Environmental and Ecological Statistics **1**, 95–107 (1994)

Measuring biological diversity

ANDREW R. SOLOW

Marine Policy Center, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

STEPHEN POLASKY

Department of Agricultural and Resource Economics, Oregon State University, Corvallis, OR 97331-3601, USA

Received September 1993

The diversity of a set of species refers to the joint dissimilarity of the species in the set. This paper discusses the measurement of diversity from the set of pairwise distances between the species in the set. A measure called the effective number of species is developed from a non-parametric probability inequality and is shown to have a simple interpretation in terms of comparing linear experiments.

Keywords: comparison of linear experiments; diversity; effective number of species; Gallot's inequality

1. Introduction

There is widespread concern that human activities – most notably tropical deforestation – are contributing to a large-scale reduction in biological diversity through the extinction of plant and animal species (e.g. Wilson, 1992). While there is general agreement that the conservation of diversity is important, it is also important to recognize that the conservation of diversity must compete for attention and resources with other worthwhile social, environmental, and economic goals. As long as the resources for the conservation of diversity are scarce, they must be allocated across conservation projects. To allocate these resources for maximal impact on diversity, it is necessary to move beyond generalities to a more precise definition of diversity and, specifically, to measure diversity. This problem is addressed in a small but growing literature (e.g. Vane-Wright *et al.*, 1991; Eiswerth and Haney, 1992; Weitzman, 1992; Solow *et al.*, 1993). In this paper, we discuss some of the issues in measuring diversity and we propose a new measure.

For the purposes of this paper, the diversity of a set of species can be described as the joint dissimilarity of the species in the set. The information available for constructing a measure of diversity is the set of pairwise distances between the species in the set. These distances can be based on morphological or behavioural differences, or on more refined (although not necessarily more informative) molecular biological methods. This use of the term diversity differs from that in ecology (e.g. Pielou, 1975), where diversity is a property of the relative abundances of species without regard to the differences between them.

Before proceeding, a word is in order about the interest that this problem may hold for statisticians. First, in a broad sense, the problem of measuring diversity can be viewed as characterizing an aspect of the distribution of points in space. It is, therefore, related to standard problems in multivariate analysis, although the aspect of interest – namely, diversity – is somewhat non-standard. Second, one of the approaches described in this paper constructs a diversity measure from a non-parametric probability inequality. Interestingly, this approach leads to a measure that has a straightforward interpretation in the context of comparing linear experiments.

2. A decision-making framework

Making effective conservation decisions depends on factors other than the measurement of diversity. Before turning to the measurement problem, it is useful to outline a simple decision-making framework in which a specific measure can be embedded. The reason for this digression is that a key aspect of species conservation – interactions between species – is most conveniently treated outside the measurement question and, for completeness, some indication of how it can be handled is in order.

Suppose that the total set of species under consideration is T. In principle, T could consist of either a set of target species or all the species on Earth. Extinctions partition T into a set of extinct species X and a set of surviving species Y. The pattern of extinctions is uncertain and can be characterized by a probability distribution. The aim of a conservation strategy C is to influence this distribution. Without further specification, let the diversity of a set of species S be $D(S)$. A reasonable basis for evaluating C is the expected diversity of the surviving species:

$$E_C(D(Y)) = \sum D(y)p_C(y) \tag{1}$$

where the summation extends over all subsets y of T and $p_C(y)$ is the probability under C that $Y = y$.

Many important ecological interactions are subsumed in $p_C(y)$. In particular, a myopic strategy that seeks to conserve a set of highly diverse species without also conserving the species on which they depend will have low expected diversity. To ensure that mistakes of this kind are not made when species interactions are poorly understood, the best instrument of species conservation may be the conservation of the habitat in which the species live.

3. The measurement of pure diversity

To implement the decision-making framework outlined in the previous section, it is necessary to specify the diversity measure $D(S)$. One measure of the diversity of a set of species is the number of species in the set. One problem with this measure – which is called *species richness* – is that it does not take account of differences between species. For example, a set consisting of four species of ant is in some sense less diverse than a set consisting of one species of ant, one species of elephant, and one species of fern.

Let the distance between two species s_i and s_j be d_{ij}. It is natural to equate the diversity of a set consisting of s_i and s_j to an increasing function of d_{ij}. For example, Fig. 1 shows two sets, each consisting of two species represented as points. Because the distance between the species in set S_1 is less than the distance between the species in set S_2, S_2 is more diverse than S_1. One way to think about the measurement of diversity is as an extension of the notion of distance to more than two points.

To facilitate the discussion, it is helpful, as in Fig. 1, to have a graphical representation of $n > 2$ species that preserves pairwise distances. If the pairwise distances are metric, then n species can be represented by points in Euclidean space of dimension $\leqslant n - 1$. Even if the distances are not metric, the species can be approximately represented by points in Euclidean space via non-metric scaling (e.g. Kruskal, 1964). If the pairwise distances satisfy the stronger ultrametric condition, then the species can be represented as the terminal nodes of a rooted tree. Even if the distances are not ultrametric, the species can be approximately represented in a tree (e.g. Sneath and Sokal, 1973).

Constructing a sensible measure of diversity is not as easy as it may seem. As before, let

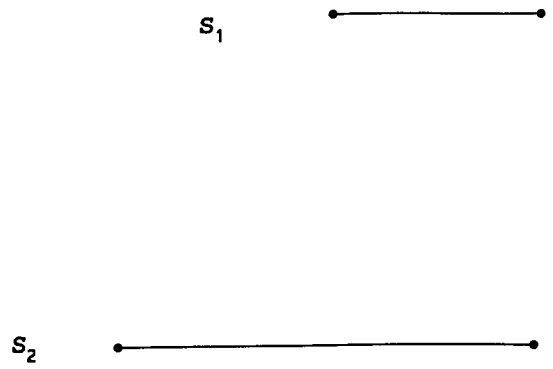

Figure 1. The pair of species in S_1 is less diverse than the pair of species in S_2.

$T = (s_1, s_2, \ldots, s_N)$ be the total set of species and let d_i be the average distance between species s_i and all other species in T. Eiswerth and Haney (1992) suggested that the diversity of a subset S of T be measured by

$$EH(S) = \sum_{s_i \in S} d_i. \tag{2}$$

To see why this measure will not work, consider the four species represented in Fig. 2. In this case, the species are represented in a symmetric, unrooted tree, with the distance between two species given by the sum of the lengths of the branches connecting them. For this configuration,

$$d_1 = d_2 = 2(2a + b + c)/3,$$
$$d_3 = d_4 = 2(a + b + 2c)/3.$$

Under this measure, the diversity of the pair (s_1, s_2) is greater than the diversity of the pair (s_1, s_3), although $d_{12} < d_{13}$. The fundamental problem with this measure is that, in calculating d_i, no distinction is made between species that will be lost and those that will survive.

To narrow the search for sensible measures of diversity, it is useful to set out some requirements for such measures. Three natural requirements are the following. First, diversity should not be decreased by the addition of a species. That is, if $S \subset S'$, then $D(S) \leqslant D(S')$. This is called *monotonicity in species*. Second, diversity should not be increased by the addition of a species that is identical to a species already in the set. This means that, for metric distances, $D(S \cup s_0) = D(S)$ if and only if $d_{0i} = 0$ for some $s_i \in S$. Weitzman (1992) referred to this as *twinning*. Third, diversity should not be decreased by an unambiguous increase in the distances between species. Specifically,

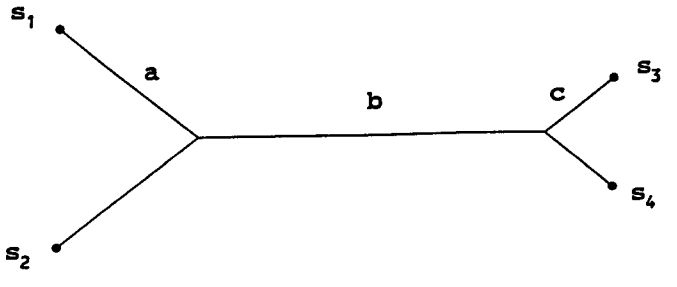

Figure 2. Under $EH(S)$, the pair (s_1, s_2) is more diverse than the pair (s_1, s_3), even though $d_{12} < d_{13}$.

for a one-to-one mapping of S onto S' such that $d_{ij} \leqslant d_{i'j'}$, with at least one strict inequality, $D(S) \leqslant D(S')$. This is called *monotonicity in distance*.

The first measure to satisfy these requirements was proposed by Weitzman (1992). To begin with, Weitzman assumed that the diversity of a single species is 0 and defined the distance between a single species s_0 and a set of species S as the nearest-neighbour distance,

$$d(s_0, S) = \min_{s_i \in S} d_{0i}. \tag{3}$$

With this definition, the measure is given by

$$W(S) = \max_{s_i \in S}[W(S - s_i) + d(s_i, S - s_i)], \tag{4}$$

where $S - s_i$ is the set formed by omitting s_i from S.

A heuristic motivation for this measure is the following. It would seem natural to require that $D(S \cup s_0) = D(S) + d(s_0, S)$. Moreover, this would provide an algorithm for calculating $D(S)$: starting with any species in S, $D(S)$ could be calculated by adding the remaining species one at a time and incrementing diversity by the nearest-neighbour distance. Unfortunately, the results of this calculation depend on the order in which the species are considered. The maximization in (4) removes this ambiguity.

One attractive feature of this measure is that, in the case where the species can be represented exactly in a tree, it corresponds to the length of the tree. This seems natural and convenient. One drawback of this measure is that, outside the ultrametric case, it is not strictly monotone in distance, in the sense that it need not increase with an unambiguous increase in distances. For example, in the case of three species, Weitzman's measure corresponds to the sum of the maximum and the minimum of the three pairwise distances. It is, therefore, unaffected by changes in the intermediate pairwise distance.

4. A utilitarian approach

The discussion so far has essentially assumed that diversity is desirable and has focused on constructing a measure with reasonable properties. A different view is that it is not so much diversity *per se* that is valuable, but the benefits that diversity provides. For example, one justification for species conservation is that some species may provide a future medical benefit. In this section, we explore the implications of this argument for the measurement of diversity.

Suppose that interest in conservation arises from the possibility that species will provide a specific benefit in the future. The essential property of this benefit is that having more than one species that provide it is no better than having a single species that provides it. An example of such a benefit is a cure for a disease and, for concreteness, we will use this as a metaphor.

Consider a set of species $S = (s_1, s_2, \ldots, s_n)$ and let B_i be the event that s_i is a cure. The event that S contains a cure is

$$B(S) = \bigcup_{i=1}^{n} B_i.$$

Because the expected benefit for S is the product of the value of a cure and the probability $p(S) = \Pr(B(S))$, $p(S)$ provides a basis for comparing S to other sets of species.

In the absence of specific information, it is reasonable to assume that $\Pr(B_i) = p$, $i = 1, 2, \ldots, n$, with p unknown. We will make the further assumption that

$$\Pr(B_i \mid B_j) = p + (1 - p)f(d_{ij}), \tag{5}$$

where f is a known function satisfying the following conditions;

$$f(0) = 1, \qquad f(\infty) = 0, \qquad f' \leqslant 0.$$

Under this model, the conditional probability of B_i given B_j declines from 1 to p as d_{ij} increases from 0 to ∞. It follows from (5) that

$$\Pr(B_i \cap B_j) = p^2 + p(1-p)f(d_{ij}). \tag{6}$$

For consistency, assume that the distances are metric and that the function f is positive definite. An example of such a function is $f(d) = \exp(-\theta d)$, $\theta > 0$. It may be helpful to think of $f(d)$ in the following way. Consider the binary random variable $I_i = 1$ if B_i and 0 otherwise, $i = 1, 2, \ldots, n$. Then $f(d_{ij})$ is the correlation between I_i and I_j.

In general, it is not possible to find $p(S)$ from the univariate and bivariate marginal probabilities. It is, however, possible to place a lower bound on $p(S)$. For arbitrary events A_i, $i = 1, 2, \ldots, n$, Gallot (1966) showed that

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \geqslant \sup_c (c'P_1P_1'c)/(c'P_2c),$$

where c is an arbitrary n-vector, $P_1 = (\Pr(A_1) \ldots \Pr(A_n))'$ and $P_2 = [\Pr(A_i \cap A_j)]$, $i, j = 1, 2, \ldots, n$. In terms of the model outlined above, this implies that

$$p(S) \geqslant \sup_c (1 + p(1-p)(c'Fc)/(c'Pc))^{-1}$$

where $F = [f(d_{ij})]$, $i, j = 1, 2, \ldots, n$, and P is an $n \times n$ matrix with all elements equal to p^2. In general, the elements of F depend on S, although this is suppressed in the notation. Provided that F is non-singular, it can be shown that

$$\begin{aligned} \sup_c (c'Pc)/(c'Fc) &= p^2 e'F^{-1}e \\ &= p^2 V(S) \end{aligned} \tag{7}$$

where e is an n-vector of 1's (e.g. Gantmacher, 1959). A similar result involving a generalized inverse of F follows from Kounias (1968) in the case where F is singular. Since the lower bound on $p(S)$ is an increasing function of $V(S)$, different sets of species can be compared in terms of this measure, with larger values corresponding to greater lower bounds.

The measure $V(S)$ has some appealing properties. If $f(d_{ij}) = 0$ for all $i \neq j$ (i.e. the species are unrelated), then F is the identity matrix and $V(S)$ is equal to n (i.e. the number of species in S). If $f(d_{ij})$ approaches 1 for all i, j (i.e. the species are perfectly related), then $V(S)$ also approaches 1. As discussed below, while it is possible for $V(S)$ to exceed n, it is conjectured for a reasonably constrained family of functions f that $V(S)$ lies between 1 and n. In a sense, $V(S)$ can be interpreted as the effective number of species in S.

Some intuition about $V(S)$ can be gained by exploiting a connection to the comparison of linear experiments (e.g. Hansen and Torgerson, 1974). Consider a set of observations:

$$Y_i = \mu + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

where μ is the unknown mean and ε_i, $i = 1, 2, \ldots, n$, are zero mean errors with covariance matrix F. It may be helpful to think of the observations as being taken at locations in space, with F reflecting spatial covariance in the error process. If only a subset of the observations are to be retained to estimate μ, a natural criterion for comparing different subsets is the variance of the generalized least squares estimator $(e'F^{-1}e)^{-1}$, which is the reciprocal of $V(S)$. We will return to this connection below.

We now take up the question: does $V(S)$ satisfy the three requirements for a diversity measure? Let $S = (s_1, s_2, \ldots, s_n)$ with

$$V(S) = \sup_{c_n} (c_n' P_n c_n)/(c_n' F_s c_n)$$

in obvious notation. Let $S' = S \cup s_{n+1}$ and let $c_{n+1}^* = (c_n' 0)'$. Since

$$\begin{aligned} V(S) &= \sup_{c_n} (c_{n+1}^{*\prime} P_{n+1} c_{n+1}^*)/(c_{n+1}^{*\prime} F_{s'} c_{n+1}^*) \\ &\leqslant \sup_{c_{n+1}} (c_{n+1}' P_{n+1} c_{n+1})/(c_{n+1}' F_{s'} c_{n+1}) \\ &= V(S'), \end{aligned}$$

$V(S')$ cannot be smaller than $V(S)$. This ensures that $V(S)$ is monotone in species.

Let S and S' be as above with $d_{1,n+1} = 0$. This means that the first and last rows and columns of $F_{s'}$ are identical. For any $(n+1)$-vector c_{n+1},

$$\begin{aligned} &(c_{n+1}' P_{n+1} c_{n+1})/(c_{n+1}' F_{s'} c_{n+1}) \\ &\quad = (c_n' P_n c_n)/(c_n' F_s c_n) \end{aligned} \tag{8}$$

where the first element of c_n is equal to the sum of the first and last elements of c_{n+1} and all other elements of c_n are equal to the corresponding elements of c_{n+1}. It follows that $V(S')$, which is the supremum over c_{n+1} of the left-hand side of (8), is equal to $V(S)$, which is the supremum over c_n of the right-hand side of (8). This ensures that $V(S)$ satisfies twinning.

In general, $V(S)$ is not monotone in distance. That is, it is possible to construct a positive definite matrix F with positive elements such that $e'F^{-1}e$ does not increase with a decrease in one of the off-diagonal elements of F. This possibility is discussed in Eaton (1992) in the context of comparing linear experiments. In terms of the situation outlined above, this means that it is possible to reduce the variance of the estimator of μ by increasing the correlation between two observations (leaving the other correlations fixed). Briefly, this seemingly paradoxical result arises from the possibility that μ can be estimated without error if F becomes singular in a certain way. When F is nearly singular in this way, an increase in correlation that moves F closer to singularity in this way can also reduce variance.

It is possible to show in the 3×3 case that a sufficient condition for monotonicity in distance is that $f_{ij} \geqslant f_{ik} f_{jk}$ for all i, j, k (i.e. all partial correlations are non-negative). Since the triangle inequality ensures that this condition is met for exponential f, we conjecture that $V(S)$ is both monotone in distance and lies between 1 and n for this choice of f.

The main disadvantage of $V(S)$ is that it assumes knowledge of the function f. In some cases, there may be sufficient information to approximate this function reasonably well. In other cases, it may be best to assume that $f(d)$ has a simple parametric form and to view $V(S)$ as a family of measures indexed by the parameter. Alternatively, a single measure can be found by integrating over a specified prior distribution for this parameter. It is also possible, as noted below, to place rough bounds on $V(S)$ by assuming that $f(d_{ij})$ is either 0 or 1.

5. An example

In this section, a simple example of the application of the measures discussed above is presented. The data used in this illustration were taken from Rodman (1991). They consist of pairwise distances between 26 species of plants that produce glucosinolate (sulfur-containing compounds

Table 1. Taxa of glucosinolate-producing plants and putative relatives

Taxon	*Code*	*Taxon*	*Code*
Akaniaceae	AKA	Tovariaceae	TOV
Bataceae	BAT	Tropaeolaceae	TRO
Brassicaceae	BRA	Balsaminacea	BAL
Bretschneideraceae	BRE	Celastraceae	CEL
Capparaceae	CAP	Centrospermae	CEN
Cariacaceae	CAR	Dilleniaceae	DIL
Drypetes	DRY	Euphorbiaceae	EUP
Gyrostemonaceae	GYR	Flacouticeae	FLA
Limnanthaceae	LIM	Geraniaceae	GER
Moringaceae	MOR	Koeberliniaceae	KOE
Pentadiplandraceae	PEN	Oxalidaceae	OXA
Resedaceae	RES	Passifloraceae	PAS
Salvadoraceae	SAL	Sapindaceae	SAP

related to mustard oils that have been identified as potential cancer-fighting agents). The species are listed in Table 1. The distances were based on an analysis of 96 characteristics. The species are displayed graphically in Fig. 3. To construct this figure, Rodman (1991) applied principal coordinate analysis (e.g. Gower, 1966) to the matrix of pairwise distances and plotted the species along the first two principal axes.

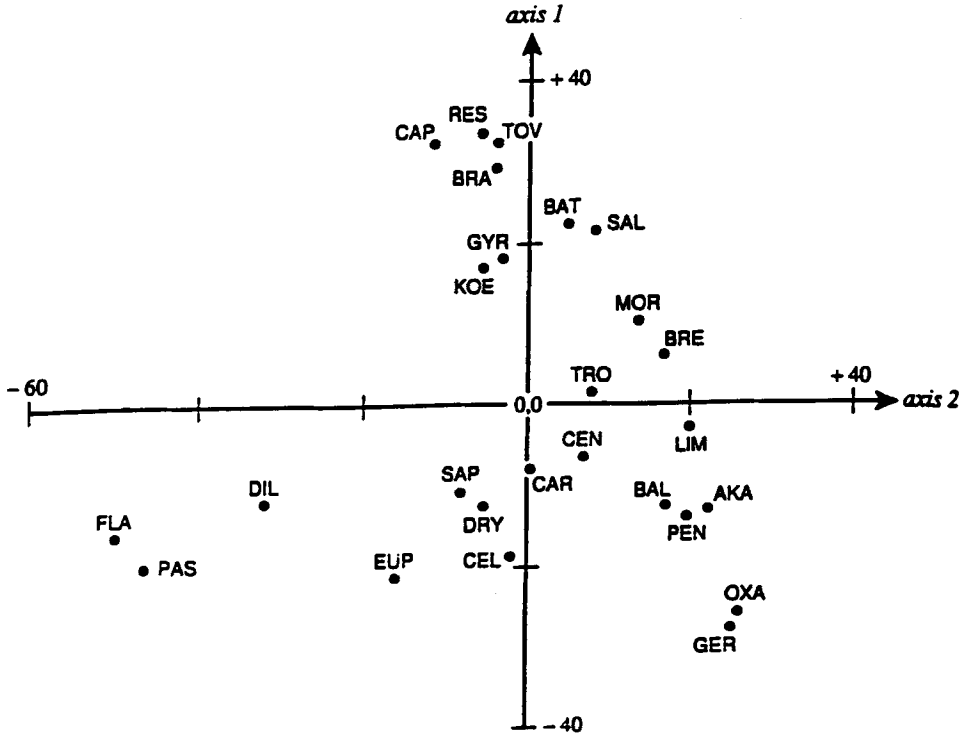

Figure 3. Locations of 26 species of glucosinolate-producing plants along the first two principal axes (Rodman, 1991).

Table 2. Values of $W(S)$ and $V(S)$ for S_1 = all 26 species; S_2 = (FLA, RES, GER), S_3 = (CAP, RES, TOV, BRA, KOE, GYR, BAT, SAL, MOR).

		$V(S)$	
S	$W(S)$	$\theta = 0.1$	$\theta = 0.5$
S_1	106.73	2.62	10.55
S_2	38.82	2.32	3.00
S_3	25.81	1.50	3.66

In Table 2, the values of $W(S)$ and $V(S)$ are given for three sets of species: S_1 = all 26 species, S_2 = (FLA, RES, GER), and S_3 = (CAP, RES, TOV, BRA, KOE, GYR, BAT, SAL, MOR). In qualitative terms, S_2 is small, but dissimilar, while S_3 is large, but similar. In calculating $V(S)$, we assumed that $f(d) = \exp(-\theta d)$ and two choices of θ were considered: $\theta = 0.1$ and 0.5. Loosely speaking, the effect of increasing θ on $V(S)$ is to give more weight to the number of species and less to their dissimilarity. The measure $W(S)$ attaches no weight to the number of species except through its effect on the accumulation of inter-species distances. In this sense, $W(S)$ is similar to $V(S)$ with $\theta = 0.1$ (e.g. they give the same ranking of S_2 and S_3). In contrast, the ranking is different with $\theta = 0.5$, since $V(S)$ is more strongly influenced by the number of species. In fact, for $\theta = 0.5$, the species in S_2 are effectively independent, so that their effective number is equal to their actual number.

6. Discussion

The diversity measures discussed in this paper are clearly best suited for situations in which there is extensive information about the species of interest. Even in such special situations, questions remain. For example, in addition to establishing the conditions under which $V(S)$ is monotone in distance, it would be useful to have some idea of the tightness of Gallot's inequality.

In practice, situations in which distances have been measured for all species pairs are exceptional. For example, many conservation decisions concern large habitats containing large numbers of species from different groups and certain pairwise distances are unavailable or unreliable. It is still possible, in such cases, to place bounds on $V(S)$. For example, if distance data are available within genera but not between genera, then an upper bound for $V(S)$ is the sum of the effective number of species in the genera (i.e. corresponding to the case where $f(d_{ij}) = 0$ for s_i and s_j in different genera).

It should be clear that the problem of measuring diversity remains very much open. The main contribution of the work outlined in this paper may lie in its formalization of this problem. The proposed measures are clearly not satisfactory in all respects. Further effort is needed to understand the behaviour of these measures and to develop improved measures.

References

Eaton, M.L. (1992). A group action on covariances with applications to the comparison of linear normal experiments. Unpublished mimeo, Department of Statistics, University of Minnesota.

Eiswerth, M.E. and Haney, J.C. (1992). Allocating conservation expenditures across habitats: Accounting for inter-species genetic distinctiveness. *Ecological Economics*, **5**, 235–250.

Gallot, S. (1966). A bound for the maximum of a number of random variables. *Journal of Applied Probability*, **3**, 556–558.

Gantmacher, F.R. (1959). *The Theory of Matrices*. (Chelsea, New York).

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **29**, 115–129.

Hansen, O.H. and Torgersen, E.N. (1974). Comparison of linear normal experiments. *Annals of Statistics*, **2**, 367–373.

Kounias, E.G. (1968). Bounds for the probability of a union, with applications. *Annals of Mathematical Statistics*, **39**, 2154–2158.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.

Pielou, E.C. (1975). *Ecological Diversity*. (Wiley Interscience, New York).

Rodman, J.E. (1991). A taxonomic analysis of glucosinolate-producing plants, Part I: Phenetics. *Systematic Botany*, **16**, 598–618.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy*. (Freeman, San Francisco).

Solow, A.R., Polasky, S., and Broadus, J.M. (1993). On the measurement of biological diversity. *Journal of Environmental Economics and Management*, **24**, 60–68.

Vane-Wright, R.I., Humphries, C.J., and Williams, P.H. (1991). What to protect – Systematics and the agony of choice. *Biological Conservation*, **55**, 235–254.

Weitzman, M.L. (1992). On diversity. *Quarterly Journal of Economics*, **107**, 363–405.

Wilson, E.O. (1992). *The Diversity of Life*. (Harvard University Press, Cambridge).