Oecologia (Berl) (1981) 50: 296–302

Oecologia

Similarity Indices, Sample Size and Diversity

Henk Wolda

Smithsonian Tropical Research Institute, P.O. Box 2072 Balboa, Republic of Panama

Summary. The effect of sample size and species diversity on a variety of similarity indices is explored. Real values of a similarity index must be evaluated relative to the expected maximum value of that index, which is the value obtained for samples randomly drawn from the same universe, with the diversity and sample sizes of the real samples. It is shown that these expected maxima differ from the theoretical maxima, the values obtained for two identical samples, and that the relationship between expected and theoretical maxima depends on sample size and on species diversity in all cases, without exception. In all cases but one (the Morisita index) the expected maxima depend strongly to fairly strongly on sample size and diversity. For some of the more useful indices empirical equations are given to calculate the expected maximum value of the indices to which the observed values can be related at any combination of sample sizes. It is recommended that the Morisita index be used whenever possible to avoid the complex dealings with effects of sample size and diversity; however, when previous logarithmic transformation of the data is required, which often may be the case, the Morisita-Horn or the Renkonen indices are recommended.

1. Introduction

It is often desirable to make comparisons between faunal or floral samples taken at different times, in different places, or by different techniques, whether by the investigator, by predators or by herbivores. In making such comparisons it seems profitable to take advantage of the existence of similarity indices, many of which have been developed in this century. Some of these indices merely take into account the presence or absence of species in the samples while others incorporate information on the relative abundance of the species. The preferable index in a given case depends on the questions asked and the kind of data available. However, it is known that at least some of the indices depend on sample size (Williams 1949; Mountford 1962; Morisita 1959), while the diversity of the samples may also have an effect (Williams 1949). Huhta (1979) tested a long series of similarity indices using real data and found that the results depend largely on the index chosen, which suggests the dangerous possibility that one can choose an index to demonstrate whatever one wants the data to show, without necessarily being able to prove that this is indeed what they do show. Thus, careful evaluation fo the various coefficients is essential.

To interpret a given value of a similarity index one must compare it with its maximum value. As that maximum value, one usually takes the theoretical maximum value, which is the value obtained when comparing two identical samples. That value usually is one. However, this does not seem to be a very useful procedure. When comparing two samples one wants to test the null hypothesis that they are random samples from the same fauna against the alternative hypothesis that they are samples from different faunas. The expected maximum value of the similarity index should, therefore, be the value of that index for two random samples from the same fauna. This value, as will be shown, can be very different from the theoretical maximum value.

This paper examines the effects of sample size and diversity on the expected maximum values of a number of similarity indices, using computer-generated samples taken from insect faunas that are distributed according to the log series with four different diversities (Fisher et al., 1943). Included in the analysis are indices known to be dependent on sample size to evaluate the extent and effect of that influence and to explore the possibility of taking it into account. Also included are indices that have been claimed to be almost independent of sample size (Mountford 1962; Morisita 1959; Horn 1966).

It will be shown that the expected maximum of all indices but one are rather strongly affected by sample size and diversity. In some indices this influence is greater than in others and an attempt will be made to deal with these influences.

2. Materials and Methods

Simulation experiments were carried out on an Infotek-improved HP-9830 desk computer. Four hypothetical faunas, distributed according to the log series, were specified. In the log series distribution the relation between the number of individuals and the number of species is given by:

$$S=\alpha\ln\left[1+\frac{N}{\alpha}\right]$$

where S is the number of species, N the number of individuals and α the diversity coefficient. Each fauna from which the samples were to be drawn had 100,000 individuals and the number of species varied from 150 ($\alpha=17.3$), 380 ($\alpha=50.0$), 580 ($\alpha=81.5$) to 750 ($\alpha=110.1$). The highest diversity used here is slightly less than that observed in a sample of Homoptera collected over four years of light-trapping in the tropical forest on Barro Colorado Island, Panama (Wolda, in press 1981) and is therefore feasible.

0029-8549/81/0050/0296/$01.40

Random samples were taken from each of these four faunas without replacement; they were of five different sizes (100, 200, 500, 1,000 and 5,000 individuals) with five replicates for each of the four smallest sample sizes and two replicates for size 5,000. (A copy of the computerprogram used to draw random samples out of a specified fauna, which is distributed according to the log series, can be obtained on request.)

For each diversity, similarity indices were calculated between all possible pairs of samples, and mean and standard deviation of the indices were calculated for each combination of sample sizes.

In the similarity indices included in this study:

a = the number of species in sample 1
b = the number of species in sample 2
c = the number of species in common between 1 and 2
d = the number of species absent in both 1 and 2
k = the number of species in 1 and 2 combined $= a+b-c$.
n_{ji} = the number of individuals of species i in sample j
N_j = the number of individuals in sample j
p_{ji} = the proportion of species i in sample $j = n_{ji}/N_j$

A number of binary coefficients are used when the number of species absent in both samples can be specified. Of these I have selected only one example:

1) Baroni-Urbani and Buser (1976)

$$S_0 = \frac{\sqrt{cd}+c}{\sqrt{cd}+a+b-c}.$$

Of the other binary coefficients where the number of species absent in both samples to be compared need not be known, I have selected three examples:

2) Czekanowski (1913), better known as Sørensen (1948)

$$QS = \frac{2c}{a+b}.$$

3) Mountford (1962)

$$I = \frac{2c}{2ab-(a+b)c}.$$

4) Association index, Dice (1945)

$$M = \frac{c}{\min(a,b)}.$$

For a survey of other binary indices see Clifford and Stephenson (1975) and Cheetham and Hazel (1969).

I have selected several of the coefficients that take into account the relative abundance of the species:

5) Bray and Curtis (1957)

$$1-BC = 1-\frac{\sum |n_{1i}-n_{2i}|}{\sum (n_{1i}+n_{2i})}$$

6) Bray and Curtis (1957) after logarithmic transformation of the data $[\ln(n_{ij}+1)]$. Same formula as (5).

7) Canberra metric (Lance and Williams 1976)

$$1-CM = 1-\frac{1}{k}\sum \frac{|n_{1i}-n_{2i}|}{(n_{1i}+n_{2i})}.$$

Double zero records are ignored. When one element in a pair equals zero there can be a problem (Clifford and Stephenson 1975, p. 58) so that the zero in such cases is replaced by 0.2. The index is also calculated without this replacement.

8) Canberra metric after logarithmic transformation of the data $[\ln(n_{ji}+1)]$. Same formula as in (7).

9) Squared Euclidian distance (Clifford and Stephenson 1975, p. 65)

$$1-D^2 = 1-\sum (p_{1i}-p_{2i})^2.$$

10) Squared Euclidian distance after logarithmic transformation of the data $[\ln(n_{ji}+1)]$. Formula as in (9).

11) Percentage similarity (Renkonen 1938)

$$PS = \sum \min(p_{1i}, p_{2i}).$$

12) Percentage similarity after logarithmic transformation of the data $[\ln(n_{ji}+1)]$. Formula as in (11).

13) Morisita index (Morisita, 1959)

$$C_\lambda = \frac{2\sum n_{1i}n_{2i}}{(\lambda_1+\lambda_2)N_1N_2}, \quad \text{where} \quad \lambda_j = \frac{\sum n_{ji}(n_{ji}-1)}{N_j(N_j-1)}.$$

14) Simplified Morisita index (Horn 1966)

C_λ as in (13) except $\lambda_j = \frac{\sum n_{ji}^2}{N_j^2}$.

15) Simplified Morisita index after logarithmic transformation of the data $[\ln(n_{ji}+1)]$. Formula as in (14).

16) Index of overlap (Horn, 1966)

$$R_0 = \frac{H'_{\max}-H'_{\mathrm{obs}}}{H'_{\max}-H'_{\min}}$$

where H' is the Shannon-Weaver diversity index.

17) Product-moment correlation coefficient. See any book on statistics for the formula.

18) Product-moment correlation coefficient after logarithmic transformation of the data $[\ln(n_{ji}+1)]$.

19) Kendall rank correlation coefficient. For methods of calculation see Ghent (1963; 1972) or any book on statistics.

20) Kendall rank correlation coefficient with $\sum(P+Q)$ in the denominator (Ghent 1972).

21) Annual variability, including only data pairs with both $n_{ji} > 4$ (Wolda 1978).

$$AV = \mathrm{Var}(\log n_{2i}/n_{1i})$$

22) Annual variability for all n_{ji} values (Wolda 1978). Same formula as in (21).

Results

One useful property of a similarity index is that it increases linearly from some fixed minimum to some fixed, finite maximum. As a criterion I have used two samples of 100 species each, each species represented by only one individual. I varied the number of species both samples have in common (c) and calculated the values of the various indices. As c increases from zero to 100, most indices increase linearly from zero to their theoretical maximum of one (Fig. 1), but this is not true for all indices. The Squared Euclidian Distance (SED) invariably has a very high value. Even when the two samples are completely different, i.e. $c=0$, this index is 0.98. For this reason this index is not very useful when comparing two different faunal samples and will be discussed no further.

Mountford's index (I) remains at low values over a broad range of values of c, then sharply increases at high values of c to reach 0.99 at $c=99$, and shoots up to infinity when $c=100$. This does not argue in favour of the Mountford index.

The product-moment correlation coefficient does increase linearly with increasing c, but from -1 to zero instead of from zero to $+1$. Moreover, when the number of individuals is not the same in all species these limits change to -1 to $+1$. In other words, the theoretical limits of the correlation coefficient are not fixed but depend on the kind of samples at hand, which does not make it attractive as a similarity index.

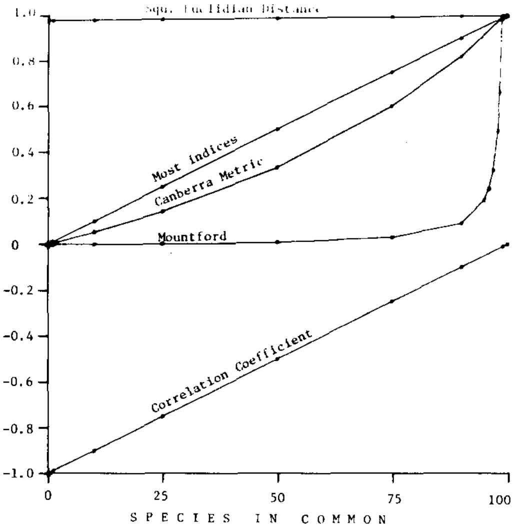

Fig. 1. Linearity of similarity indices. Two samples of 100 species each are compared, each species with one individual. The number of species in common (*c*) is plotted along the abscissa, the value of the indices along the ordinate

The Kendall rank correlation coefficient also varies in the above example (Fig. 1) between -1 at $c=0$ and zero at $c=100$, but because of the large number of ties the coefficient is indeterminate at intermediate values of *c* unless the species are ordered in a fixed way and it is known which species are present in each sample.

The Canberra metric does increase from zero to one as *c* increases from zero to 100, but in a non-linear manner. A similar test was done for the Annual Variability (AV) measure. Again two samples of 100 species each were compared, but the number of individuals in each species was taken as 5. The resulting curve was convex with its maximum value at $c=0$, i.e. when both samples are completely different.

All other indices increase linearly between zero and one as *c* increases from zero to 100, which seems to give them an advantage over the indices in which this is not the case. However, it is not the theoretical maximum to which one should relate a given index, but the expected maximum under the null hypothesis that the two samples are random samples from the same fauna. The computer-generated faunal samples were used to explore the behaviour of various similarity indices and the effects of sample size and species diversity.

The diversity of the faunas from which the samples were drawn was specified, but since this is usually an unknown quantity, one has to estimate the diversity from the samples at hand. As expected, diversity coefficients such as the Shannon-Weaver index H' varied with sample size; however, in the log series distribution the coefficient α is presumed independent of sample size. This was tested using the samples present and the result are given in Table 1. The diversity of the samples was found to be that of the source fauna, with no effect of sample size. There is, of course, an effect on the variance so that the sample estimate of α becomes less reliable the smaller the sample.

Within each fauna, similarity indices were calculated for all possible combinations of samples, which gave a total of 231 values for each index. For each combination of sample sizes a mean value was calculated for each index. These means are based on one value for the combination of sample sizes 5,000 × 5,000, and on 10 values for all other combinations with 5,000 and for all combinations not including 5,000, where the two samples are equal in size. For all other combinations of sample sizes, i.e. the combinations where the two sample sizes are different and do not include 5,000, there are 25 values. For the smallest and largest diversities tested ($\alpha=17.31$ and $\alpha=110.1$) the results are plotted in Fig. 2. The figures for the intermediate diversities are not shown here ($\alpha=50.0$ and $\alpha=81.54$), but are intermediate between these two extremes.

Table 1. The relation between sample size and the α-diversity in random samples taken from a log series distribution

Sample size	*N*	α-diversity		
		Mean	SD	Range
Expected diversity 17.31:				
5,000	2	17.82	0.148	17.71- 17.92
1,000	5	17.61	1.870	13.99- 18.78
500	5	16.90	0.844	16.16- 18.22
200	5	18.96	3.277	15.61- 23.55
100	5	18.20	1.293	17.18- 20.35
Expected diversity 50.0:				
5,000	2	48.56	0.962	47.88- 49.24
1,000	5	50.49	2.739	46.13- 52.31
500	5	49.50	4.798	44.85- 57.05
200	5	47.43	5.145	43.58- 51.91
100	5	48.45	7.839	36.26- 55.04
Expected diversity 81.54:				
5,000	2	81.73	3.825	79.02- 84.43
1,000	5	85.36	3.664	79.27- 88.44
500	5	81.52	10.243	72.65- 97.75
200	5	81.28	11.451	70.83-100.32
100	5	74.69	12.215	60.42- 88.72
Expected diversity 110.1:				
5,000	2	114.85	2.475	113.1 -116.6
1,000	5	114.56	4.318	109.1 -120.7
500	5	110.80	11.929	92.7 -126.0
200	5	113.26	10.267	105.1 -126.7
100	5	114.0	24.557	88.9 -154.5

The expected maxima of the similarity indices tested (results given in Fig. 2) are invariably strongly dependent on sample size, to the extent that this effect cannot be ignored, and the effect increases with faunal diversity. Apart from both Annual Variability measures (21 and 22), which have an erratic relation between sample size and the index, the relationships found fall into two broad types. In one type the effects of the smallest (*S*) and the largest (*L*) samples are additive in the sense that an increase in either one of them causes an increase in the value of the index (e.g. indices 11, 14, 16, 17, 18, 19 and 20). The other type of relationship shows an increase in the value of the index with an increase in *S*, but a decrease with an increase in *L* (most other indices in Fig. 2). In the latter type it is the ratio between the two sample sizes that is important. This difference is important since in the former type, the one in which the effects of *S* and *L* are additive, one can confidently extrapolate to index values at larger sample sizes, while in the latter this would not be possible if the sample sizes were very different. For two random samples from the same fauna, one of 5,000,000 and one of 5,000 individuals, one could predict that Renkonen's PS (11) value would be near unity, but one could not make such a prediction for the Bray-Curtis (5) index.

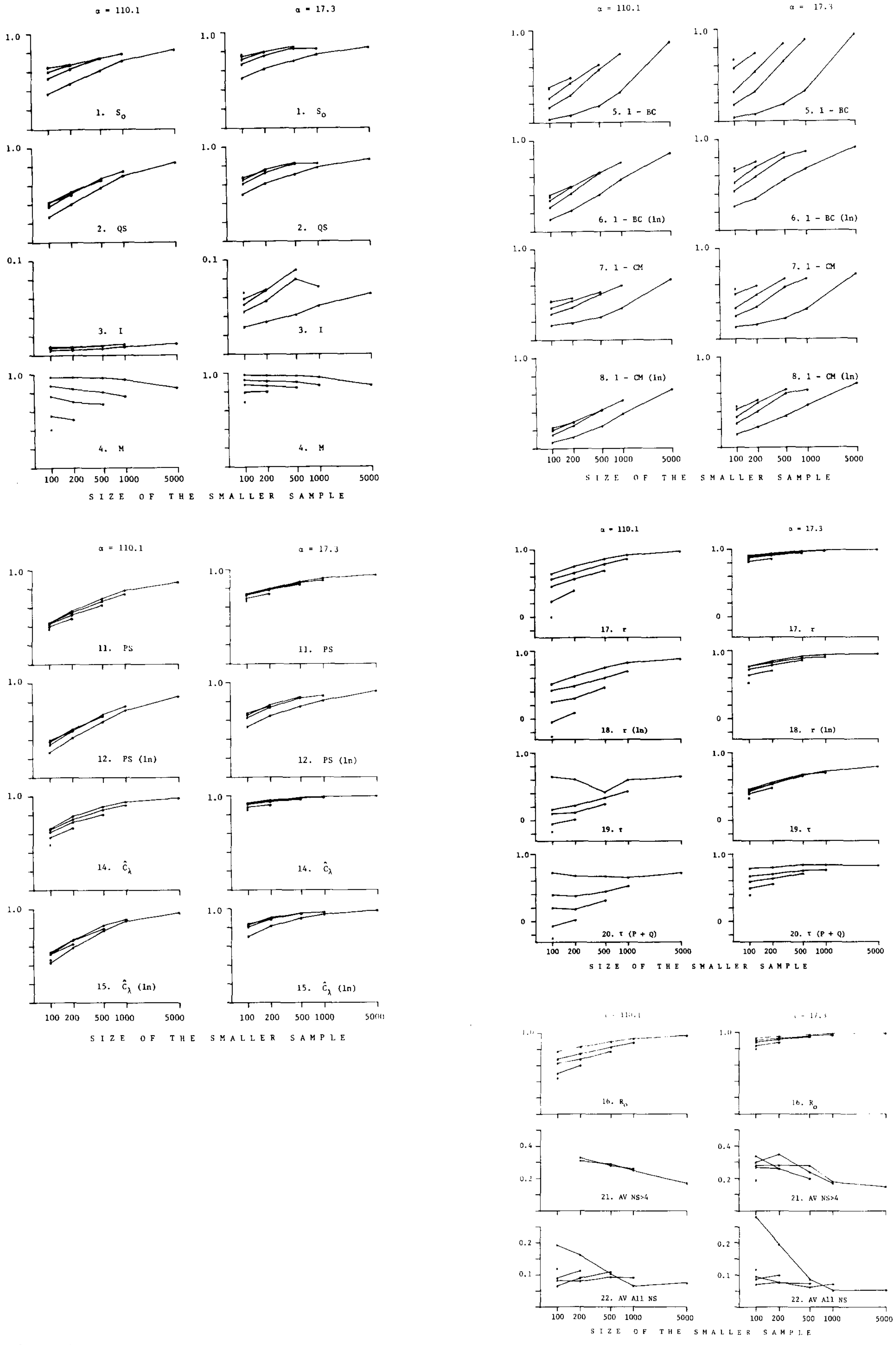

Fig. 2. The value of the expected maximum value of a number of similarity indices as a function of the size of the smaller sample (S, plotted along the abscissa) and the larger sample (L). Plots for samples with the same L are connected by a line. The longest line refers to $L = 5{,}000$, the next longest line to $L = 1{,}000$, etc. At $L = 100$ there is only one point. The indices are indicated by their symbol and by the same number used in Materials and Methods where these indices are described. For each index the graph on the left refers to samples from a very diverse fauna ($\alpha = 110.1$) and the one on the right to a much less diverse fauna ($\alpha = 17.3$)

Table 2. Equations approximating the relation between the maximum values of some similarity indices and sample size, i.e. the number of individuals in the smaller (S) and the larger (L) of the two samples compared. The goodness of fit, i.e. the relation between the simulation results given in Fig. 1 and the results using these equations is given as r^2 (in percentages)

Index	Alpha Diversity	Regression equation	$100 \cdot r^2$
1. Baroni-Urbani and Buser	17.31	$S_0 = 1.190 - 1.563\ S^{-0.265} - 389 \cdot 10^{-7} L$	92.6
	50.0	$S_0 = 1.190 - 2.108\ S^{-0.301} - 389 \cdot 10^{-7} L$	87.7
	81.54	$S_0 = 1.208 - 2.204\ S^{-0.288} - 432 \cdot 10^{-7} L$	86.5
	110.1	$S_0 = 1.213 - 2.651\ S^{-0.312} - 438 \cdot 10^{-7} L$	90.2
2. Czekanowski (Sørensen)	17.31	$QS = 1.148 - 2.146\ S^{-0.322} - 301 \cdot 10^{-7} L$	96.1
	50.0	$QS = 1.130 - 3.292\ S^{-0.364} - 264 \cdot 10^{-7} L$	84.2
	81.54	$QS = 1.137 - 3.375\ S^{-0.347} - 281 \cdot 10^{-7} L$	95.2
	110.1	$QS = 1.125 - 4.170\ S^{-0.375} - 251 \cdot 10^{-7} L$	97.4
4. Association index, Dice	17.31	$M = 0.592 - 0.0256 \ln S + 0.0632 \ln L$	87.8
	50.0	$M = 0.340 - 0.0372 \ln S + 0.1020 \ln L$	90.3
	81.54	$M = 0.254 - 0.0426 \ln S + 0.1153 \ln L$	90.0
	110.1	$M = 0.020 - 0.0348 \ln S + 0.1377 \ln L$	92.5
11. Renkonen	17.31	$PS = 1 - 1.642\ S^{-0.405} - 4.282\ L^{-0.866}$	99.0
	50.0	$PS = 1 - 2.410\ S^{-0.384} - 2.754\ L^{-0.719}$	99.1
	81.54	$PS = 1 - 2.810\ S^{-0.375} - 0.645\ L^{-0.438}$	99.0
	110.1	$PS = 1 - 3.111\ S^{-0.375} - 0.640\ L^{-0.470}$	98.7
12. Renkonen (log)	17.31	$PS_{\ln} = 1.121 - 2.827\ S^{-0.392} - 245 \cdot 10^{-7} L$	98.0
	50.0	$PS_{\ln} = 1.109 - 3.507\ S^{-0.388} - 222 \cdot 10^{-7} L$	96.7
	81.54	$PS_{\ln} = 1.099 - 4.240\ S^{-0.393} - 198 \cdot 10^{-7} L$	97.5
	110.1	$PS_{\ln} = 1.098 - 4.088\ S^{-0.379} - 197 \cdot 10^{-7} L$	98.4
14. Morisita-Horn	17.31	$\hat{C}_\lambda = 1 - 3.663\ S^{-0.823} - 11.382\ L^{-1.070}$	97.8
	50.0	$\hat{C}_\lambda = 1 - 8.761\ S^{-0.803} - 18.165\ L^{-1.020}$	99.8
	81.54	$\hat{C}_\lambda = 1 - 10.651\ S^{-0.765} - 10.668\ L^{-0.882}$	99.3
	110.1	$\hat{C}_\lambda = 1 - 9.425\ S^{-0.718} - 5.959\ L^{-0.758}$	99.0
15. Morisita-Horn (log)	17.31	$\hat{C}_{\lambda\ln} = 1.096 - 5.538\ S^{-0.640} - 197 \cdot 10^{-7} L$	95.6
	50.0	$\hat{C}_{\lambda\ln} = 1.070 - 9.434\ S^{-0.658} - 144 \cdot 10^{-7} L$	95.2
	81.54	$\hat{C}_{\lambda\ln} = 1.039 - 9.414\ S^{-0.624} - 77 \cdot 10^{-7} L$	95.8
	110.1	$\hat{C}_{\lambda\ln} = 1.071 - 10.520\ S^{-0.622} - 139 \cdot 10^{-7} L$	96.5
16. Overlap, Horn	17.31	$R_0 = 1 - 1.247\ S^{-0.631} - 6.486\ L^{-0.835}$	98.1
	50.0	$R_0 = 1 - 1.799\ S^{-0.539} - 9.393\ L^{-0.772}$	98.7
	81.54	$R_0 = 1 - 1.802\ S^{-0.485} - 5.825\ L^{-0.639}$	98.8
	110.1	$R_0 = 1 - 2.556\ S^{-0.517} - 7.040\ L^{-0.646}$	97.9

Both the product-moment correlation coefficient and the Kendall rank correlation coefficient (17, 18, 19, 20) show a strong effect of sample size, to the extent that the expected maximum values can be negative at higher diversities, which makes them unpleasant to work with. The effects of sample size on the Bray-Curtis index (5, 6) and the Canberra metric (7, 8) is also excessive. In the binary indices the Mountford index (3) has such low expected maximum values (note the scale of the ordinate) that it is also almost useless.

The Morisita index (Morisita 1959), is not included in Fig. 2. Morisita performed simulation experiments not unlike the more detailed and extensive ones presented here and concluded that his index 'is almost uninfluenced by the sizes of N_1 and N_2 unless either or both of N_1 and N_2 are small.' This index has been criticized for having a maximum not equal to one, but of 'about one' (Horn 1966). In fact, the theoretical maximum of the index is always larger than one and is strongly dependent on sample size (Fig. 3). This theoretical maximum can only be calculated for two samples of equal size and, therefore, is unknown for combinations of samples of different sizes. However, the expected maximum values are 'about one' and are independent of sample size, which confirms Morisita's results. The actual value of the expected maximum is not exactly unity and may even tend to be a little above one, but it should be close enough to one so that this index can be used without corrections for effects of sample size. The uncertainties in not having a fixed upper limit for the index equal to one are outweighed by the problems of correcting the other indices for effects of sample size and diversity.

In spite of the obvious advantage of the Morisita index, it may be desirable to use some other index. For instance, if the data need to be transformed to logarithms before using an index or if a binary index is required, the values observed should be related to the expected maximum of that index at the observed diversity and sample sizes. That expected maximum can be calculated by simulating a number of samples of the desired sizes and diversity with a computer. This procedure is, however, very laborious and the information contained in Fig. 2 can be used to estimate the expected maximum of the index concerned. The more useful indices were selected and an equation was fitted to the data expressing the index as a function of the smallest (S) and largest (L) sample size. The results are given Table 2 together with the goodness

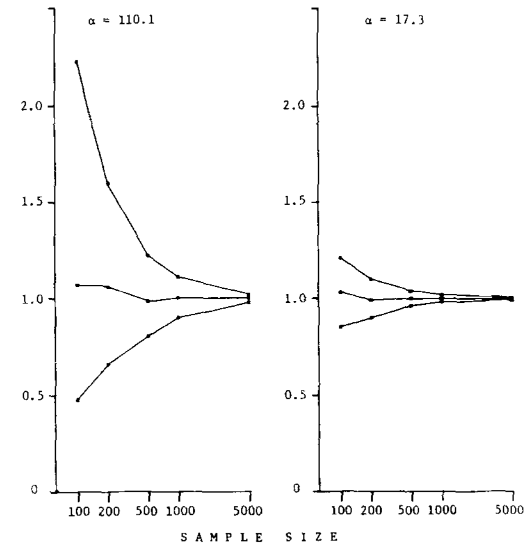

Fig. 3. The Morisita index as a function of sample size at two diversities. The theoretical maximum value obtained when comparing two identical samples is highly dependent on sample size (*top line*). The expected maximum obtained when comparing two random samples from the same fauna is independent of sample size and diversity and is about one (*middle line*). The expected maximum expressed as a fraction of the theoretical maximum is plotted in the bottom line

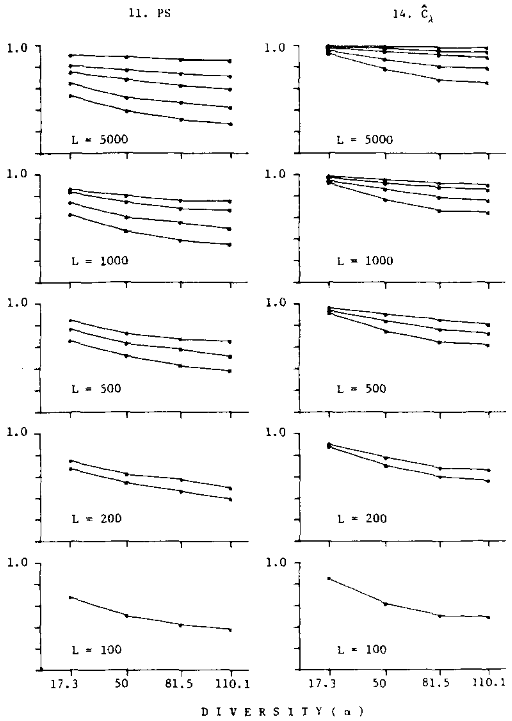

Fig. 4. Relation between the expected maximum value of some similarity indices (ordinate) and diversity (α, abscissa). Renkonen's *PS* left and the Morisita-Horn index on the right. For each larger sample size (*L*) there is a graph and within each graph a line for each smaller sample size (*S*). The lowest line in each graph is for $S=100$, the next lowest line, if there is one, is for $S=200$, then $S=500$, $S=1,000$ and $S=5,000$

of fit (r^2) of the equations to the data of Fig. 2. For the non-binary indices the fit is good to superb and for the binary indices reasonable to good. Several types of regressions were tried and the ones given gave the best fit.

As shown for two selected indices in Fig. 4, the relation between diversity and the effect of sample size is not complicated, and linear intrapolation between two adjacent diversities covered in Table 2 should cause no problem. The same holds for the other indices. One can, therefore, determine the expected maximum at the two diversities adjacent to the diversity of the samples and intrapolate.

Discussion

The Baroni-Urbani and Buser index requires that the number of species absent in both samples be known, which makes it normally impractical. However, if such extra information is available, such as in the present simulation experiments, it does not improve the effect of sample size on the index.

The difference between the Czekanowski (Sørensen) index *QS* and the association index (Dice) *M* is considerable. For example, if there are two samples, one with 150 and the other with 30 species and they have 30 species in common, *M* is 1 but *QS* only 0.333. In the Czekanowski index the value of *c*, the number of species both samples have in common, is compared with the total number of species in both samples and in the association index, with the number of species in the smallest sample only. If, therefore, there is reason to believe that the difference between the two samples is completely or largely due to differences in sample size, the association index might be preferred. If, however, there is reason to believe that the difference between the samples is not due to differences in sample size but is real, one may find the Czekanowski index more useful.

The fit of the equations for the binary indices in Table 2 is not perfect, so another index might be preferred. In many cases, however, binary indices are the only ones that can be used as no information on the relative abundance of the species is available. One often has only lists of species to compare. In such cases the interpretation of the (binary) similarity index is difficult as no correction for sample size is possible and no estimate of alpha diversity can be obtained.

Among the non-binary indices the Morisita index has a major advantage in that it is independent of sample size and diversity, except possibly for very low sample sizes (Morisita 1959), but then sampling error will be such that a similarity index may not convey much information anyway. If other indices are preferred, one can consider Renkonen's *PS* with or without previous logarithmic transformation of the data, the Morisita-Horn index with or without a previous logarithmic transformation, and the Horn index of overlap. The equations in Table 2 help to evaluate the indices found within the range of sample sizes and diversities used in this paper.

Extrapolation of the present results to larger sample sizes is no problem with Renkonen's index, the Morisita-Horn index, and the Horn overlap index, provided the data are not previously transformed to logarithms. Extrapolation with such transformation is possible if both sample sizes are large, but if one is much smaller than the other this should be discouraged until the effects of sample size at those larger sizes have been established through simulation experiments. Extrapolation to smaller sample sizes should also be avoided. If the samples are smaller than 100, there is no point in calculating a similarity index because sampling error makes the results meaningless unless the diversity of the samples is very low.

In all experiments reported in the present paper, the samples compared were drawn from faunas with the same diversity; no comparisons have been made of samples with different diversities. However, such differences are not uncommon. To deal with those samples, the average diversity of the two samples may be used in calculations. The results should be acceptable unless, perhaps, the difference in diversity between the two samples is very large.

In all simulations the faunas were assumed to be distributed according to the log series. Field samples may differ from this distribution (Wolda and Fisk in preparation), but this deviation should not affect the results greatly. With data which are very different from a log series, care should be taken in applying the present results. However, the distribution in most cases should be close enough to a log series to enable one to use the present results as an approximation.

In field samples a positive correlation has sometimes been observed between the value of a similarity index and the number of individuals in the samples. In his study on the seasonal occurrence of coprophagous beetles Hanski (1980) found such a correlation for the Renkonen's index and suggested various interesting biological consequences of this correlation. He also mentioned the 'possibility of an artefact' and indeed this correlation is an inherent property of this index and does not warrant any biological conclusions.

Morisita (1959) showed that his index is virtually independent of sample size, and the present paper not only confirms this but also shows that it is independent of diversity. Moreover, to the best of my knowledge, it is the only index that has these properties. Therefore, it deserves a much greater popularity than it enjoys at present. Investigators comparing different faunal or floral samples should use this index in preference to the other indices with their problems of sample size and diversity. However, this index is very sensitive to changes in abundance of the more common species and often produces an erratic picture. In such cases a logarithmic transformation of the data is desirable and then the Morisita-Horn index or Renkonen's index are recommended, using the equations in Table 2 to correct for effects of sample size.

Different species usually behave differently and can have strongly variable effects on their resources such as food. A numerical difference, therefore, between two localities in one species may be much more 'important' than the difference in another species. Differences in 'importance' between species have not been considered here, but the difference between samples could be weighed according to some attribute of the species like size, dependent on the questions being asked.

Whatever the index used, one would like to be able to test hypotheses about the values obtained. Regrettably, to the best of my knowledge, such statistical tests are not yet available, although a beginning has been made for the Annual Variability measure (Leigh in preparation). It is hoped that the present paper may help persuade statisticians to come to the rescue.

References

Baroni-Urbani C, Buser MW (1976) Similarity of binary data. Syst Zool 25:251-259

Bray JR, Curtis JT (1957) An ordination of the upland forest communities in southern Wisconsin. Ecol Monogr 27:325-349

Cheetham AH, Hazel JE (1969) Binary (presence-absence) similarity coefficients. J Palaeont 43:1130-1136

Clifford HT, Stephenson W (1975) An introduction to numerical classification. Academic Press, New York-San Francisco-London

Czekanowski J (1913) Zarys Metod Statystycnck. Warsaw: E. Wendego. See also: Coefficient of racial likeness and durchschnittliche Differenz. Anthropol Anz 9:227-249 (1922)

Dice LR (1945) Measures of the amount of ecological association between species. Ecology 26:297-302

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12:42-58

Ghent AW (1963) Kendall's "Tau" coefficient as an index of similarity in comparisons of plant and animal communities. Canad Entomol 95:568-575

Ghent AW (1972) A graphic computation procedure for Kendall's Tau suited to extensive species-density comparisons. Am Midl Nat 87:459-471

Hanski I (1980) Spatial variation in the timing of the seasonal occurrence in coprophagous beetles. Oikos 34:311-321

Horn H (1966) Measurement of "overlap" in comparative ecological studies. Am Nat 100:419-424

Huhta V (1979) Evaluation of different similarity indices as measures of succession in arthropod communities of the forest after clearcutting. Oecologia (Berl) 41:11-23

Lance GN, Williams WT (1967) Mixed-data classificatory programs. I. Agglomerative systems. Aust Comput J 1:15-20

Mountford MD (1962) An index of similarity and its application to classificatory problems. In: PW Murphy (ed), Progress in Soil Zoology. Butterworths, London, pp. 43-50

Morisita M (1959) Measuring of interspecific association and similarity between communities. Mem Fac Sci Kyushu Univ, Ser E. Bio., 3:65-80

Renkonen O (1938) Statistisch-ökologische Untersuchungen über die terrestische Käferwelt der finnischen Bruchmoore. An Zool Soc Zool-Bot Fenn Vanamo 6:1-231

Sørensen TA (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on Danish commons. K dan Vidensk Selsk Biol Skr 5:1-34

Williams CB (1949) Jaccard's generic coefficient and coefficient of floral community, in relation to the logarithmic series and the index of diversity. Ann Bot 13:53-58

Wolda H (1978) Fluctuations in abundance of tropical insects. Am Nat 112:1017-1045

Wolda H (1981) Seasonality of leafhoppers (Homoptera) on Barro Colorado Island. In: EG Leigh, AS Rand, DM Windsor (eds), Ecology of a tropical forest: Seasonal rhythms and longterm changes. Smithsonian Press, Washington (in press)

Wolda H, Fisk FW (1981) Seasonality of tropical insects II. Blattaria (cockroaches) from the seasonal tropics of Panama. J Anim Ecol (in press)

Received November 3, 1980