic, and similarity indices. However, his review included only five similarity indices used in aquatic systems. Heltshe (1988) has derived an exact expression for the two-sample jackknife estimate of the simple matching coefficient and the Jaccard index of similarity when quadrat sampling procedures are employed. He demonstrated that these estimators were not affected by quadrat size as long as the sampling area was fixed. In this report we confine our attention to the use of only one index which we believe has utility for the purposes at hand. However, we emphasize that other published indices are appropriate for the methodology to be demonstrated herein, but we strongly suggest using indices which are not affected by sample size (group size).

In terms of an example, our attention will first be focused on SIMI, a similarity measure proposed by Stander (1970) and illustrated by Smith et al. (1986). Another index of similarity (I) described initially by Mountford (1962) is based on the logarithmic series distribution and the assumption that the samples are drawn from the same assemblage (community). Mountford's index has some desirable properties for certain applications, and these are described in another report. The same holds true for HR, the heterogeneity ratio introduced by Kobayashi (1987)

Brock (1977) made a critical comparison of two community similarity indices and concluded that investigators should not rely on one index. Instead, he stated that they

should consider several which have different areas of sensitivity. Kobayashi (1987) studied fifteen previously used similarity indices for the effects of variable sample size. He found only three (NESS-Grassle and Smith 1976; C_λ - Morisita 1957; and $C_{\lambda'}$ - Morisita 1971) to be relatively independent of sample size. He also described a new index, heterogeneity ratio (HR), which was sample size independent but assumes the samples were taken from a community (assemblage) described by the log series distribution.

In attempting to summarize the results of the above-mentioned studies, the following statements and suggestions are considered useful to users of similarity indices.

1) Most similarity indices tend to ignore both the time series and response pattern information in the raw data. They are usually defined on an arbitrary pair of sample vectors. These vectors are identifiable as representing particular time and location samples. The samples may be selected for comparison by a similarity index to reflect a potential change between two time periods at the same location or a difference between two locations at the same time period. However, many indices are algebraically symmetric, so the fact that one was taken before the other has no impact on the calculation. The functions assigning pairwise attribute similarity scores are symmetric, so that:

$$g(x_{lk}, x_{l'k}) = g(x_{l'k}, x_{lk})$$

2) Some similarity indices are relatively unaffected by differences in sample size. Such indices include Morisita's C_λ and $C_{\lambda'}$, Grassle and Smith's NESS, Kobayashi's HR and Mountford's I, among others. The above indices are suggested for future application and further consideration, because they have satisfied one important criterion since sample size is not necessarily fixed in the case of some fishery surveys, and comparative studies often involve variable sample sizes.

3) Greater general use of jackknife and bootstrap methods for estimating confidence intervals is also recommended for the following reasons. Precise variance estimates can be developed which need not be symmetric and which yield accurate confidence

intervals. Equal numbers of replicates are not required. However, some replication is required for both jackknife and bootstrap methods to be applied successfully.

4) Jackknife and bootstrap methods may be used with virtually any data and similarity index as long as it is replicated. However, see item 2) above. When sample sizes are small and variability is high, the jackknife seems to perform better than the bootstrap.

5) The variance of a similarity measure (index) depends on the number of replicates and the number of organisms counted per replicate. When variability between replicates is the largest component of variability, a large number of replicates is required to provide reliable estimates of similarity measures.

6) When similarities are close to 0.0 or to 1.0, the distributions of the various estimates may be skewed. Skewness does not seem to be a problem if values of similarity range from 0.1 - 0.9, unless the sample sizes are small (< 6).

Approach and Example—

For the purposes at hand, assume that there are two multispecies fish assemblages with m_i (i = 1,2) replicate samples (trawl hauls, trap lifts, gillnet sets, etc.) from each and T species or species groups are considered. The analysis of the trawl hauls results in data of the form n_{ijk} = the abundance of species k in replicate j of community i.

Let $n_{ij\cdot} = \sum_k n_{ijk}$ be the total count for the jth replicate of the assemblage i. Let $N_i = \sum_j n_{ij\cdot}$ be the total count of individuals in the sample from assemblage i and $n_{i\cdot k}$ be the total number of species k in the sample from community i. The general data that arise in a specific experimental comparison are shown in Table 1.

The estimates of SIMI for each pair of assemblages is based on the total count vectors $n_1 = (n_{i\cdot 1}, n_{1\cdot 2}, \ldots n_{i\cdot T})$, the abundance of each species summed over all replicates. Proportions, $P_{i\cdot k} = \frac{n_{i\cdot k}}{N_i}$ and SIMI are estimated as:

Table 1. An example of the form data to be collected for two multispecies fish assemblages—which could have been separated by space or time.

	Taxa for Assemblage 1				
Replicate	1	2	...	T	Totals
1	m_{111}	m_{112}	...	n_{11T}	$n_{11.}$
2	m_{121}	n_{122}	...	n_{12T}	$n_{12.}$
3	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
m_1	n_1m_11	n_1m_12	...	$n_1m_1{}^T$	$n_1m_{1.}$
Totals	$n_{1.1}$	$n_{1.2}$		$n_{1.T}$	N_1

	Taxa for Assemblage 2				
Replicate	1	2	...	T	Totals
1	m_{211}	m_{212}	...	n_{21T}	$n_{21.}$
2	m_{221}	n_{222}	...	n_{22T}	$n_{22.}$
3	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
m_2	n_2m_21	n_1m_22	...	$n_2m_2{}^T$	$n_2m_{2.}$
Totals	$n_{2.1}$	$n_{2.2}$	...	$n_{2.T}$	n_2

$$SIMI = \Sigma P_{1k} P_{2k} / \sqrt{\Sigma P^2_{1k} \Sigma P^2_{2k}}$$

by

$$\hat{SIMI} = \Sigma p_{1k} p_{2k} / \sqrt{\Sigma p^2_{1k} \Sigma p^2_{2k}}$$

Capital letters indicate true proportions and lower case p's indicate the estimates from the sample data.

It is clear that an estimate of the mean score SIMI is possible from Table 1 data—as well as a measure of variance. However, the jackknife and bootstrap are recommended alternate methods, because they don't depend upon use of the multinomial distribution, and the bootstrap can produce asymmetric confidence intervals.

Using the notation of Smith et al. (1986), we let $G = G(\vec{P_1}, \vec{P_2})$ denote a general measure of similarity where $\vec{P_1} = (P_{i1}, \ldots P_{iT})$ is the vector of proportional abundances for community i and $P_{i \cdot k}$ refers to the proportion of species k in community i. The estimate of similarity (G) is based on n_{ijk}, the abundance of species k in replicate j for community (assemblage) i. The estimate of $P_{i \cdot k}$ is:

$$p_{ik} = \sum_{j=1}^{m_i} n_{ijk} / \sum_{k=1}^{T} \sum_{j=1}^{m_i} n_{ijk} = n_{ik}/N_i \qquad (3)$$

where N_i is the total count for community (assemblage) i. To estimate the similarity between communities (G), use

$$\hat{G} = G(p_1, p_2)$$

Jackknife—The two sample jackknife estimate is determined by calculating G with different replicates removed each time. Let G'_{-k} denote the estimate of G when the kth replicate is removed from community 1 and G''_{-k} denote the estimate where the kth replicate is removed from community 2. Pseudo-values are then generated by sequentially removing replicates from community 1 and computing

$$\hat{G}_k' = (m_1 - 1/2)\ \hat{G} - (m_1 - 1)\ \hat{G}'_{-k} \tag{4}$$

and for community 2:

$$\hat{G}_k'' = (m_2 - 1/2)\ \hat{G} - (m_2 - 1)\ \hat{G}''_{-k} \tag{5}$$

The jackknife estimate of G is calculated based on these pseudo-values as:

$$\hat{G}_J = 1/m_1 \sum_{k=1}^{m_i} G_k' + 1/m_2 \sum_{k=1}^{m_2} \hat{G}_k'' \tag{6}$$

The estimate of the variance is:

$$s^2(\hat{G}_J) = 1/m_1\ \hat{\sigma}_1^2 + 1/m_2\ \hat{\sigma}_2^2 \tag{7}$$

where:

$$\hat{\sigma}_1^2 = 1/(m_1-1) \sum_{k=1}^{m_i} (\hat{G}'_k - 1/m_1 \sum \hat{G}'_k)^2 \tag{8}$$

and $\hat{\sigma}_2^2$ is defined similarly. Construction of the jackknife estimate in this way automatically adjusts for bias.

Bootstrap—This method estimates the sampling distribution of G by repeatedly sampling with replacement from the actual data. The bootstrap proceeds as follows:

a) Using the data, sample with replacement m_1 replicates from the m_1 replications for community 1 and sample with replacement m_2 replicates from the m_2 replications for community 2 (i.e., sample from the original data, with replacement). This sample is called the bootstrap sample.

b) Compute the bootstrap estimate of G based on the bootstrap sample which is denoted $\hat{G}_{(i)}$.

c) Repeat a) and b) b times to get estimates of G.

d) Calculate the bootstrap estimate of G and its sampling variance, i.e.,

$$\hat{G}_B = \frac{1}{b} \sum_{i=1}^{b} \hat{G}_{(i)} \tag{9}$$

and

$$s^2_B = (b-1)^{-1} \sum_{i=1}^{b} (\hat{G}_{(i)} - G_{(B)})^2 \tag{10}$$

It has been suggested that b values lie in the range of 50-200 for estimating the variance. For confidence intervals, b should be ~ 1000. The bootstrap method also gives an estimate of the bias in estimating $\hat{G}$ · Bias $(\hat{G}) = \hat{G}_B - \hat{G}$. To adjust for bias, form:

$$\hat{G}^* = \hat{G} \cdot \text{Bias} - (\hat{G}) = 2\ G - G_B \tag{11}$$

Confidence Intervals—

For the jackknife, confidence intervals are symmetric and are given by:

$$SIMI_J \pm t_c \cdot s\,(SIMI_J) \tag{12}$$

where t_c is the critical value from the t table with $m_1 + m_2 - 2$ degrees of freedom and confidence level $1 - \alpha$.

Symmetric or asymmetric intervals may be computed for the bootstrap methods described by Smith (1985). Due to the simplicity of the derivatives of the proportional similarity measures:

$$PS = \sum_{j=1}^{r} \min(P_{1j}, P_{2j}) = 1 - 1/2 \sum_{j=1}^{r} \left| P_{1j} - P_{2j} \right| \tag{13}$$

This formula can be reduced if the weights are all equal. Because $g_{1j} = 0$ when $P_{1j} > P_{2j}$, contribution to the variance comes only from those proportions for species 1 that are less than or equal to those for species 2. Hence, these proportions can be combined. If $P_{1j} \neq P_{2j}$ then:

$$V(PS) = V \sum_{j} {}_{j} + V \sum_{K} {}_{j} \quad P_{2k} \tag{14}$$

Here J refers to indices when $p_{1j} < p_{2j}$ and K to those indices where $p_{2k} < p_{1k}$. Hence, to compute the variance simply combine the data into two groups.

The SIMCONBA.BAS Program Documentation—

The SIMCONBA.BAS program was developed to apply both jackknife and bootstrap techniques to generate critical comparisons of similarity indices. The program is based on the work of Smith et al. (1986). It was written in Microsoft's QuickBasic version 4.0. It is designed to calculate similarity values for any two communities. The SIMI similarity coefficient is used at present, but any similarity coefficient could be implemented.

$$\text{SIMI} = \sum_{k=1}^{T} p_{1k} \cdot p_{2k} \Big/ \sqrt{\left(\sum_{k=1}^{T} p_{1k}^2 \cdot \sum_{k=1}^{T} p_{2k}^2\right)} \qquad (15)$$

where T is the total number of species included, p_{1k} and p_{2k} are the proportions of species k (based on the sum of that species abundances over all replicates) in communities 1 and 2, respectively. This SIMI value is symbolized by G in most of the following discussion.

The SIMI values and their variances are determined by jackknifing and bootstrapping. Ninety-five percent confidence intervals are also generated by the jackknife and bootstrap methods. For comparison, the simple SIMI values, comparing each possible combination of replicates and the mean of those values, are also generated.

Program Operation—

The proportional representation of each species from each community is calculated by summing the abundance of each species over all replicates within a community and dividing by the total number of individuals observed in that community. These proportions are estimates of the true proportions of each species in each community and are used in Equation (11) to calculate the SIMI value.

The jackknife estimate proceeds by removing one replicate from community 1, while leaving community 2 intact, and recalculating the SIMI value (G'_{-k}). It then replaces that replicate in community 1, removes a replicate from community 2, and calculates a

second SIMI value (G''_{-k}). Each of these values is then used to calculate a 'pseudo' value according to the following:

$$G'_k = (m_1 - 1/2)G - (m_1 - 1)G'_{-k} \tag{16}$$

$$G''_k = (m_2 - 1/2)G - (m_2 - 1)G''_{-k}' \tag{17}$$

where m_1 is the number of replicates from community 1, m_2 is the number of replicates from community 2, and G is the original SIMI value calculated by Equation (11).

This process is repeated, each time replacing the previous and removing the next consecutive replicate, until all replicates have been removed once. A pair of 'pseudo' values is generated at each interaction. These are then used to calculate the final jackknife estimate of SIMI:

$$G_J = 1/m_1 \sum_{k=1}^{m_1} G'_k + 1/m_2 \sum_{k=1}^{m_2} G''_k \tag{18}$$

They are also used in the determination of the variance of this estimate:

$$\sigma^2_1 = 1/(m_1-1) . \sum_{k=1}^{m_1} (G'_k - 1/m_1 . \sum_{k=1}^{m_1} G'_k)^2 \tag{19}$$

$$\sigma^2_2 = 1/(m_2-1). \sum_{k=1}^{m_2} (G''_k - 1/m_2 . \sum_{k=1}^{m_2} G''_k)^2 \tag{20}$$

estimating the variance by:

$$s_2 (G_J) = 1/m_1 . \sigma^2_1 + 1/m_2 . \sigma^2_2 \tag{21}$$

The 95 percent confidence interval is symmetric and is determined by the following:

$$G_J \pm t_c . s(G_J), \tag{22}$$

where t_c is the critical value from the t-table with $m_1 + m_2$-2 degrees of freedom and a confidence level of 1-α (i.e., $\alpha = 0.95$). Smith et al. (1986) describe a transformation using the inverse cosine which can be applied when the SIMI estimate is near zero or one, to account for the skewed distribution.

The bootstrap estimate proceeds by generating two sets of random numbers. These random numbers are uniformly distributed between 1 and the number of replicates for each

community. Each number represents the number of a replicate (from the original data) that will be used as a given replicate in the bootstrap sample. Each bootstrap sample contains the same number of replicates as the original data set. However, some of those original replicates may be represented more than once in each bootstrap sample.

After these randomly chosen representatives of each community are generated, they are used to calculate a new SIMI value ($G_{(i)}$). This process is repeated many times (1,000 or more is best). The bootstrap estimate of the true SIMI value (G_B) is then calculated as the mean of these values. The variance of this estimate is determined by the following:

$$s^2_B = (b - 1)^{-1} \cdot \sum_{i=1}^{b} (G_{(i)} - G_B)^2 \qquad (23)$$

where b is the number of bootstrap samples used to calculate G_B.

The 95 percent confidence interval is asymmetric and is bounded by the 2.5th and 97.5th percentiles of the sorted values. This program uses a bubblesort to order the bootstrapped values of SIMI ($G_{(i)}$). It is a dependable algorithm, but is probably the slowest step in the program. A shellsort or quicksort routine may prove to be faster.

A symmetric confidence interval can also be generated based on the t-statistic. However, that is not included in this program at present.

Smith et al. (1986) also discuss bias encounted in the bootstrap estimate. The program provides adjusted and biased values for both the bootstrap and jackknife estimates.

Data—

As input, the program uses estimates of the abundance of each species in each replicate from each of the two communities in question. These data may be entered interactively at run time with the keyboard or via a previously constructed file. If a file is used, it must conform to the following format:

LINE 1: TITLE (enclosed in double quotation marks ["])

LINE 2: The number of species included in the data set (= number of columns of data)

LINE 3: The number of replicates from community 1 and community 2 (in that order and separated by a comma or a blank).

LINE 4: The abundance of each species observed in the FIRST replicate from the FIRST community (each separated by a comma or a blank).

LINE 5: The abundance of each species observed in the SECOND replicate from the FIRST community (each separated by a comma or a blank).

.
.
.

LINE X: The abundance of each species observed in the FIRST replicate from the SECOND community (each separated by a comma or a blank).

.
.
.

A file of this form may be constructed by the program after keyboard entry of the data.

The data may be in the form of either absolute abundances or proportions. The program makes frequent reference to the original raw abundance data throughout its operation. Due to this, abundance data is generated from any data entered as proportions. The assumption is made that exactly 1,000 individuals were sampled by each replicate. The abundance data generated in this way can also be saved to a file if the original proportion data were entered through the keyboard.

The number of replicates taken is not usually the same from different communities. Therefore, this program allows the use of data sets which represent communities containing different numbers of replicates.

Process—

Many of the operations are interactive and can consume a considerable amount of time—even on a fast microcomputer. For this reason, there are many symbols displayed on the screen to reassure the user that the program is operating and not caught in an infinite loop.

Output—

Output may be sent to the screen, the printer, or a file. Three sets of output are generated by this program (Table 5). The first is the simple SIMI values without variance measurements. A table filled with a SIMI value for each possible combination (pairwise) of a replicate in community 1 with a replicate in community 2 is printed. The mean for all the values in this table is printed. Next, the SIMI value, based on the sum of species abundances over all replicates (G), is printed. This is followed by the output of the jackknife and bootstrap procedures. Each contains its SIMI estimate, that estimate adjusted for bias, the associated variance, and the 95 percent confidence interval. For the jackknife estimate, this interval is around the adjusted value. The jackknife estimate also includes the t-value used and the degrees of freedom associated with it. The bootstrap output displays the number of bootstrap samples used to provide the estimate.

Comparison of Program Performance with Test Data—

TEST.DTA—Bootstrap estimates used 1,000 bootstrap samples. Table 2 of Smith et al. (1986)(which is shown as our Table 2) offers some hypothetical data (TEST.DTA) and their results. These results include the simple SIMI values from all possible pairs of replicates from each community—as well as the mean of these values and the SIMI based on the abundances summed over replicates (G).

The SIMCONBA program exactly reproduced those results (see Table 3). Smith et al. also give bias adjusted similarity measures for the jackknife and bootstrap estimates and

Table 2 Hypothetical test data from Table 2 of Smith et al. (1985) and comparative results with the program described herein

		Species Type					Total
Community	Replicate	A	B	C	D	E	
1	1	75	75	250	100	0	500
	2	175	75	175	75	0	500
	3	100	100	150	125	25	500
	Totals	350	250	575	300	25	1500
2	1	125	100	125	150	0	500
	2	125	75	75	75	150	300
	3	100	100	75	75	150	500
	Totals	350	275	275	300	300	1500

SIMI Measures of above Community 1

	Replicate	1	2	3
	1	0.863	0.918	0.980
Community 2	2	0.008	0.732	0.780
	3	0.615	0.700	0.790

Table 3. Test data from Smith et al. 1986 and results compared

SIMI values for each combination of replicates

	Community 1		
	1	2	3
1	0.863911	0.919329	0.980409
2	0.607785	0.732422	0.780644
3	0.618412	0.700841	0.789669

Mean of these similarity coefficients: .776647

Similarity coefficient based on the averages of all replicates: .852024

Jackknife estimate of similarity with 95 percent confidence interval:

t-value = 2.132, with 4 df.

similarity coefficient:	0.814571
similarity coefficient adjusted for bias:	0.889477
variance:	0.015759

0.6218373 < JACKSIMI < 1.157116

Bootstrap estimate of similarity with 95 percent confidence interval:

1000 bootstrap samples used

Similarity coefficient:	0.826716
Similarity coefficient adjusted for bias:	0.877332
Variance:	0.007648

0.662016 < BOOTSIMI < 0.96700

Comparative Results

Data Set	Analysis	Result	SIMCONBA	Smith et al.
TEST.DTA	Simple	MEANSIMI	0.78	0.78
	SSIMI	XSIMI	0.85	0.85
	jackknife	similarity	0.814	*0.811
		adj.-sim.	0.889	0.889
		variance	0.0157	N/A
	t = 2.132	lower bound	0.622	0.54
	(4, 0.05)	upper bound	1.157	1.23
	bootstrap	similarity	0.827	*0.82
		adj.-sim.	0.877	0.876
		variance	0.0157	N/A
	t = 2.132	lower bound	0.622	0.73
	(4, 0.05)	upper bound	0.967	1.04

*estimated from bias adjusted value

values for the associated 95 percent confidence intervals. In Table 2 Smith et al.'s values of the unadjusted estimates were backcalculated from the bias adjusted values that they reported.

The program's jackknife and bootstrap estimates of SIMI closely matched those of Smith et al. However, the confidence intervals were narrower than those reported. In the case of the bootstrap, the values were close to those reported, and the random nature of the process or differences in machines used may explain most of this difference. Machine differences might also contribute to the deviations of the jackknife estimate, but it was the t-value that seemed to be most responsible for these.

Samar Sea Data—The Samar Sea data set (SAMAR.DTA, Table 4) consisted of three replicates (cruises) of demersal fish species abundances from one station (#22) at two points in time. The estimates of similarity between these two communities displayed a wide range, from 0.282 (smallest simple SIMI value) to 0.873 (largest simple SIMI value)(Table 5). The jackknife (unbiased 0.545, biased 0.701) and bootstrap (0.822) estimates indicated that the true value was probably in the upper end of this range, but their confidence intervals (0.627, 1.06 [jackknife]; 0.466, 0.846 [bootstrap] were wide. Smith et al. (1986) state that in situations where the variability is large and the sample size is small, the jackknife performs better than the bootstrap. This situation clearly exists for the Samar Sea data set used here. However, the jackknife did not perform better than the bootstrap. Both clearly showed the high variability in the system.

These results suggest that the community at station 22 did not change much over the time period between surveys. However, the variability was high and the true similarity may be lower than is indicated from this analysis. A larger data set with more replicates would help clarify the situation.

Table 4. Data from six cruises at two different sites in the Samar Sea, Philippines (five major species)

Assemblage	Cruise	Spp67	Spp77	Spp106	Spp117	Spp179	Totals
1	97	375	920	453	277	89	2114
	98	718	110	372	8	99	1307
	99	1736	203	448	0	54	2441
	Totals	2829	1233	1273	285	242	5862
2	107	36	983	999	30	84	2132
	108	2514	337	3878	355	369	7453
	109	796	458	2196	453	498	13148

Species codes:

67 — *Lelognathus bindus*

77 — *Loligo* spp.

106 — *Pentaprion longimanus*

117 — *Pria canthus tagenus*

177 — *Upeneus moluceensis*

Table 5. Data from six cruises at two different times at station 22 in the Samar Sea, Philippines

SIMI values for each combination of replicates

	Community 1		
	1	2	3
1	0.873	0.445	0.282
2	0.596	0.867	0.736
3	0.646	0.737	0.572

Mean of these similarity coefficients: 0.639382

Similarity coefficient based on the averages of all replicates: 0.773120

Jackknife estimate of similarity with 95 percent confidence interval:

t-value = 2.132, with 4 df.

similarity coefficient:	0.701589
similarity coefficient adjusted for bias:	0.844650
variance:	0.010389

0.627338 < JACKSIMI < 1.061961

Bootstrap estimate of similarity with 95 percent confidence interval:

1000 bootstrap samples used

Similarity coefficient:	0.728795
Similarity coefficient adjusted for bias:	0.817444
Variance:	0.007888

0.471423 < BOOTSIMI < 0.857882

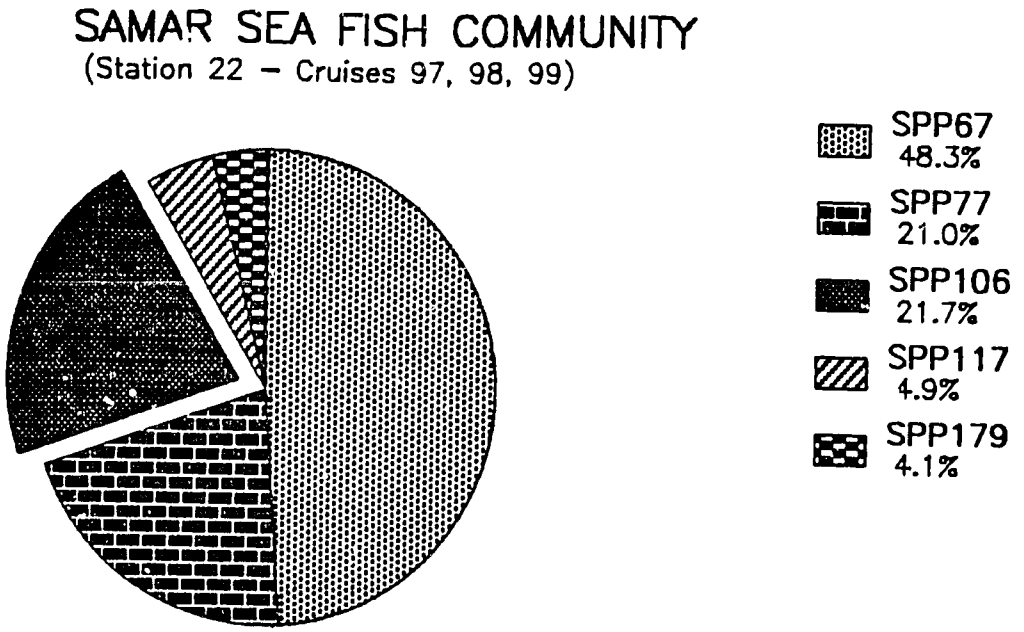
If one looks at the species abundances graphically (see Figure 1), it is clear that the dominance of the community shifted from species 67 to species 106, while the proportions represented by the other species remained relatively constant.

Conclusions—

Similarity values are important tools for gauging change in complex natural (as well as artificial) communities. When replicate samples from the communities being compared are available, we may take advantage of some powerful, but computationally-intensive methods, to more accurately measure their similarity and the variability around that measure.

This variability can be used to judge the reliability of the similarity measure and to test for significant differences between the similarity of other community pairs.

As in all research, the extent to which the available data accurately describe the system of interest must be considered. Despite the powerful nature of these techniques to get the most out of the data, the results will only be as good as those data. If the data set is insufficient, then the variability will probably be too large to draw any definitive conclusions about the communities being investigated.



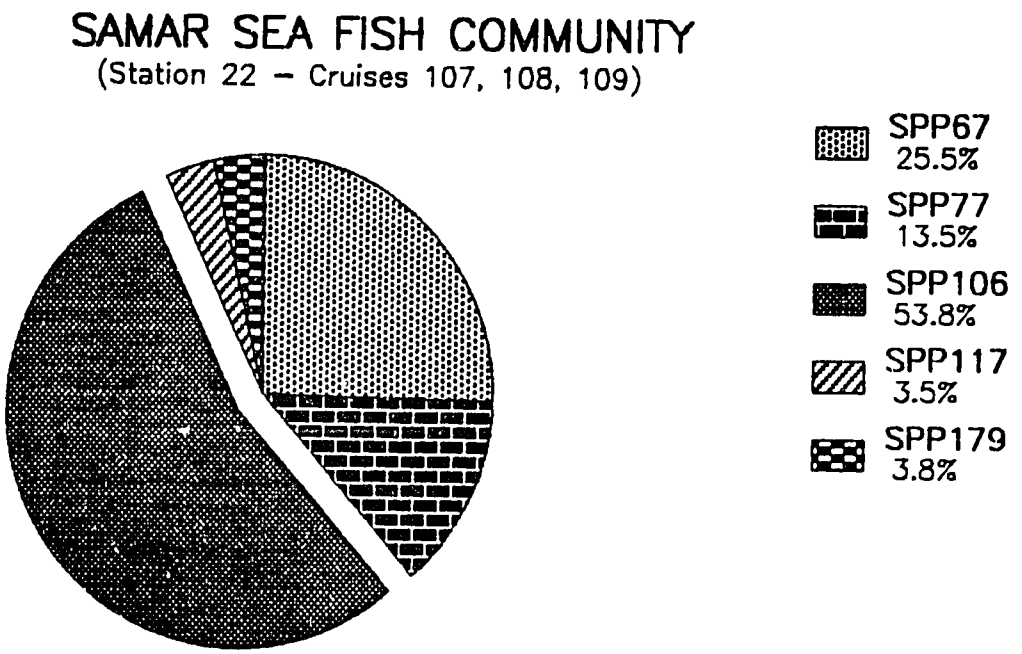


Figure 1. Graphical portrayal of species abundances from a portion of the Samar Sea fish community taken by trawl sampling.

References

Brock, A.R. 1977. Comparison of community similarity indexes. Journ. W.P.C.F. 48:2488-2493.

Czarnecki, D.B. 1979. Epipetic and epilithic diatom assemblages in Montezuma Wall National Monument. Arizona. J. Phycol. 15:246-352.

Grassle, J.F., and W. Smith. 1976. A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. Oecologia 25:13-22.

Heltshe, J.F. 1988. Jackknife estimate of the matching coefficient of similarity. Biometrics 44:447-460.

Johnston, J.W. 1976. Similarity indices I: What do they measure. Battelle Northwest Labs. BNWL-LIS2 NRC-1. 100 pp + appendix.

Kobayashi, S. 1987. Heterogeneity ratio: A measure of Beta-diversity and its use in community classification. Ecol. Res. 2, 101-111.

Kohn, A.J. and A.C. Riggs. 1982. Sample size dependence in increases of proportional similarity. Mar. Ecol. Progress Ser. 9:149-151.

Morisita, M. 1971. Composition of the I_s-index. Res. Popul. Ecol. 13:1-27.

_____. 1959. Measuring of interspecific association and similarity between communities. Man. Fac. Sci. Kyushu Univ. Ser. E (Biol) 3:65-80.

Mountford, M.D. 1962. An index of similarity and its application to classification problems. In: (P.W. Murphy, ed). Progress in Soil. Zoology. pp. 43-50. Butterworths, London.

Nemec, A.F. L., and R.O. Brinkhurst. 1988. Using the bootstrap to assess statistical significance in the cluster analysis of species abundance data. Can. J. Fish. Aquatic Sci. 45:965-970.

Pielou, E.C. 1979. Interpretation of paleoecological similarity indices. Paleobiology 5:435-443.

Smith, E.P. 1985. Statistical comparison of weighted overlap measures. Trans. Amer. Fish. Soc. 114:250-257.

Smith, E.P., R.B. Genter, and J. Cairns, Jr. 1986. Confidence intervals for the similarity between algal communities. Hydrobiologia 139:237-245.

Washington, H.G 1984. Diversity, biotic and similarity indices. Water Res. 18(6): 651-694.

Appendix

```
10 PRINT "                    ***** SIMCONBA.BAS *****"
20 PRINT "                    ( BASICA version)"
30 PRINT
40 PRINT "                            by Jim McKenna"
50 PRINT "                            25 March 1989"
60 PRINT " This program will accept data on the abundance of a set of species"
70 PRINT " at two sites (communities), with at least two replicates from each"
80 PRINT " site.  These data will then be analyzed for similarity using the "
90 PRINT " SIMI coefficient of Stander (1970):"
100 PRINT "                 SIMI = ΣP1k P2k/√(ΣP1k² ΣP2k²)"
110 PRINT
120 PRINT " The variance and 95% confidence intervals are determined using"
130 PRINT " the bootstrap and jackknife methods."
140 PRINT
150 PRINT " This program is based on the work of Smith, Genter, and Cairns (1986)"
160 PRINT " It was written in Microsoft's QuickBasic, version 4.0"
170 PRINT
180 PRINT " Smith, E.P.; Genter, R.B.; Cairns, J.Jr. Confidence intervals "
190 PRINT " for the similarity between algal communities. Hydrobiologia "
200 PRINT " 139:237-245;1986."
210 PRINT
220 PRINT "ABSTRACT:"
230 PRINT "  While researchers commonly use similarity measures to compare "
240 PRINT " algal communities, very few researchers have considered the "
250 PRINT " variability of these estimated measures.  This paper discusses"
260 PRINT " a recent method for estimating the variance of and confidence"
270 PRINT " intervals for similarity measures proposed by Johnson & Millie "
280 PRINT " (1982, Hydrobiologia 89:3-8). Applications of this method to data "
290 PRINT " have produced confidence intervals that are too narrow.  Two "
300 PRINT " alternative methods, the jackknife method and the bootstrap, are"
310 PRINT " shown to provide superior estimates of the variability."
320 PRINT
330 PRINT
340 PRINT "Strike any key to continue ...": WHILE INKEY$ = "": WEND
350 PRINT " The program compares only two communities at a time"
360 PRINT " The data are accepted only as keyboard input at the present"
370 PRINT " and are expected to be the absolute abundances for each species."
380 PRINT
390 PRINT
400  '
410 INPUT "Are your data absolute abundances (A) or proportions (P) "; ANS1$
420 INPUT "Are your data in a file "; ANS2$
430  ' ...
440  '
450 INPUT " How many species are included in the data "; SPP%
460 INPUT " How many replicates were taken from community 1 "; M1%
470 INPUT " How many replicates were taken from community 2 "; M2%
480  '
490 IF M1% > M2% THEN MX% = M1% ELSE MX% = M2%
540  '
550 DIM SIMI(M1%, M2%):  'ARRAY to hold the values when each rep. used separately
560 DIM COMM1(MX%, SPP%), COMM2(MX%, SPP%):  ' ARRAYS holding raw data for each community
570 DIM PROP1(MX%, SPP%), PROP2(MX%, SPP%):  ' ARRAYS holding proportions for each community
580 DIM NN1J(MX%), NN2J(MX%):  ' ARRAYS holding the total # ind. in each replicate
```

```
590 DIM NN1K(SPP%), NN2K(SPP%):  ' ARRAYS holding the total for each spp. in each comm.
600 DIM XPROP1(SPP%), XPROP2(SPP%):  ' ARRAYS holding the proportions by totals of each spp.
610 DIM RD1(M1%), RD2(M2%):  ' ARRAYS to hold random numbers for bootstrap est.
620 DIM BOOT1(M1%, SPP%), BOOT2(M2%, SPP%):  ' ARRAYS to hold bootstrap samples
630 DIM JN1K(SPP%), JN2K(SPP%):  ' ARRAYS to hold total for each in each jackknife comm.
640 DIM JPROP1(SPP%), JPROP2(SPP%):  ' ARRAYS to hold proportions from jackknifed comm.
650 DIM PSEUDO1(MX%), PSEUDO2(MX%):  'ARRAYS to hold pseudo values from jackknife est.
660  ' N1, N2 are the total # of ind. of all spp. from community 1 & 2, respectively
670  '
680  'Initialize the raw data arrays
690 FOR I% = 1 TO MX%
700         PRINT ".";
710         FOR J% = 1 TO SPP%
720                 COMM1(I%, J%) = 0
730                 COMM2(I%, J%) = 0
740         NEXT J%
750 NEXT I%
760 PRINT
770  ' Initialize other arrays
780 GOSUB 2810:   'Sub. to initialize proportion arrays etc.
790  ' Initialize overall proportions
800 FOR I% = 1 TO SPP%
810         XPROP1(I%) = 0
820         XPROP2(I%) = 0
830 NEXT I%
840  '
850  ' Data entry
860  ' ANS1$ = "A": ANS2$ = "N": ' Temporary - until input routine is improved
870 IF (ANS2$ = "Y" OR ANS2$ = "y") AND (ANS1$ = "A" OR ANS1$ = "a") THEN GOSUB 1990
880 IF (ANS2$ = "Y" OR ANS2$ = "y") AND (ANS1$ <> "A" AND ANS1$ <> "a") THEN GOSUB 2200:   'SUB. TO READ
PROPORTIONS
940 IF (ANS2$ <> "Y" AND ANS2$ <> "y") AND (ANS1$ = "A" OR ANS1$ = "a") THEN GOSUB 2450:   'SUB. TO INPUT #'s
FROM KEYBOARD
950 IF (ANS2$ <> "Y" AND ANS2$ <> "y") AND (ANS1$ <> "A" AND ANS1$ <> "a") THEN GOSUB 2610:  'SUB. TO INPUT
PROPORTIONS FROM KEYBOARD
1000  '
1010  ' Compute totals for each replicate (total # ind. for all spp. - each rep.)
1020 FOR I% = 1 TO SPP%
1030         FOR J% = 1 TO M1%
1040                 NN1J(J%) = NN1J(J%) + COMM1(J%, I%)
1050         NEXT J%
1060         FOR J% = 1 TO M2%
1070                 NN2J(J%) = NN2J(J%) + COMM2(J%, I%)
1080         NEXT J%
1090 NEXT I%
1100  ' Compute the proportion for each species w/in each rep. from each comm.
1110 FOR I% = 1 TO SPP%
1120         FOR J% = 1 TO M1%
1130                 PROP1(J%, I%) = COMM1(J%, I%) / NN1J(J%)
1140         NEXT J%
1150         FOR J% = 1 TO M2%
1160                 PROP2(J%, I%) = COMM2(J%, I%) / NN2J(J%)
1170         NEXT J%
1180 NEXT I%
```

```
1190  '
1200  ' Calculate SIMI values for each combination of replicates
1210 FOR I% = 1 TO M1%
1220        FOR J% = 1 TO M2%
1230                PART1 = 0: PART2A = 0: PART2B = 0:  ' REINITIALIZE
1240                FOR K% = 1 TO SPP%
1250                        PART1 = PART1 + (PROP1(I%, K%) * PROP2(J%, K%))
1260                        PART2A = PART2A + PROP1(I%, K%) ^ 2
1270                        PART2B = PART2B + PROP2(J%, K%) ^ 2
1280                NEXT K%
1290                SIMI(I%, J%) = PART1 / SQR(PART2A * PART2B)
1300                TOTSIMI = TOTSIMI + SIMI(I%, J%)
1310                PRINT USING "#.### "; SIMI(I%, J%),
1320        NEXT J%
1330 NEXT I%
1340 PRINT
1350 MEANSIMI = TOTSIMI / (M1% * M2%):  ' MEAN of SIMI values
1360  '
1370  'The FUNCTION SIMILRTY is the procedure to compute SIMI from averages of reps.
1380 GOSUB 2810:  ' INITIALIZE arrays and variables
1390  ' Re-initialize overall species proportions
1400 FOR I% = 1 TO SPP%
1410        XPROP1(I%) = 0
1420        XPROP2(I%) = 0
1430 NEXT I%
1440  ' The SIMI value when all replicates are averaged
1450  ' QB4: XSIMI = SIMILRTY(COMM1(), COMM2(), SPP%, M1%, M2%, NN1K(), NN2K(), N1, N2, XPROP1(), XPROP2(),
PART1, PART2A, PART2B)
1460  '
1470  'FUNCTION SIMILRTY (COMM1(), COMM2(), SPP%, M1%, M2%, NN1K(), NN2K(), N1, N2, XPROP1(), XPROP2(), PART1,
PART2A, PART2B)
1480  'Compute totals for each species for each community
1490 FOR II% = 1 TO SPP%
1500        FOR JJ% = 1 TO M1%
1510                NN1K(II%) = NN1K(II%) + COMM1(JJ%, II%)
1520        NEXT JJ%
1530        FOR JJ% = 1 TO M2%
1540                NN2K(II%) = NN2K(II%) + COMM2(JJ%, II%)
1550        NEXT JJ%
1560 NEXT II%
1570  ' Calc. grand total for each community
1580 FOR II% = 1 TO SPP%
1590        N1 = N1 + NN1K(II%)
1600        N2 = N2 + NN2K(II%)
1610 NEXT II%
1620  'Calc. proportions for each species w/in each community
1630 FOR II% = 1 TO SPP%
1640        XPROP1(II%) = NN1K(II%) / N1
1650        XPROP2(II%) = NN2K(II%) / N2
1660 NEXT II%
```

```
1670  ' Calc. SIMI value for this pair of communities
1680 FOR KK% = 1 TO SPP%
1690         PART1 = PART1 + (XPROP1(KK%) * XPROP2(KK%))
1700         PART2A = PART2A + XPROP1(KK%) ^ 2
1710         PART2B = PART2B + XPROP2(KK%) ^ 2
1720 NEXT KK%
1730 XSIMI = PART1 / SQR(PART2A * PART2B)
1740  '
1750  ' QB4: END FUNCTION
1760 INPUT "Do you want the results printed to the printer (P), a file (F), or the screen (S)"; ANS4$
1770 IF ANS4$ = "P" OR ANS4$ = "p" THEN UNIT$ = "LPT1:" ELSE IF ANS4$ = "F" OR ANS4$ = "f" THEN INPUT "Enter a
name for the output file"; UNIT$ ELSE UNIT$ = "SCRN:"
1840 OPEN "O", 2, UNIT$
1850 IF TITL$ = "" THEN INPUT "ENTER a title for this run"; TITL$
1860 PRINT #2, TITL$
1870  '
1880  'Print out results thus far:
1890 GOSUB 3030:   '1st print out routine
1900  '
1910  ' Calculate jackknife est. and variance
1920 GOSUB 3290:  ' jackknife procedure
1930  '
1940  ' Calculate bootstrap est. and variance
1950 GOSUB 4530:  ' bootstrap procedure
1960  '
1970 END
1980  '
1990   ' SUB. to read data (numbers) from a file
2000 INPUT "Enter the name of the data file"; NAM1$
2010 OPEN "I", 1, NAM1$
2020 INPUT #1, TITL$: PRINT TITL$
2030 INPUT #1, SPP%: PRINT SPP%
2040 INPUT #1, M1%, M2%: PRINT M1%, M2%
2050 FOR I% = 1 TO M1%
2060         FOR J% = 1 TO SPP%
2070                INPUT #1, COMM1(I%, J%): PRINT COMM1(I%, J%);
2080         NEXT J%
2090         PRINT
2100 NEXT I%
2110 FOR I% = 1 TO M2%
2120         FOR J% = 1 TO SPP%
2130                INPUT #1, COMM2(I%, J%): PRINT COMM2(I%, J%);
2140         NEXT J%
2150         PRINT
2160 NEXT I%
2170 CLOSE 1
2180 RETURN
2190  '
2200   ' SUB. to read data (PROPORTIONS) from a file
2210 INPUT "Enter the name of the data file"; NAM1$
2220 OPEN "I", 1, NAM1$
2230 INPUT #1, TITL$: PRINT TITL$
2240 INPUT #1, SPP%: PRINT SPP%
2250 INPUT #1, M1%, M2%: PRINT M1%, M2%
```

```
2260 PRINT "To generate abundances from these proportions it is assumed"
2270 PRINT " that exactly 1000 individuals were sampled by each replicate."
2280 FOR I% = 1 TO M1%
2290        FOR J% = 1 TO SPP%
2300                INPUT #1, PROP1(I%, J%): PRINT PROP1(I%, J%);
2310                COMM1(I%, J%) = PROP1(I%, J%) * 1000:  ' CONVERT to abundance.
2320        NEXT J%
2330        PRINT
2340 NEXT I%
2350 FOR I% = 1 TO M2%
2360        FOR J% = 1 TO SPP%
2370                INPUT #1, PROP2(I%, J%): PRINT PROP2(I%, J%);
2380                COMM2(I%, J%) = PROP2(I%, J%) * 1000:  ' CONVERT to abundances
2390        NEXT J%
2400        PRINT
2410 NEXT I%
2420 CLOSE 1
2430 RETURN
2440  '
2450   ' SUB. TO READ ABSOLUTE ABUNDANCE DATA ENTRY FROM KEYBOARD
2460 FOR I% = 1 TO SPP%
2470        FOR J% = 1 TO M1%
2480                PRINT "Enter the abundance of species "; I%; "in replicate"; J%; "from community 1 :";
2490                INPUT COMM1(J%, I%)
2500        NEXT J%
2510        FOR J% = 1 TO M2%
2520                PRINT "Enter the abundance of species "; I%; "in replicate"; J%; "from community 2 :";
2530                INPUT COMM2(J%, I%)
2540        NEXT J%
2550 NEXT I%
2560 PRINT
2570 INPUT "Do you want to save these data in a file"; ANS3$
2580 IF ANS3$ = "Y" OR ANS3$ = "N" THEN GOSUB 5800
2590 RETURN
2600  '
2610   ' SUB. TO READ ABSOLUTE ABUNDANCE DATA ENTRY FROM KEYBOARD
2620 PRINT "To generate abundances from these proportions it is assumed"
2630 PRINT " that exactly 1000 individuals were sampled by each replicate."
2640 FOR I% = 1 TO SPP%
2650        FOR J% = 1 TO M1%
2660                PRINT "Enter the proportion of species "; I%; "in replicate"; J%; "from community 1 :";
2670                INPUT PROP1(J%, I%)
2680                COMM1(J%, I%) = PROP1(J%, I%) * 1000:  ' CONVERT to abundance
2690        NEXT J%
2700        FOR J% = 1 TO M2%
2710                PRINT "Enter the proportion of species "; I%; "in replicate"; J%; "from community 2 :";
2720                INPUT PROP2(J%, I%)
2730                COMM2(J%, I%) = PROP2(J%, I%) * 1000:  ' CONVERT to abundance
2740        NEXT J%
2750 NEXT I%
2760 PRINT
2770 INPUT "Do you want to save these data in a file"; ANS3$
2780 IF ANS3$ = "Y" OR ANS3$ = "N" THEN GOSUB 5800
2790 RETURN
```

```
2800  '
2810   ' SUB. to initialize the proportion arrays and some other variables
2820 FOR I% = 1 TO SPP%
2830        FOR J% = 1 TO MX%
2840                PROP1(J%, I%) = 0
2850                PROP2(J%, I%) = 0
2860        NEXT J%
2870        NN1K(I%) = 0
2880        NN2K(I%) = 0
2890        JN1K(I%) = 0
2900        JN2K(I%) = 0
2910 NEXT I%
2920  '
2930 PART1 = 0
2940 PART2A = 0
2950 PART2B = 0
2960 TOTSIMI = 0
2970 N1 = 0
2980 N2 = 0
2990 JN1 = 0
3000 JN2 = 0
3010  '
3020 RETURN
3030   ' SUB. to print out the results of the first part
3040 PRINT #2, : PRINT
3050 PRINT #2, "   ##### SIMI values for each combination of replicates #####   "
3060 PRINT #2,
3070 PRINT #2, TAB(25); " COMMUNITY 1"
3080 FOR I% = 1 TO M1%
3090        PRINT #2, TAB(I% * 10); I%;
3100 NEXT I%
3110 PRINT #2,
3120 PRINT #2, TAB(10); STRING$(60, "_")
3130 FOR J% = 1 TO M2%
3140        PRINT #2, TAB(2); J%; TAB(6); "|";
3150        FOR I% = 1 TO M1%
3160                PRINT #2, TAB(I% * 10 - 2);
3170                PRINT #2, USING "#.### "; SIMI(I%, J%);
3180        NEXT I%
3190        PRINT #2,
3200 NEXT J%
3210 PRINT #2,
3220 PRINT #2, "MEAN OF THESE SIMILARITY COEFFICIENTS:"; MEANSIMI
3230 PRINT #2,
3240 PRINT #2, "SIMILARITY COEFFICIENT BASED ON THE AVERAGES OF ALL REPLICATES:"; XSIMI
3250 PRINT #2,
3260 PRINT "Strike any key to contirue ...": WHILE INKEY$ = "": WEND
3270 RETURN
3280  '
```

```
3290  'sub. to calc. jackknife estimates
3300 PRINT "Jackknifing ..."
3310  ' One replicate from one community is removed, a SIMI value is calclulated,
3320  ' then a replicate from the other community is removed and another SIMI value
3330  ' is calculated. These are combined to give one of the pseudo-SIMI values.
3340  ' This is repeated until all replicates have been removed once.
3350  '
3360 FOR K% = 1 TO MX%:  ' Keep track of the replicate that is being removed.
3370        GOSUB 2810:   ' RE-initialize summing arrays
3380        FOR G% = 1 TO 2:  'Keep track of which community the rep. is removed.
3390                FOR I% = 1 TO SPP%
3400                        PRINT "J";
3410  ' Calculate the total # ind. for each spp. in all reps. being used.
3420                        FOR J% = 1 TO MX%
3430                                IF J% = K% GOTO 3490:  ' SKIP the removed rep.
3440                                IF G% = 1 THEN JN1K(I%) = JN1K(I%) + COMM1(J%, I%) ELSE JN2K(I%) =
JN2K(I%) + COMM2(J%, I%)
3490                        NEXT J%
3500                NEXT I%
3510                PRINT
3520  ' Calc. total # of ind./community (JN1;JN2)
3530                FOR I% = 1 TO SPP%
3540                        PRINT "T";
3550                        IF G% = 1 THEN JN1 = JN1 + JN1K(I%) ELSE JN2 = JN2 + JN2K(I%)
3600                NEXT I%
3610                PRINT
3620  ' Calc. proportions for each species
3630                FOR I% = 1 TO SPP%
3640                        PRINT "%";
3650                        IF G% = 1 THEN JPROP1(I%) = JN1K(I%) / JN1 ELSE JPROP2(I%) = JN2K(I%) / JN2
3700                NEXT I%
3710                PRINT
3720  ' Calculate G -k
3730                PART1 = 0: PART2A = 0: PART2B = 0
3740                FOR I% = 1 TO SPP%
3750                        PRINT "G";
3760                        IF G% = 2 THEN GOTO 3810
3770                                PART1 = PART1 + (JPROP1(I%) * XPROP2(I%))
3780                                PART2A = PART2A + JPROP1(I%) ^ 2
3790                                PART2B = PART2B + XPROP2(I%) ^ 2:  'original prop.
3800                                GOTO 3850
3810                                PART1 = PART1 + (XPROP1(I%) * JPROP2(I%))
3820                                PART2A = PART2A + XPROP1(I%) ^ 2:  'original prop.
3830                                PART2B = PART2B + JPROP2(I%) ^ 2
3850                NEXT I%
3860                XXX = PART1 / SQR(PART2A * PART2B)
3870                IF G% = 1 THEN G1 = XXX ELSE G2 = XXX
3920                PRINT
3930        NEXT G%
3940  ' Calc. pseudo-SIMI values
3950        PSEUDO1(K%) = (M1% - .5) * XSIMI - (M1% - 1) * G1
3960        PSEUDO2(K%) = (M2% - .5) * XSIMI - (M2% - 1) * G2
3970 NEXT K%
3980  '
```

```
3990  ' Sum the pseudo values
4000 TOTPSEU1 = 0
4010 TOTPSEU2 = 0
4020 FOR I% = 1 TO MX%
4030        PRINT "Σ";
4040        TOTPSEU1 = TOTPSEU1 + PSEUDO1(I%)
4050        TOTPSEU2 = TOTPSEU2 + PSEUDO2(I%)
4060 NEXT I%
4070 PRINT
4080  '
4090  'CALC. jackknife est. - automatically adjusts for bias
4100 JACK = (1 / M1%) * TOTPSEU1 + (1 / M2%) * TOTPSEU2
4110  '
4120  ' BIASED ESTIMATE
4130  ' (bias = JACK - XSIMI)
4140 JACKADJ = 2 * XSIMI - JACK
4150  '
4160  'Est. variance
4170  '  est. var. for pseudo1
4180 SIGA = 0
4190 FOR I% = 1 TO M1%
4200        SIGA = SIGA + (PSEUDO1(I%) - (1 / M1%) * TOTPSEU1) ^ 2
4210 NEXT I%
4220 SIG1 = (1 / (M1% - 1)) * SIGA:  ' VARIANCE FOR PSEUDO1
4230  '  est. var. for pseudo2
4240 SIGA = 0
4250 FOR I% = 1 TO M2%
4260        SIGA = SIGA + (PSEUDO2(I%) - (1 / M2%) * TOTPSEU2) ^ 2
4270 NEXT I%
4280 SIG2 = (1 / (M2% - 1)) * SIGA:  ' Variance for pseudo2
4290  '
4300  ' final variance estimate
4310 JACKVAR = (1 / M1%) * SIG1 + (1 / M2%) * SIG2
4320  '
4330  'Confidence interval and printout
4340 PRINT "Degrees of freedom = "; M1% + M2% - 2
4350 PRINT "Confidence level (1-α) = ";
4360 PRINT USING "#.###"; 1 - .95;
4370 PRINT " for a 95% interval"
4380 INPUT "Please enter the t value based on these parameters"; T
4390 PRINT #2, : PRINT #2,
4400 PRINT #2, "JACKKNIFE estimate of similarity w/ 95% confidence interval"
4410 PRINT #2,
4420 PRINT #2, TAB(10); "t-VALUE = "; T; ", WITH "; M1% + M2% - 2; " df."
4430 PRINT #2, TAB(10); "SIMILARITY COEFFICIENT :"; TAB(52); JACKADJ
4440 PRINT #2, TAB(10); "SIMILARITY COEFFICIENT adjusted for bias :"; JACK
4450 PRINT #2, TAB(10); "VARIANCE :"; TAB(52);
4460 PRINT #2, USING "#.######"; JACKVAR
4470 PRINT #2, TAB(15); (JACK - SQR(JACKVAR) * T); "< JACKSIMI <"; (JACK + SQR(JACKVAR) * T)
4480 PRINT #2,
4490 PRINT "Strike any key to continue ... ": WHILE INKEY$ = "": WEND
4500  '
4510 RETURN
4520  '
```

```
4530  'Sub. to calculate the bootstrap est. of similarity and variance
4540 INPUT "How many bootstrap samples do you want to generate "; B%
4550 DIM GB(B%)
4560  '
4570 FOR K% = 1 TO B%:  '  ****** START of bootstrap sample loop *******
4580        PRINT "B";
4590  ' generate random #'s (uniform between 1 and max.) for the reps. to be used
4600        RANDOMIZE TIMER:  ' seed random # generator
4610        FOR G% = 1 TO 2:  ' Keep track of the community being sampled
4620               IF G% = 1 THEN MX% = M1% ELSE MX% = M2%
4670               FOR I% = 1 TO MX%
4680                      IF G% = 1 THEN RD1(I%) = INT(RND * MX% + 1) ELSE RD2(I%) = INT(RND * MX% + 1)
4730               NEXT I%
4740  ' Generate bootstrap community
4750               FOR I% = 1 TO MX%
4760                      FOR J% = 1 TO SPP%
4770                              IF G% = 1 THEN BOOT1(I%, J%) = COMM1(RD1(I%), J%) ELSE BOOT2(I%, J%) =
COMM2(RD2(I%), J%)
4820                      NEXT J%
4830               NEXT I%
4840        NEXT G%
4850  ' GOTO sub. to recalculate the similarity as was done in the first place.
4860        GOSUB 2810:   'INITIALIZE arrays and variables
4870  '     Re-initialize overall species proportions
4880        FOR I% = 1 TO SPP%
4890               XPROP1(I%) = 0
4900               XPROP2(I%) = 0
4910        NEXT I%
4920  '
4930  ' QB4:  GB(K%) = SIMILRTY(BOOT1(), BOOT2(), SPP%, M1%, M2%, NN1K(), NN2K(), N1, N2, XPROP1(), XPROP2(),
PART1, PART2A, PART2B)
4940  '              FUNCTION SIMILRTY (BOOT1(), BOOT2(), SPP%, M1%, M2%, NN1K(), NN2K(), N1, N2, XPROP1(),
XPROP2(), PART1, PART2A, PART2B)
4950  '              Compute totals for each species for each community
4960               FOR II% = 1 TO SPP%
4970                      FOR JJ% = 1 TO M1%
4980                              NN1K(II%) = NN1K(II%) + BOOT1(JJ%, II%)
4990                      NEXT JJ%
5000                      FOR JJ% = 1 TO M2%
5010                              NN2K(II%) = NN2K(II%) + BOOT2(JJ%, II%)
5020                      NEXT JJ%
5030               NEXT II%
5040                 Calc. grand total for each community
5050               FOR II% = 1 TO SPP%
5060                      N1 = N1 + NN1K(II%)
5070                      N2 = N2 + NN2K(II%)
5080               NEXT II%
5090                Calc. proportions for each species w/in each community
5100               FOR II% = 1 TO SPP%
5110                      XPROP1(II%) = NN1K(II%) / N1
5120                      XPROP2(II%) = NN2K(II%) / N2
5130               NEXT II%
```

```
5140  '                Calc. SIMI value for this pair of communities
5150                FOR KK% = 1 TO SPP%
5160                        PART1 = PART1 + (XPROP1(KK%) * XPROP2(KK%))
5170                        PART2A = PART2A + XPROP1(KK%) ^ 2
5180                        PART2B = PART2B + XPROP2(KK%) ^ 2
5190                NEXT KK%
5200         GB(K%) = PART1 / SQR(PART2A * PART2B)
5210  '
5220  ' QB4: END FUNCTION
5230 NEXT K%
5240 PRINT
5250  '
5260  ' Calc. the mean of bootstrap similarity values
5270 TOTGB = 0
5280 FOR I% = 1 TO B%
5290         PRINT "Σ";
5300         TOTGB = TOTGB + GB(I%)
5310 NEXT I%
5320 PRINT
5330 BOOT = TOTGB / B%
5340  '
5350  'Calc. the variance
5360 DIFFGB = 0
5370 FOR I% = 1 TO B%
5380 PRINT "-";
5390         DIFFGB = DIFFGB + (GB(I%) - BOOT) ^ 2
5400 NEXT I%
5410 PRINT
5420 BOOTVAR = DIFFGB / (B% - 1)
5430  '
5440  ' BIAS AND ADJUSTMENT
5450  ' {bias = BOOT - XSIMI}
5460 GADJ = 2 * XSIMI - BOOT
5470  '
5480  'Calc. the confidence interval for bootstrap
5490  'sort the GB's - bubble sort (this is a slow method)
5500 FOR I% = 1 TO B% - 1
5510         PRINT "<>";
5520         FOR J% = I% + 1 TO B%
5530                IF GB(J%) > GB(I%) GOTO 5580
5540                        TEMP = GB(I%)
5550                        GB(I%) = GB(J%)
5560                        GB(J%) = TEMP
5580         NEXT J%
5590 NEXT I%
5600 PRINT
5610  ' PRECENTILES
5620 LOWER% = .025 * B%:  ' Depth of the 2.5 %ile of the list of bootstrap samples
5630 UPPER% = .975 * B%:  ' Depth of the 97.5 %ile of the list of bootstrap samples
5640 GB25TH = GB(LOWER%)
5650 GB975TH = GB(UPPER%)
5660  '
```

```
5670  'PRINTOUT
5680 PRINT #2, : PRINT #2,
5690 PRINT #2, "BOOTSTRAP estimate of similarity with 95% confidence interval"
5700 PRINT #2,
5710 PRINT #2, TAB(10); B%; " BOOTSTRAP SAMPLES USED"
5720 PRINT #2, TAB(10); "SIMILARITY COEFFICIENT :"; TAB(52); BOOT
5730 PRINT #2, TAB(10); "SIMILARITY COEFFICIENT adjusted for bias :"; GADJ
5740 PRINT #2, TAB(10); "VARIANCE :"; TAB(52);
5750 PRINT #2, USING "#.######"; BOOTVAR
5760 PRINT #2, TAB(15); GB(LOWER%); "< BOOTSIMI <"; GB(UPPER%)
5770 PRINT #2,
5780 RETURN
5790  '
5800   ' SUB. TO SAVE ENTERED DATA IN A FILE
5810 INPUT "ENTER the name for the output file"; NAM2$
5820 OPEN "O", 2, NAM2$
5830 INPUT "Enter a title for this data set"; TITL$
5840 PRINT #2, CHR$(34); TITL$; CHR$(34)
5850 PRINT #2, SPP%
5860 PRINT #2, M1%, M2%
5870 FOR I% = 1 TO M1%
5880        FOR J% = 1 TO SPP%
5890                PRINT #2, TAB(5 * J%); COMM1(I%, J%);
5900        NEXT J%
5910 NEXT I%
5920 FOR I% = 1 TO M2%
5930        FOR J% = 1 TO SPP%
5940                PRINT #2, TAB(5 * J%); COMM2(I%, J%);
5950        NEXT J%
5960 NEXT I%
5970 PRINT #2, ""
5980 PRINT #2, "The format of this data set is as follows:"
5990 PRINT #2, " TITLE"
6000 PRINT #2, " # OF SPECIES"
6010 PRINT #2, " # OF REPLICATES FROM COMMUNITY 1, # OF REPLICATES FROM COMMUNITY 2"
6020 PRINT #2, " THE ABUNDANCE OF EACH SPECIES FROM REPLICATE 1, COMMUNITY 1"
6030 PRINT #2, " THE ABUNDANCE OF EACH SPECIES FROM REPLICATE 2, COMMUNITY 1"
6040 PRINT #2, " THE ABUNDANCE OF EACH SPECIES FROM REPLICATE 3, COMMUNITY 1"
6050 PRINT #2, " ..."
6060 PRINT #2, " THE ABUNDANCE OF EACH SPECIES FROM REPLICATE 1, COMMUNITY 2"
6070 PRINT #2, " THE ABUNDANCE OF EACH SPECIES FROM REPLICATE 2, COMMUNITY 2"
6080 PRINT #2, " THE ABUNDANCE OF EACH SPECIES FROM REPLICATE 3, COMMUNITY 2"
6090 PRINT #2, " ..."
6100 CLOSE 2
6110 RETURN
```