Vol. 9: 147–151, 1982 | MARINE ECOLOGY – PROGRESS SERIES
Mar. Ecol. Prog. Ser. | Published July 31

Sample Size Dependence in Measures of Proportional Similarity

Alan J. Kohn and Alan C. Riggs

Department of Zoology, University of Washington, Seattle, Washington 98195, USA

ABSTRACT: The 2 commonly presented measures of proportional similarity differ in that one (PS_I) is independent of the relative sizes of the samples being compared, but the other (PS_D) is not. Using hypothetical data sets in which the proportions of entries in each sample remain constant but the ratio of total sample sizes is varied, we characterize the differences between these measures. Both give the same value when the 2 sample sizes are equal. When the ratio of sample sizes $\neq 1$, the value of PS_D is usually less than the value of PS_I. Since PS_D is often affected more by sample size ratio than by proportionate compositional similarity (it may be a function of sample size ratio only), we suggest the use of PS_I, except when it is intended that the index reflect sample size differences.

INTRODUCTION

Two measures of proportional similarity formalized by Whittaker (1952) are often used to compare composition and relative abundance of species in 2 assemblages, and the overlap in use of a set of resources by two co-occurring species. In recent critical reviews of matrices of similarity, 2 quite different approaches have led to the conclusion that one of these measures, here designated PS_I, is often preferable to other indices of similarity that have been proposed. Pielou (1979) developed an objective test that resulted in the conclusion that PS_I was superior to several other indices based on quantitative and presence-absence data. Bloom (1981) showed that other indices deviate considerably, in different directions, and asymmetrically from PS_I. Abrams (1980) listed several additional advantages of this measure over others proposed to estimate overlap in resource use.

Our purposes are (1) to point out that while one of Whittaker's formulas gives values independent of sample sizes, the other, here designated PS_D, is a function of relative sample size as well as of similarity in composition and frequency; (2) to show how varying relative sample size affects the distribution of the latter measure.

In his original formalization of the sample size-independent measure (PS_I), first used by Renkonen (1938), Whittaker (1952) noted its applicability to assessing similarity of population distributions of 2 species across samples as well as similarity of samples with respect to component species. Several recent studies have used equivalent formulas as measures of proportional similarity of species composition of marine communities (Rex, 1977; Pielou, 1979; Weinstein et al., 1980) and of resource utilization by 2 species (Huey et al., 1974; Sale, 1974; Price and Willson, 1976; Hanski, 1978; Hurlbert, 1978; Sabo and Whittaker, 1979; Leviten and Kohn, 1980) or by 2 size classes within a species (Leviten, 1978; Davies et al., 1979). The formula of the sample size-independent measure is

$$PS_I = 1 - 0.5 \sum_{i=1}^{s} \left| p_{x,i} - p_{y,i} \right| = \Sigma \min\,(p_{x,i},\, p_{y,i}) \qquad (1)$$

where $p_{x,i}$ = proportion of species i in sample X, or the proportional utilization by species x of the i^{th} resource category, and there are S species or resource categories. The $p_{y,i}$ are defined similarly for sample Y. In terms of the hypothetical data presented in Table 1, $p_{x,i} = \frac{x_i}{X}$ and $p_{y,i} = \frac{y_i}{Y} \cdot \sum_i p_{x,i} = \sum_i p_{y,i} = 1$.

As Whittaker (1952: 11) pointed out, 'The index thus measures the extent to which two samples are alike in composition, the fraction of their totals in which they are alike in percentages of individuals of the various species, and ranges from zero for samples with no species in common to 1.00 or 100 percent for identical samples.' $PS_I = 1$ for 2 samples with identical propor-

© Inter-Research/Printed in F. R. Germany

0171-8630/82/0009/0147/$ 02.00

Table 1. Initial composition of pairs of assemblages in hypothetical data sets used in demonstrating the dependence of PS_D on relative sample size (Figs. 1 and 2). For comparisons of distributions of species over resource states, *i* designates resource states and x_i, y_i are the numbers of species X and Y using each resource state

Figs. 1, 2	Species (*i*)	Number of ind. of species *i* in assemblages X	Y	1.0 x_i	y_i
	1	x_1	y_1	10	10
	2	x_2	y_2	20	20
	3	x_3	y_3	30	30
	4	x_4	y_4	40	40
	5	x_5	y_5	50	50
	S	X	Y	X = 150	Y = 150
				Z = 300	

Fig. 1: The same species are present in both samples; their proportions differ

Initial PS_D	0.8		0.6		0.5		0.2		0.16			
	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
Initial samples	50	5	10	50	50	17	1	1	10	90	1	1
	50	65	20	40	50	17	3	3	10	90	3	81
	50	60	30	30	17	50	9	27	90	10	6	27
	40	50	40	20	17	50	27	81	90	10	27	6
	30	40	50	10			81	9			81	3
	X = 220	Y = 220	X = 150	Y = 150	X = 134	Y = 134	X = 121	Y = 121	X = 200	Y = 200	X = 118	Y = 118
	Z = 440		Z = 300		Z = 268		Z = 242		Z = 400		Z = 236	

Fig. 2: Both species composition and proportions differ in the 2 samples

Initial PS_D	0.8		0.5		0.4		0.3		0.2		0.17	
	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
Initial samples	0	23	0	50	70	0	80	0	85	0	90	10
	20	20	20	40	70	20	80	20	85	10	90	10
	30	30	30	30	20	20	20	20	20	20	10	90
	40	40	40	20	20	80	10	80	5	80	5	90
	23	0	50	0	20	80	10	80	5	90	5	0
	X = 113	Y = 113	X = 140	X = 140	X = 200	Y = 200	X = 200	Y = 200	X = 200	Y = 200	X = 200	Y = 200
	Z = 226		Z = 280		Z = 400		Z = 400		Z = 400		Z = 400	

tions of all species, regardless of the absolute numbers of individuals in the samples.

Whittaker (1952) derived Expression (1) as a special case of a sample size-dependent measure (PS_D):

$$PS_D = 1 - \frac{\Sigma \mid x_i - y_i \mid}{\Sigma\ (x_i + y_i)} = \frac{2\ \Sigma \min\ (x_i,\ y_i)}{\Sigma\ (x_i + y_i)} \qquad (2)$$

where x_i and y_i = numbers of the i^{th} species in the 2 samples. Pielou (1975: 100) noted that this measure attains its maximum value only if the total number of individuals, as well as the abundance of species, is the same in both samples. She did not state this explicitly in her 1977 book, but it is the important difference between Expressions (1) and (2). Although he proposed (2) for use 'when sample size is standardized for equal areas', i. e. not explicitly for identical sample sizes, Whittaker (1952) recognized that values of (2) depend on the relative size of the samples.

Gallaher and Blake (1977) demonstrated the mathematical relationships between (1) and (2); they indicated a preference for the latter. A number of recent studies have used (2) (Dauer and Simon, 1975; Boesch, 1977; Dean, 1981) or its complement, a measure of dissimilarity (Bernstein et al., 1978; Schoener and Greene, 1981; earlier citations in Clifford and Stephenson, 1975). Bloom (1981) gave Eq. (2) but transformed it to the equivalent of Eq. (1) for his analysis.

METHOD

To illustrate how relative sample size affects the values of the 2 measures, we generated a series of hypothetical data sets (Table 1) in a manner similar to those of Gallaher and Blake (1977) and Hurlbert (1978). In each plot of these (Figs. 1, 2) the proportions of species within each sample remain constant, and the ratio of one total sample size to the other is varied by the factors shown on the abscissa.

RESULTS

In all cases, the initial value (at $Y/X = 1$) of $PS_D = PS_I$. In a comparison of 2 identical samples, the initial value of $PS_D = PS_I = 1$. If the intrasample proportions x_i/X and y_i/Y are held constant as Y/X increases from 1, the resultant curve (PS_D starting at 1.0 in Figs. 1 and 2) is a limit curve composed of the maximum possible values of PS_D for $Y/X \geq 1.0$. The function describing this curve is $PS_{D_{max}} = 2X/X + Y$. It is evident (Figs. 1, 2) that the limit curve decreases with increasing Y/X, and that the decline is sharpest in the range $1 < Y/X < 6$. When 2 sample sizes vary by a factor of 3, for example, the maximum possible value of PS_D is 0.5, and when

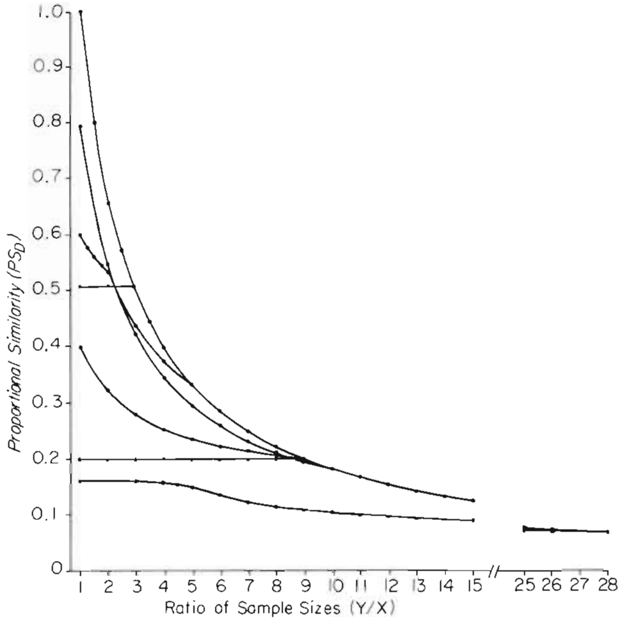

Fig. 1. Plots of data in Table 1 showing the effect of relative sample size on values of proportional similarity PS_D. The upper or limit curve, whose initial value is $PS_D = PS_I = 1.0$, is of the form $PS_{D_{max}} = 2X/X + Y$. The shapes of the other curves vary because, as the ratio of sample sizes increases, the number of $x_i > y_i$ increases in a manner depending on the distribution of individuals in X and Y. At $Y/X = 1$, $PS_D = PS_I$ in all curves. For all sample size ratios, PS_I is invariant, remaining at the initial value. The distributions have the same species in X and Y and were arbitrarily generated (Table 1) to give initial values ($Y/X = 1$) of 1.0, 0.8, 0.6, 0.5, 0.4, 0.2, and 0.16. Under these conditions all curves join the limit curve

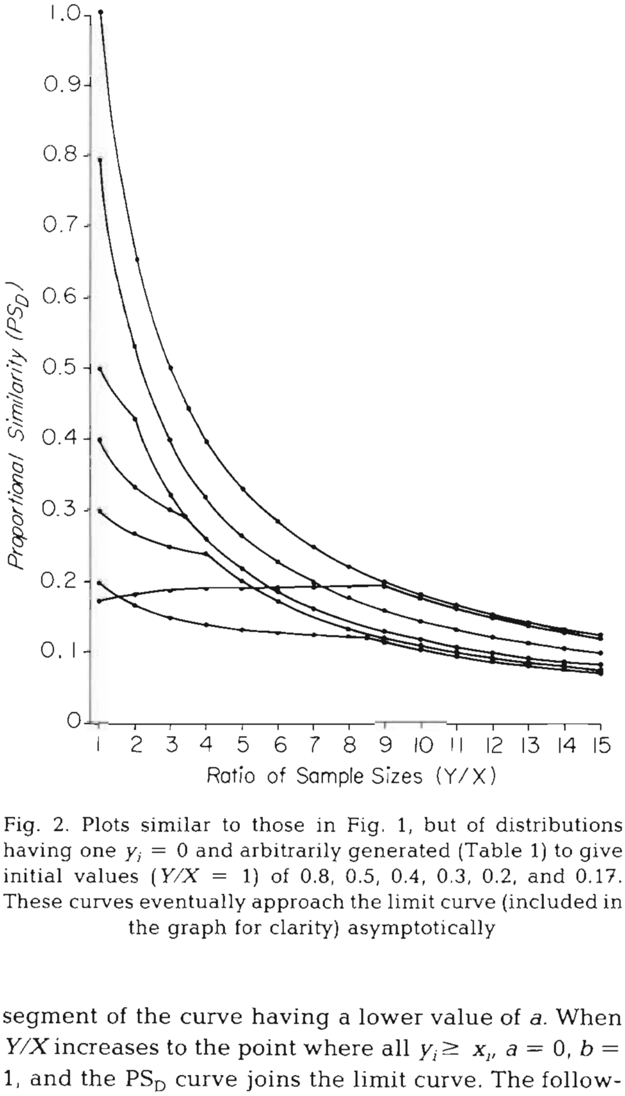

Fig. 2. Plots similar to those in Fig. 1, but of distributions having one $y_i = 0$ and arbitrarily generated (Table 1) to give initial values ($Y/X = 1$) of 0.8, 0.5, 0.4, 0.3, 0.2, and 0.17. These curves eventually approach the limit curve (included in the graph for clarity) asymptotically

they vary by a factor of 9, it is only 0.2. In comparisons of 2 samples with non-identical proportions of the same species (curves with initial values < 1.0), the higher the initial value of PS_D, the more steeply it is likely to decline with diverging sample sizes ($Y/X > 1$) (Fig. 1).

If both samples contain the same species, but in different proportions, the PS_D curves join the limit curve (Fig. 1). If both species composition and proportions differ in the 2 samples, the PS_D curves approach the limit curve asymptotically (Fig. 2). At lower Y/X values, both types of PS_D curves vary considerably. Their paths consist of one or more segments, each segment approaching the limit curve at a constant rate a, such that:

$$\frac{PS_D}{PS_{D_{max}}} = a\frac{Y}{X} + b \qquad (3)$$

For each segment, a is a function of the number of x_i, y_i, pairs in which $x_i > y_i$, as well as the difference $x_i - y_i$ in these pairs. As Y/X increases, the number of pairs having $x_i > y_i$ decreases; each decrease initiates a new segment of the curve having a lower value of a. When Y/X increases to the point where all $y_i \geq x_i$, $a = 0$, $b = 1$, and the PS_D curve joins the limit curve. The following 2 examples illustrate how these variables affect the values of PS_D:

(1) The distribution with initial $PS_D = 0.6$ contains 2 pairs with $x_i > y_i$ (x_4, y_4 and x_5, y_5) (Table 1). When $Y/X = 2$, $x_4 = y_4$ and only $x_5 > y_5$. This causes a transition point in the curve at $Y/X = 2$ (Fig. 1). When $Y/X = 5$, $x_5 = y_5$, causing a second transition point as the limit curve is joined, since all $y_i \geq x_i$.

(2) Even though 2 samples may be very similar in all other respects, one initial $x_i >> y_i$ (e.g. x_1, y_1 in the distribution with initial value 0.8 in Fig. 1) prevents the PS_D curve from joining the limit curve until the sample size ratio Y/X equals the initial value of that x_i/y_i (10 in this example: Fig. 1).

When one or more species present in X are absent from Y (Fig. 2), the curves do not join the limit curve, since at least one $y_i = 0$ and it can never exceed the corresponding $x_i > 0$. In the latter case, at that value of Y/X where $y_i \geq x_i$ in all pairs with $y > 0$, $a = 0$ and the curves begin a regular decline, each with its value a constant ratio ($b_{(\mathrm{final})} < 1$) of the corresponding value of the limit curve. This ratio depends only on the fraction of individuals in sample X belonging to species also represented in sample Y.

$$b_{(\mathrm{final})} = \frac{\Sigma\, x_i \text{ in all pairs with } y > 0}{X} \qquad (4)$$

As in Fig. 1, the courses of the left portions of the curves in Fig. 2 vary. For example, the curve with the lowest initial PS_D value (0.17) achieves the highest value at $Y/X = 7.3$.

DISCUSSION

One measure of the proportional similarity of 2 samples, PS_I, uses data initially standardized to remove any effect of different sample sizes. The other, PS_D, varies with differences in sample sizes of the 2 data sets being compared, or with differences in population densities if equal-sized quadrats are sampled. In the latter case, PS_D may reflect differences in total population density or resource abundance between two quadrats. Its use is justified only if the user intends such differences to contribute to the value of the index.

Several recent texts fail to mention PS_I, giving only PS_D (Southwood, 1976; Pielou, 1977) or its complement (Clifford and Stephenson, 1975), a dissimilarity measure. The use of PS_D is also recommended by Gallaher and Blake (1977), who presented values of both measures for one data set with sample size ratio $Y/X = 2.7$. This gave the results $PS_I = 0.850$ and $PS_D = 0.513$. Gallaher and Blake (1977: 264) stated: 'These results suggest that the relative index PS_I is dominated by compositional similarity; whereas, the absolute index PS_D is weighted by distributional or geometrical differences.' However, our analysis (Fig. 1) shows that while their value of PS_I is 85 % of the maximum possible similarity of the 2 samples, their value of PS_D is actually 95 % of maximum possible similarity ($PS_{D_{max}} = 0.54$ at $Y/X = 2.7$).

We have demonstrated that while PS_I is independent of sample size differences, PS_D (1) generally decreases with more disparate sample sizes, (2) rapidly approaches a limit curve with more disparate sample sizes, and (3) is often a measure more of the size ratio of the 2 samples than their proportionate compositional similarity; in the extreme case it is a function of sample size ratio only. Moreover, because of the disproportionate effects of certain x_i, y_i pairs, curves of PS_D plotted against sample size ratio may vary widely in shape (Table 1, Figs. 1, 2). In some cases (e. g. curves with initial values 0.8, 0.6, and 0.5 in Fig. 1), samples with disparate PS_D values at one ratio of sample sizes ($Y/X = 1$ in this case) may all have the same value at a different sample size ratio ($Y/X = 2.2$); at values of $Y/X > 2.2$, their order is reversed and remains so until they superimpose on the limit curve. Figs. 1 and 2 show that these undesirable features of PS_D may prevail at sample size ratios between 1 and 3, a realistic range for data in ecological analyses (e. g. Huey et al., 1974; Gallaher and Blake, 1977). Some studies (e. g. Wiegert, 1974; Hurlbert, 1978) have included more disparate sample sizes.

We recommend against use of PS_D because we consider the properties listed above undesirable. PS_I is completely independent of sample size, decreasing only with more disparate species frequencies. It should be used when one wishes to exclude the effects of differing sample sizes. It also seems generally more appropriate to use PS_I as a measure of similarity of resource use, where one is usually interested in the mean similarity of individuals of the different species without regard to sample size.

Acknowledgements. This work was supported by NSF grant DEB 77-24430. We thank R. B. Huey, E. C. Pielou, T. W. Schoener and the late R. H. Whittaker for reading the manuscript and for helpful comments.

LITERATURE CITED

Abrams, P. (1980). Some comments on measuring niche overlap. Ecology 61: 44–49

Bernstein, B. B., Hessler, R. R., Smith, R., Jumars, P. A. (1978). Spatial dispersion of benthic Foraminifera in the abyssal central North Pacific. Limnol. Oceanogr. 23: 401–416

Bloom, S. A. (1981). Similarity indices in community studies: potential pitfalls. Mar. Ecol. Prog. Ser. 5: 125–128

Boesch, D. F. (1977). A new look at the zonation of benthos along the estuarine gradient. In: Coull, B. C. (ed.) Ecology of marine benthos. University of South Carolina Press, Columbia, S. C., pp. 245–266

Clifford, H. T., Stephenson, W. (1975). An introduction to numerical classification, Academic Press, New York

Dauer, D. M., Simon, J. L. (1975). Lateral or along-shore distribution of the polychaetous annelids of an intertidal, sandy habitat. Mar. Biol. 31: 363–370

Davies, R. W., Wrona, F. J., Linton, L. (1979). A serological study of prey selection by *Helobdella stagnalis* (Hirudinoidea). J. Anim. Ecol. 48: 181–194

Dean, T A. (1981). Structural aspects of sessile invertebrates as organizing forces in an estuarine fouling community. J. exp. mar. biol. Ecol. 53: 163–180

Gallaher, E. E., Blake, N. J. (1977). On equivalent forms of Whittaker's similarity index. J. theor. Biol. 68 : 259–265

Hanski, I. (1978). Some comments on the measurement of niche metrics. Ecology 59: 168–174

Huey, R. B., Pianka, E. R., Egan, M. E., Coons, L. W. (1974). Ecological shifts is sympatry. Ecology 55: 304–316

Hurlbert, S. H. (1978). The measurement of niche overlap and some relations. Ecology 59: 69–77

Leviten, P. J. (1978). Resource partitioning by predatory gastropods of the genus *Conus* on subtidal Indo-Pacific coral reefs: the significance of prey size. Ecology 59: 614–631

Leviten, P. J., Kohn, A. J. (1980). Microhabitat resource use, activity patterns, ans episodic catastrophe: *Conus* on tropical intertidal reef rock benches. Ecol. Monogr. 50: 55–75

Pielou, E. C. (1975). Ecological diversity, John Wiley & Sons, New York

Pielou, E. C. (1977). Mathematical ecology, Wiley, New York

Pielou, E. C. (1979). Interpretation of paleoecological similarity matrices. Paleobiology 5: 435–443

Price, P. W., Willson, M. F. (1976). Some consequences for a parasitic herbivore, the milkweed longhorn beetle, *Tetraopes tetrophthalmus*, of a host-plant shift from *Asclepias syriaca* to *A. verticillata*. Oecologia 25: 331–340

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

Rex, M. A. (1977). Zonation in deep-sea gastropods: the importance of biological interactions to rates of zonation. In: Keegan, B. F., Ceidigh, P. O., Boaden, P. J. S. (eds.) Biology of benthic organisms. Pergamon Press, New York, pp. 521–530

Sabo, S. R., Whittaker, R. H. (1979). Bird niches in a subalpine forest: an indirect ordination. Proc. natn. Acad. Sci. U.S.A. 76: 1338–1342

Sale, P. F. (1974). Overlap in resource use, and interspecific competition. Oecologia 17: 245–256

Schoener, A., Greene, C. H. (1981). Comparison between destructive and nondestructive sampling of sessible epibenthic organisms. Limnol. Oceanogr 26: 770–774

Southwood, T. R. E. (1976). Ecological methods, Chapman and Hall, London

Weinstein, M. P., Weiss, S. L., Walters, M. F. (1980). Multiple determinations of community structure in shallow marsh habitats, Cape Fear river estuary, North Carolina, USA. Mar. Biol. 58: 227–243

Whittaker, R. H. (1952). A study of summer foliage insect communities in the Great Smoky Mountains. Ecol. Monogr. 22: 1–44

Wiegert, R. G. (1974). Litterbag studies of microarthropod populations in three South Carolina old fields. Ecology 55: 94–102

This paper was submitted to the editor; it was accepted for printing on December 8, 1981