Journal of Applied Ecology 1998, **35**, 523–531

A taxonomic distinctness index and its statistical properties

K.R. CLARKE and R.M. WARWICK

Centre for Coastal and Marine Sciences, Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK

Summary

1. For biological community data (species-by-sample abundance matrices), Warwick & Clarke (1995) defined two biodiversity indices, capturing the structure not only of the distribution of abundances amongst species but also the taxonomic relatedness of the species in each sample. The first index, taxonomic diversity (Δ), can be thought of as the average taxonomic 'distance' between any two organisms, chosen at random from the sample: this distance can be visualized simply as the length of the path connecting these two organisms, traced through (say) a Linnean or phylogenetic classification of the full set of species involved. The second index, taxonomic distinctness (Δ^*), is the average path length between any two randomly chosen individuals, conditional on them being from different species. This is equivalent to dividing taxonomic diversity, Δ, by the value it would take were there to be no taxonomic hierarchy (all species belonging to the same genus). Δ^* can therefore be seen as a measure of pure taxonomic relatedness, whereas Δ mixes taxonomic relatedness with the evenness properties of the abundance distribution.

2. This paper explores the statistical sampling properties of Δ and Δ^*. Taxonomic diversity is seen to be a natural extension of a form of Simpson's index, incorporating taxonomic (or phylogenetic) information. Importantly for practical comparisons, both Δ and Δ^* are shown not to be dependent, on average, on the degree of sampling effort involved in the data collection; this is in sharp contrast with those diversity measures that are strongly influenced by the number of observed species.

3. The special case where the data consist only of presence/absence information is dealt with in detail: Δ and Δ^* converge to the same statistic (Δ^+), which is now defined as the average taxonomic path length between any two randomly chosen species. Its lack of dependence, in mean value, on sampling effort implies that Δ^+ can be compared across studies with differing and uncontrolled degrees of sampling effort (subject to assumptions concerning comparable taxonomic accuracy). This may be of particular significance for historic (diffusely collected) species lists from different localities or regions, which at first sight may seem unamenable to valid diversity comparison of any sort.

4. Furthermore, a randomization test is possible, to detect a difference in the taxonomic distinctness, for any observed set of species, from the 'expected' Δ^+ value derived from a master species list for the relevant group of organisms. The exact randomization procedure requires heavy computation, and an approximation is developed, by deriving an appropriate variance formula. This leads to a 'confidence funnel' against which distinctness values for any specific area, pollution condition, habitat type, etc., can be checked, and formally addresses the question of whether a putatively impacted locality has a 'lower than expected' taxonomic spread. The procedure is illustrated for the UK species list of free-living marine nematodes and sets of samples from intertidal sites in two localities, the Exe estuary and the Firth of Clyde.

Key-words: biodiversity, randomization test, sampling effort, unbiasedness, variance estimate.

Journal of Applied Ecology (1998) **35**, 523–531

Correspondence: Dr K. R. Clarke (fax: 01752 633101; e-mail: b.clarke@pml.ac.uk).

Introduction

It is increasingly recognized (e.g. Harper & Hawksworth 1994) that adequate measures of biodiversity within a particular taxonomic group should not be merely functions of the number of species present and their relative abundances, but should also include information on the 'relatedness' of these species. There is now a substantial literature (Faith 1994; Humphries, Williams & Vane-Wright 1995 and referrals therein) on measures incorporating, principally, phylogenetic relationships amongst species and their possible use in selecting species or reserves of greatest conservation priority. Vane-Wright, Humphries & Williams (1991), Williams, Humphries & Vane-Wright (1991) and May (1990) introduced measures of distinctness based only on the topology of a phylogenetic tree, appropriate when branch lengths are entirely unknown, and Faith (1992, 1994) defined and justified a phylogenetic diversity (PD) measure based on known branch lengths: PD is simply the cumulative branch length of the full tree.

This literature does not appear, to date, to have carried over into the area of environmental monitoring and assessment, where the emphasis is not on choosing species to conserve but monitoring for environmental degradation or the benefits of remediation. The considerations here are rather different: the raw material is often a set of community samples with recorded abundances for each observed species, rather than a single species list, thought of as a complete inventory. The outcome required is not a preferential selection of species from the inventory for conservation status, but an assessment of whether sampled assemblages display some pattern in biodiversity through time or in space. Natural variation, and thus sampling properties of the resulting abundance matrices and derived indices, are of paramount importance. Also, the basic information on species relatedness is often just a Linnean taxonomy (Fig. 1), a crude approximation to a phylogeny but one that does impose an ordering of branch lengths which is interpretable and should be used. For example, even allowing that the only data available are in the form of presence/absences, measures that rely solely on topology and/or species richness would not distinguish between Fig. 2a and Fig. 2b, yet Fig. 2a clearly exhibits greater biodiversity in the sense of richness in higher taxa. Similarly, PD, applied to a Linnean classification (Faith 1994), has a focus exclusively on 'character richness' rather than 'character combinations' (in the terminology of Humphries, Williams & Vane-Wright 1995), so that PD concentrates on higher taxon richness and ignores the evenness component in diversity. Thus, PD would not distinguish between Fig. 2c and Fig. 2d, yet Fig. 2d clearly represents a less taxonomically diverse assemblage than Fig. 2c, both in the sense of possessing greater vulnerability to species loss and in potential functional inefficiency.

An over-riding consideration in a comparative biodiversity study is the extent to which a putative statistic is sensitive to differing sampling effort at different sites or times. It is well-known, and demonstrated starkly in Fig. 3a–c, that standard diversity estimates can be strongly dependent on sampling effort, particularly in so far as they are influenced by the number of species in the sample. Species richness is crucially dependent on sampling effort and it must be expected that only carefully controlled and equitable sampling studies can provide comparative data. Warwick & Clarke (1995), however, define taxonomic diversity/distinctness measures that satisfy the above requirements of incorporating higher taxon richness and evenness concepts (see the Δ^+ values in the legend to Fig. 2) but also have an apparent insensitivity to sampling effort (Fig. 3d–f).

In the Methods that follow, the construction of Δ and Δ^* is described and the link to the Simpson diversity drawn. It is demonstrated theoretically that, if Δ_m and Δ_m^* are defined as the values of Δ and Δ^* from a subset of m organisms, randomly selected from a total of n individuals, then they are either exactly (Δ_m) or

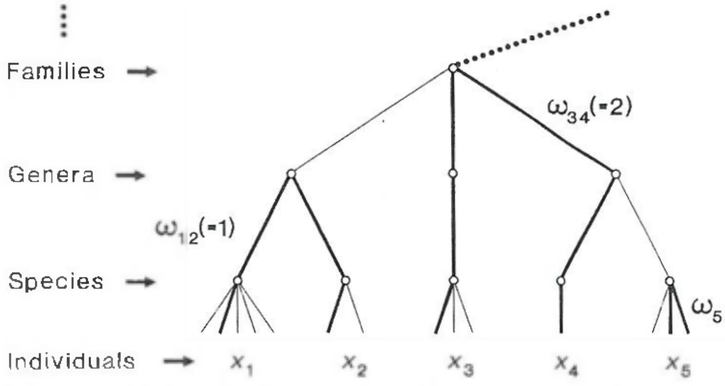

Fig. 1. Part of a taxonomic classification, showing examples of path length weights $\{\omega_{ij}\}$ used to define taxonomic diversity/distinctness measures; conventional diversity indices utilize only the species abundances $\{x_i; i = 1, \ldots, s\}$.

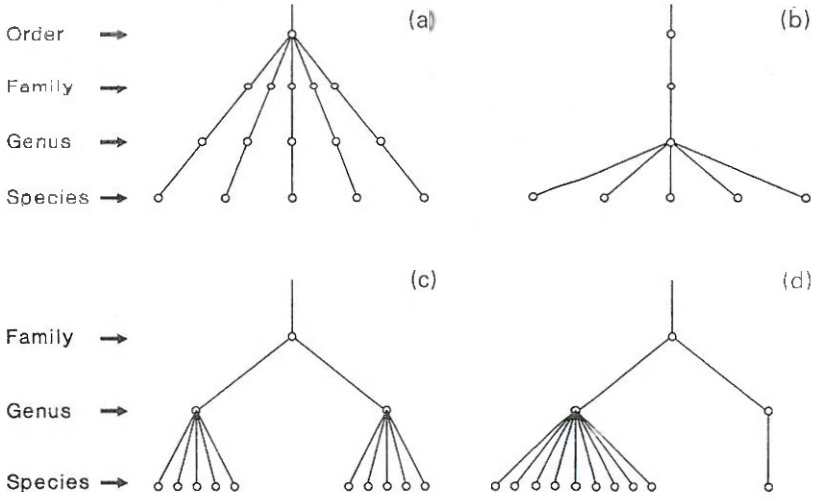

Fig. 2. Some simple, contrasting taxonomic trees for presence/absence data (i.e. ignoring species abundance information). Diversity measures based only on topology of the trees would not distinguish (a) from (b), and measures based on total branch length would not distinguish (c) from (d), but taxonomic distinctness Δ^+, based on the average of pair-wise path lengths (equation 4 of the text), does draw these distinctions. Using a simple (1, 2, 3, ...) weighting of path lengths, Δ^+ values are (a) 3·0, (b) 1·0, (c) 1·56 and (d) 1·2, placing the four configurations in the intuitively expected distinctness order.

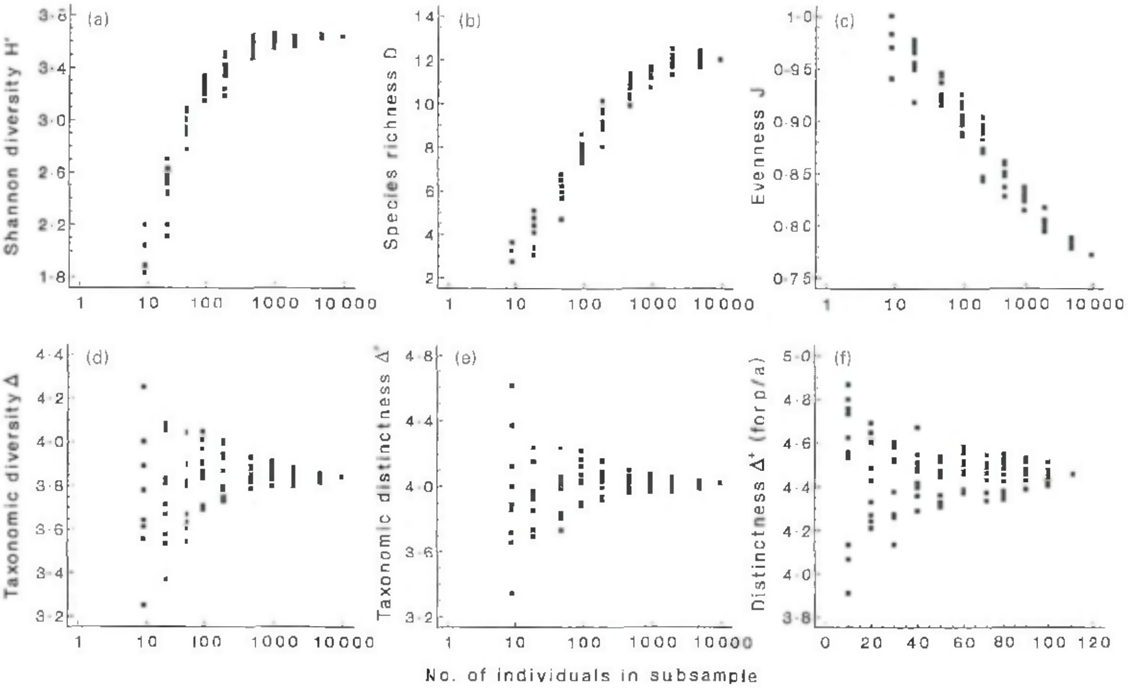

Fig. 3. Simulation study on the effects of sample size on (bio)diversity indices, using a single, composite sample of abundances of 111 nematode species (c. 10 000 individuals) from six sites in the Firth of Clyde (Lambshead 1986). Subsamples of individuals were drawn at random for 10 (logarithmically increasing) subset sizes, with 10 replicate simulations at each size, and the following indices computed: (a) Shannon diversity, (b) Margalef's d (a species richness index that attempts to adjust for sample size), (c) Pielou's J (reflecting evenness of abundances across species), (d) Δ (equation 1), (e) Δ^* (equation 3), (f) Δ^+ (equation 4). The simulations for the final plot ignored the species abundances and selected fixed numbers of *species* (from the 111) for computation of Δ^+. The conventional diversity indices are seen to be dependent on subsample size, unlike the taxonomic diversity/distinctness measures.

approximately (Δ_m^*) unbiased estimates of the respective true Δ and Δ^* for the whole sample, whatever the subset size *m*. The unbiasedness is also shown to be exact in the particular case where the data only records the presence or absence of species, not their abundances.

Methods

DEFINITION OF INDICES

'Taxonomic diversity' (Warwick & Clarke 1995) is defined, using the notation of Fig. 1, as:

$$\Delta = [\Sigma\Sigma_{i<j}\omega_{ij}x_ix_j + \Sigma_i 0.x_i(x_i - 1)/2]/$$
$$[\Sigma\Sigma_{i<j}x_ix_j + \Sigma_i x_i(x_i - 1)/2]$$
$$= [\Sigma\Sigma_{i<j}\omega_{ij}x_ix_j]/[n(n-1)/2] \qquad \text{eqn 1}$$

where x_i ($i = 1, \ldots, s$) denotes the abundance of the ith species, n ($= \Sigma_i x_i$) is the total number of individuals in the sample and ω_{ij} is the 'distinctness weight' given to the path length linking species i and j in the hierarchical classification. The double summations are over all pairs of species i and j (with $i < j$, for sake of definiteness). The first form of equation 1 exemplifies the construction of Δ as a pair-wise, averaged (weighted) path length in the diagram; the (null) second term in the numerator is included here to emphasize the zero path length defined for two individuals of the same species. In more formal statistical terms, Δ is the expected path length between any two randomly chosen individuals from the sample. The second, more succinct, form of equation 1 shows the relationship of Δ to standard diversity indices: when $\omega_{ij} = 1$ (for all $i < j$), i.e. when the taxonomic hierarchy is ignored, Δ reduces to a form of Simpson diversity (e.g. Pielou 1975), namely:

$$\Delta^{\circ} = [2\Sigma\Sigma_{i<j}p_ip_j]/(1 - n^{-1}), \quad \text{where } p_i = x_i/n,$$
$$= [(\Sigma_i p_i)^2 - \Sigma_i p_i^2]/(1 - n^{-1})$$
$$= (1 - \Sigma_i p_i^2)/(1 - n^{-1}). \qquad \text{eqn 2}$$

Indeed, the Simpson index was first constructed from the probability that any two organisms, selected at random from the full set of individuals, are from the same species (Simpson 1949). Taxonomic diversity Δ can therefore be seen as a generalization of Simpson diversity, incorporating an element of taxonomic relatedness.

This motivates the introduction of a second index, 'taxonomic distinctness', Δ* (Warwick & Clarke 1995), which is modified to remove some of the overt dependence of Δ on the species abundance distribution represented by the $\{x_i\}$. It divides the taxonomic diversity of equation 1 by the Simpson-type index of equation 2, i.e. Δ is divided by its value when the hierarchical classification collapses to the special case of all species belonging to a single genus. By definition, the resulting ratio, Δ*, must be more nearly a function of pure taxonomic relatedness of individuals. The algebraic definition of taxonomic distinctness is:

$$\Delta^* = [\Sigma\Sigma_{i<j}\omega_{ij}x_ix_j]/[\Sigma\Sigma_{i<j}x_ix_j] \qquad \text{eqn 3}$$

and an alternative way of viewing this is as the expected (weighted) path length between any two randomly chosen individuals from the sample, conditional on them being from different species. Note that, unlike Δ, the expression for Δ* is invariant to a scale change in x, so that Δ* could incorporate straightforwardly cases where the data are not counts of individuals but, say, total biomass for each species. It would also accommodate the use of transformed counts, e.g. the $\log(1 + x)$ or $x^{0{\cdot}25}$ transformations commonly used in multivariate community analysis to down-weight the contributions of dominant species (Clarke & Green 1988). A special case (in a sense, the ultimate down-weighting transformation) is the use only of presence/absence information for each species. The $\{x_i\}$ are then all thought of as equating to unity (for species that are present) and Δ and Δ* reduce to the same statistic, namely:

$$\Delta^+ = [\Sigma\Sigma_{i<j}\omega_{ij}]/[s(s-1)/2] \qquad \text{eqn 4}$$

where s is the number of species present and, for the double summation, i and j range over these s species.

The statistical results of the paper concentrate only on two cases: when the $\{x_i\}$ are (untransformed) counts and when only presence/absence information is available. They encompass any weighting scheme for the $\{\omega_{ij}\}$. The examples in this and the companion paper (Warwick & Clarke 1998), however, all use the simplest possible weights, in the context of the class of free-living marine nematodes: $\omega = 1$ (species in the same genus), 2 (same family but different genera), 3 (same suborder but different families), 4 (same order but different suborders), 5 (same subclass but different orders), 6 (different subclasses). For example, treating the subtree in Fig. 1 as a complete sample, a simple (0, 1, 2) weighting of the levels gives, from equations 1, 3 and 4, the values $\Delta = 1{\cdot}48$, $\Delta^* = 1{\cdot}82$ and $\Delta^+ = 1{\cdot}80$. Note, however, that any other set of increasing weights would honour the structure implied by a taxonomic classification. A natural refinement would be for the weights to depend on the quantitative reduction in taxon richness on moving up the hierarchy, although the number of species per genus, genera per family, etc., would need to be set globally for each faunal group, in some way, for the index to be comparable across studies.

For all three indices, the effect of the denominator terms in equations 1, 3 and 4 is to reduce or eliminate direct dependence on the number of species s: thus in Fig. 3d–f there is an apparently static mean value for Δ, Δ* and Δ⁺, whatever the subsample size (or, in the case of Fig. 3f, whatever the number of species in the subset). Naturally, the variance increases with decreasing information in all cases. The Δ statistics cannot therefore claim to represent all aspects of a community's diversity: if the results of Fig. 3 have some theoretical generality then they can be seen to partition out from species richness some combination of higher taxonomic spread and evenness, making the claim that average distinctness in a sample can be reliably estimated, while total distinctness in the sample (which clearly depends on the richness) cannot.

MEANS AND VARIANCES

For the model underlying Fig. 3d,e, the full set of species abundances $\{x_i;\ i = 1, \ldots, s\}$ and the total species richness s are thought of as fixed, and the 'true' taxonomic diversity and distinctness values are given by equations 1 and 3, with the $\{\omega_{ij}\}$ being known weights. A random subsample (without replacement) of a fixed number of organisms, m, is taken from the full set of individuals n $(=\Sigma_i x_i)$, and Δ_m and $\Delta_m{}^*$ denote the calculated taxonomic diversity and distinctness values from that subsample. The Appendix (case 1) shows quite generally that Δ_m is an unbiased estimate of Δ, and $\Delta_m{}^*$ is approximately unbiased for Δ^* (using a Taylor series), whatever the subsample size m and the structure of the trees.

An important special case is when only presence/absence information is available, and the subsamples now draw species at random (without replacement): m species from the full set of s. The distinctness for the subsample, $\Delta_m{}^+$, is again an (exactly) unbiased *estimate of* Δ^+ *for the full species set, for any* m (see the Appendix, case 2). This firms up the suggestion from Fig. 3f that the mean of a number of repeated subsamples at each size is constant, and there is no subsampling bias.

The Appendix develops these results for mean values formally, in order to set the symbolism for derivation of variance formulae, but note that there is a simpler heuristic explanation for the exact unbiasedness of Δ_m and $\Delta_m{}^+$. For example, the expectation of $\Delta_m{}^+$ is just the expected path length between two randomly selected species from a subset of m species. However, the latter subset is selected randomly from the full set of s species, and a random pair from a random sample of m species $(m > 1)$ is also a random pair from the full set of s species. By definition, Δ^+ is the expected path length for a randomly selected pair of species from the full set of s species, so it must follow that $E(\Delta_m{}^+) = \Delta^+$. Similar reasoning yields the exact unbiasedness result for Δ_m but not for $\Delta_m{}^*$, because of the conditionality clause in its definition; recourse needs to be made to the Taylor series approximation of the Appendix.

The Appendix then goes on to show that the variance of the subsample estimate $\Delta_m{}^+$ has the following form:

$$\mathrm{var}(\Delta_m{}^+) = 2(s-m)[m(m-1)(s-2)(s-3)]^{-1}$$

$$[(s-m-1)\sigma^2_{\omega} + 2(s-1)(m-2)\sigma^2_{\bar\omega}] \qquad \text{eqn 5}$$

where:

$$\sigma_{\omega}{}^2 = [(\Sigma_i\Sigma_{j(\neq i)}\omega_{ij}{}^2)/s(s-1)] - \bar\omega^2 \qquad \text{eqn 6}$$

$$\sigma_{\bar\omega}{}^2 = [(\Sigma_i\bar\omega_i{}^2)/s] - \bar\omega^2 \qquad \text{eqn 7}$$

$$\bar\omega_i = (\Sigma_{j(\neq i)}\omega_{ij})/(s-1) \qquad \text{eqn 8}$$

$$\omega = (\Sigma_i\bar\omega_i)/s = (\Sigma_i\Sigma_{j(\neq i)}\omega_{ij})/[s(s-1)] \equiv \Delta^+. \qquad \text{eqn 9}$$

These two σ^2 terms are straightforward properties of the taxonomic tree for the full species set, with $\sigma_\omega{}^2$ corresponding to the variance of all path lengths $\{\omega_{ij}\}$ between different species, and $\sigma_{\bar\omega}{}^2$ the variance of the mean path lengths $\{\bar\omega_i\}$ from each species to all others. Note that equation 5 is an exact result not a Taylor series approximation.

These sampling properties now motivate a statistical test for increase or decrease in observed taxonomic distinctness, based either on direct simulation or approximate confidence intervals (of the usual mean ± 2 SD form), constructed from the variance expression of equation 5.

Results

A PRACTICAL TEST FOR CHANGE IN TAXONOMIC DISTINCTNESS

The fact that, for presence/absence data, the distinctness estimate $(\Delta_m{}^+)$ from a subset of m species unbiasedly estimates the distinctness (Δ^+) of the full set, suggests the following test scenario for situations in which, at first sight, no valid diversity comparisons seem possible. The starting assumption is that there exists a reasonably comprehensive species list (inventory) for a region, within which certain localities are postulated to have reduced diversity. If the only data available at these localities are local species lists from one-off studies, and there is no control of the sampling effort expended in each location (or in constructing the regional inventory), then the only conventional diversity measure calculable – the number of species found at each locality – is uninterpretable. However, the above results show that one can unbiasedly compare *taxonomic distinctness at a locality with that for* the global list. For the null hypothesis of no difference, a randomization test can be performed by repeatedly subsampling species sets of size m, drawn at random from the global list, and constructing the histogram of the resulting $\Delta_m{}^+$ estimates. These will centre around the global distinctness of Δ^+ and the spread of the simulated values can be used to determine if the observed $\Delta_m{}^+$ for that locality is at variance with the null hypothesis.

Figure 4 is based on a UK species list for free-living marine nematodes $(s = 395$; see the companion paper Warwick & Clarke 1998), a nematode species list $(m = 122)$ from combined core samples taken over the course of a year at eight sandy sites in the Exe estuary, England, UK (Warwick 1971), and a further nematode species list $(m = 111)$ from six sandy sites in the Clyde estuary, Scotland, UK (Lambshead 1986). For a total of 1000 random samples of size $m = 122$ (for Fig. 4a), and a further 1000 random samples with $m = 111$ (for Fig. 4b), drawn from the global list, the $\Delta_m{}^+$ estimates give the histograms of Fig. 4a,b, showing the typically rather narrow range of distinctness values commensurate with the null hypothesis for

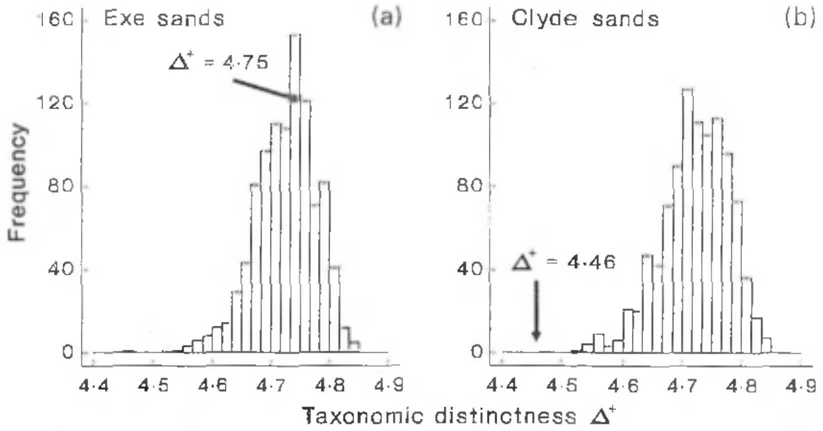

Fig. 4. Histogram of Δ^+ values for 1000 random subsamples of a fixed number m of species, from a full list of free-living marine nematodes of the UK ($s = 395$ species): (a) $m = 122$, (b) $m = 111$, corresponding to the sublist sizes for combined samples at intertidal sandy sites in the Exe and Clyde estuaries, respectively. The true Δ^+ values for both localities are also indicated: for the Clyde, the null hypothesis that the average distinctness equates with that for the UK as a whole is clearly rejected ($P < 0{\cdot}1\%$).

these subsample sizes. The true $\Delta_m{}^+$ for the Exe estuary sands, of 4·75, lies centrally to the distribution of Fig. 4a and therefore provides no evidence of a different average distinctness at this locality than in the UK region as a whole. To reject the null hypothesis, at approximately the 5% level, the true $\Delta_m{}^+$ would need to fall below the 25th lowest (of 1000) simulated $\Delta_m{}^+$ values in the histogram, or above the 25th highest. In contrast, the true $\Delta_m{}^+$ for the Clyde sands (4·46) is below this lower limit in Fig. 4b and in fact it is smaller than any of the 1000 simulated values, so there is significant evidence of a lower taxonomic distinctness here than for the UK as a whole ($P < 0{\cdot}1\%$).

The computational burden of this large number of simulations, which needs to be repeated for every locality under test (with a different species subset size), can be heavy, although not usually prohibitive. A much faster, approximate procedure is provided by the variance formula of equation 5. The constants $\sigma_\omega{}^2$ and $\sigma_\varpi{}^2$ in this expression are a function only of the tree structure of the global list (of s species) and need to be calculated only once (for all marine nematodes of the UK, for example). The variance expression is then a rather simple function of subsample size m and these constants, so that an approximate 95% confidence 'funnel' (mean ± 2 SD) can easily be constructed over the full range of m-values. Here the mean is equal to Δ^+ for the global list ($= 4{\cdot}72$ for UK marine nematodes) and the SD is the square root of the variance expression in equation 5. Figure 5 displays this funnel (the smooth, darker lines) and contrasts it with the results of extensive simulation runs (the circles, joined by lighter lines) for subset lists of $m = 10, 15, 20, 25, \ldots, 350$ species. At each point there are 1000 random selections and the circles denote the 25th lowest and 25th highest distinctness values (simulated 95% confidence limits). There is clear evidence of a left-skewed distribution for $\Delta_m{}^+$ in this case (as also for the Chilean nematode data of the companion paper; Warwick & Clarke 1998) but the normality approximation to the lower confidence limit (the important limit in practice) is good enough to suggest that this may be a useful short-cut to the full randomization procedure, in non-borderline cases, when computing power is limiting. An improved empirical fit could doubtless be constructed from an expression for the third moment of $\Delta_m{}^+$.

Discussion

As shown in Fig. 5, distinctness values for any specific locality, habitat type, pollution condition, etc., can be plotted on the confidence funnel created from a regional species list, to test for significant departures from the null hypothesis (that a particular subsample behaves, in terms of its pair-wise average distinctness, as if it were a random sample from the larger list). The companion paper, Warwick & Clarke (1998), applies and interprets this method in a range of situations.

It is perhaps surprising that a diversity test of any sort should be possible in a case where sampling effort is uncontrolled and the only data consist of presence or absence of species. Indeed, the test could not be expected to have the same sensitivity as that obtainable from a wider range of diversity measures (or multivariate analysis) calculable from abundance data in carefully standardized sampling plans. The key point to recognize here is that certain diversity features, most obviously the number of species recorded in a sample, are highly dependent on the sampling regime, and can only be straightforwardly compared under conditions of comparable sampling effort. The same caveats will apply to other diversity totals, such as PD, the total phylogenetic or taxonomic branch length in a subtree for a particular

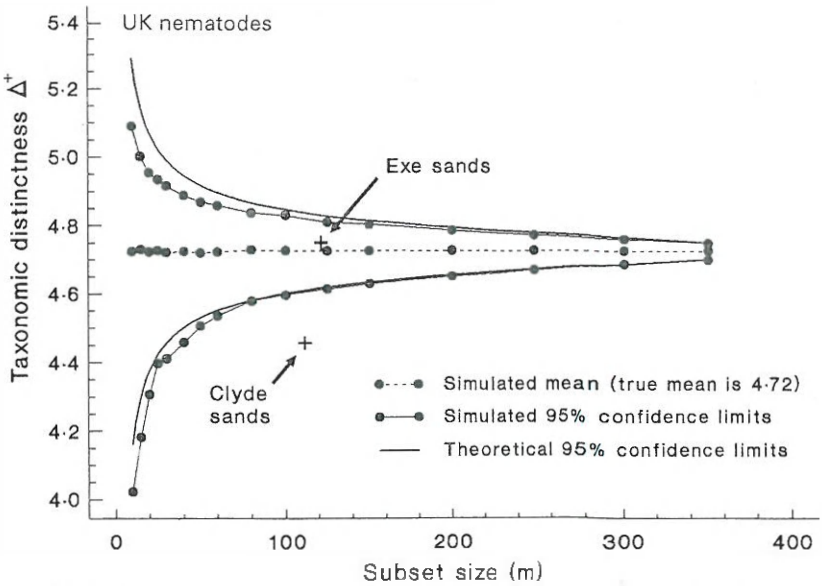

Fig. 5. Confidence funnels for the Δ^+ randomization test, from the all-UK list of marine nematode species. Circles correspond to direct randomization results for each sublist size, and smooth (thick) lines to approximate limits using the variance formula of equation 5. The dashed line gives the mean Δ^+ over each simulation, confirming the theoretical unbiasedness result ($\Delta^+ = 4{\cdot}72$ for the full set of 395 species).

locality/condition. They will not apply in general to average properties, such as the pair-wise taxonomic distinctness indices discussed here or, possibly, an average phylogenetic diversity, defined as PD/s. (Note though that, as pointed out earlier, the latter would have certain interpretational drawbacks: average PD takes the same value for Fig. 2c,d. It is also true that average PD calculated from a randomly selected sub-list of m species does not unbiasedly estimate average PD for the total list of s species, a fact that can be seen as further limiting the usefulness of this possible alternative formulation.)

Thus, for historic data and/or meta-analyses in which results from different workers are contrasted, there may be little choice but to recognize that only certain aspects of diversity, such as average taxonomic distinctness, may be validly compared. This raises a final question, on the extent to which the comparability of Δ^+ is compromised by the differing taxonomic identification skills of different workers. In fact, for Δ_m^+ to remain unbiased for Δ^+, it is not necessary to assume that all workers are equally efficient, only that taxonomic accuracy is independent of the taxonomic relatedness of the species involved. To put it another way, certain workers may miss (or mis-identify) species but, provided they do so at random across the species pool, in effect the test remains unchanged. (Whether low numbers of species are found because of low sampling effort or a low identification rate is then irrelevant to the construction of Δ^+.) Whether such an independence scenario is reasonable in practice is discussed further in the companion paper (Warwick & Clarke 1998).

Acknowledgements

This work forms part of the Marine Biodiversity project of the CCMS Plymouth Marine Laboratory, Natural Environment Research Council, UK, and is part-funded by the UK Ministry of Agriculture, Fisheries and Food (project no. AE1113). We thank Paul Harvey (University of Oxford) for useful contextual comments, and also an anonymous referee who contributed helpful insights into the derivation of results.

References

Clarke, K.R. & Green, R.H. (1988) Statistical design and analysis for a 'biological effects' study. *Marine Ecology Progress Series*, **46**, 213–226.

Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.

Faith, D.P. (1994) Phylogenetic pattern and the quantification of organismal biodiversity. *Philosophical Transactions of the Royal Society of London Series B*, **345**, 45–58.

Harper, J.L. & Hawksworth, D.L. (1994) Biodiversity: measurement and estimation. Preface. *Philosophical Transactions of the Royal Society of London Series B*, **345**, 5–12.

Humphries, C.J., Williams, P.H. & Vane-Wright, R.I. (1995) Measuring biodiversity value for conservation. *Annual Review of Ecology and Systematics*, **26**, 93–111.

Lambshead, P.J.D. (1986) Sub-catastrophic sewage and industrial waste contamination as revealed by marine nematode faunal analysis. *Marine Ecology Progress Series*, **29**, 247–260.

May, R.M. (1990) Taxonomy as destiny. *Nature*, **347**, 129–130.

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

Simpson, E.H. (1949) Measurement of diversity. *Nature*, **163**, 688.

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

Warwick, R.M. (1971) Nematode associations in the Exe estuary. *Journal of the Marine Biological Association of the United Kingdom*, **51**, 439–454.

Warwick, R.M. & Clarke, K.R. (1995) New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, **129**, 301–305.

Warwick, R.M. & Clarke, K.R. (1998) Taxonomic distinctness and environmental assessment. *Journal of Applied Ecology*, **35**, 532–543.

Williams, P.H., Humphries, C.J. & Vane-Wright, R.I. (1991) Measuring biodiversity: taxonomic relatedness for conservation priorities. *Australian Systematic Botany*, **4**, 665–679.

Received 31 August 1997; revision received 2 April 1998

Appendix

CASE 1: RANDOM SUBSAMPLES OF INDIVIDUALS

For the situation represented by Fig. 3d,e, the exact unbiasedness of Δ_m and asymptotic unbiasedness of $\Delta_m{}^*$, as estimators for Δ and Δ^*, respectively, is demonstrated under random subsampling (without replacement) of m organisms from the full set of n.

The species abundances $\{x_i;\ i = 1, \ldots, s;\ \Sigma_i x_i = n\}$ and the total number of species s are thought of as fixed, with the taxonomic diversity Δ and distinctness Δ^* for the full data set given by equations 1 and 3 of the main text. For a fixed-size (m) subset of individuals, denote the abundances of each of the s species by $\{Y_i;\ i = 1, \ldots, s;\ \Sigma_i Y_i = m\}$, capital letters reflecting the fact that these are the 'random variables'. The estimators of Δ and Δ^* from a sample of size m are:

$$\Delta_m = [\Sigma\Sigma_{i<j}\omega_{ij}Y_iY_j]/[m(m-1)/2] \qquad \text{eqn A1}$$

$$\Delta_m{}^* = [\Sigma\Sigma_{i<j}\omega_{ij}Y_iY_j]/[\Sigma\Sigma_{i<j}Y_iY_j]. \qquad \text{eqn A2}$$

The $\{Y_i\}$ are jointly hypergeometric, with probability distribution:

$$\begin{aligned}\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_s = y_s)\\ = (x_1Cy_1)(x_2Cy_2)\ldots(x_sCy_s)/(nCm) \qquad \text{eqn A3}\end{aligned}$$

and mean values:

$$E(Y_i) = mx_i/n \quad (i = 1, \ldots, s). \qquad \text{eqn A4}$$

Using the fact that the expectation of a sum of random variables is the sum of the expectations, even when non-independent, the expectation of Δ_m is:

$$E(\Delta_m) = [\Sigma\Sigma_{i<j}\omega_{ij}E(Y_iY_j)]/[m(m-1)/2] \qquad \text{eqn A5}$$

It can be shown from equation A3 that:

$$\begin{aligned}E(Y_iY_j) = [m(m-1)x_ix_j]/[n(n-1)]\\ (i, j = 1, \ldots, s;\ i \neq j) \qquad \text{eqn A6}\end{aligned}$$

so that:

$$E(\Delta_m) = [\Sigma\Sigma_{i<j}\omega_{ij}x_ix_j]/[n(n-1)/2] \equiv \Delta. \qquad \text{eqn A7}$$

In a similar way, but this time utilizing an asymptotic Taylor series expansion to express the mean of a ratio as approximately the ratio of the means, and again using equation A6:

$$\begin{aligned}E(\Delta_m{}^*) &\approx [\Sigma\Sigma_{i<j}\omega_{ij}E(Y_iY_j)]/[\Sigma\Sigma_{i<j}E(Y_iY_j)]\\ &= [\Sigma\Sigma_{i<j}\omega_{ij}x_ix_j]/[\Sigma\Sigma_{i<j}x_ix_j] \equiv \Delta^*. \qquad \text{eqn A8}\end{aligned}$$

CASE 2: RANDOM SUBLISTS OF SPECIES

The exact unbiasedness of the $\Delta_m{}^+$ estimator for Δ^+ is demonstrated for random sublists of m species drawn (without replacement) from the full list of s species. In addition, the exact variance of $\Delta_m{}^+$ is derived, as a basis for confidence funnels, such as that in Fig. 5.

This is a special case of the formulation in case 1, with abundances taking the values $x_i \equiv 1$ for all $i = 1, \ldots, s$ species present in the full set; the taxonomic distinctness Δ^+ of the full tree (395 species in the UK nematode example of Fig. 5) is given by equation 4 of the main text. For a fixed-size (m) sublist of species, drawn randomly without replacement, the random variables $\{Y_i;\ i = 1, \ldots, s;\ \Sigma_i Y_i = m\}$ now take only the values 0 or 1 ('indicator' variables), dependent on whether the ith species is absent or present, respectively, from the sublist. The taxonomic distinctness $\Delta_m{}^+$ for the sublist is defined as:

$$\Delta_m{}^+ = [\Sigma\Sigma_{i<j}\omega_{ij}Y_iY_j]/[m(m-1)/2] \qquad \text{eqn A9}$$

where the double summation is over all species $\{i, j = 1, \ldots, s;\ i < j\}$.

The joint probability distribution of the $\{Y_i\}$ now reduces to the simple case:

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_s = y_s) = 1/(sCm) \qquad \text{eqn A10}$$

i.e. all combinations of m species drawn from s are equally likely. It follows that:

$$\begin{aligned}E(Y_i) = m/s,\ E(Y_iY_j) = [m(m-1)]/[s(s-1)]\\ (i, j = 1, \ldots, s;\ i \neq j) \qquad \text{eqn A11}\end{aligned}$$

and:

$$\begin{aligned}E(\Delta_m{}^+) &= [\Sigma\Sigma_{i<j}\omega_{ij}E(Y_iY_j)]/[m(m-1)/2]\\ &= [\Sigma\Sigma_{i<j}\omega_{ij}]/[s(s-1)/2] \equiv \Delta^+ \qquad \text{eqn A12}\end{aligned}$$

establishing the unbiasedness of $\Delta_m{}^+$ as an estimator of Δ^+.

Derivation of the variance result, of equation 5 of the main text, starts from a modified form of equation A9, using the symmetry in the weights ($\omega_{ij} \equiv \omega_{ji}$) and the standard formula for the variance of a sum of random variables:

$$\begin{aligned}\mathrm{var}(\Delta_m{}^+) &= \mathrm{var}([\Sigma_i\Sigma_{j(\neq i)}\omega_{ij}Y_iY_j]/[m(m-1)])\\ &= [m(m-1)]^{-2}\Sigma_i\Sigma_{j(\neq i)}\\ &\qquad [\Sigma_k\Sigma_{r(\neq k)}\omega_{ij}\omega_{kr}\,\mathrm{cov}(Y_iY_j, Y_kY_r)] \qquad \text{eqn A13}\end{aligned}$$

where the four summations are all in the range $1, \ldots, s$. Consider now only the inner pair of summations, over (k, r), for a specific (i, j) (with $j \neq i$). Under the various combinations of subscript equivalences, e.g. $(k = i, r = j)$, $(k = i, r \neq i \text{ or } j)$, $(k \neq i, r = j)$, etc., only three different covariance terms emerge (note that in equations A14–A19 subscripts i, j, k, r all differ):

$$\text{(i)}\ \mathrm{cov}(Y_iY_j, Y_iY_j) \equiv \mathrm{var}(Y_iY_j) = a - a^2 \qquad \text{eqn A14}$$

where, because the $\{Y_i\}$ are indicator variables taking only the values (0, 1):

$$a = \mathrm{E}(Y_i^2 Y_j^2) \equiv \mathrm{E}(Y_i Y_j) = \Pr\{Y_i = 1, Y_j = 1\}$$
$$= [m(m-1)]/[s(s-1)] \quad \text{eqn A15}$$

$$\text{(ii)}\ \mathrm{cov}(Y_i Y_j, Y_i Y_r) = b - a^2 \quad \text{eqn A16}$$

where:

$$b = \mathrm{E}(Y_i^2 Y_j Y_r) = \Pr\{Y_i = 1, Y_j = 1, Y_r = 1\}$$
$$= [m(m-1)(m-2)]/[s(s-1)(s-2)] \quad \text{eqn A17}$$

$$\text{(iii)}\ \mathrm{cov}(Y_i Y_j, Y_k Y_r) = c - a^2 \quad \text{eqn A18}$$

where:

$$c = \mathrm{E}(Y_i Y_j Y_k Y_r) = [m(m-1)(m-2)(m-3)]/$$
$$[s(s-1)(s-2)(s-3)]. \quad \text{eqn A19}$$

Summing over the inner pair of subscripts (k, r) in equation A13 gives the (i, j)th term as:

$$\omega_{ij}[2(a-a^2)\omega_{ij} + 2(b-a^2)(\omega_{io} + \omega_{oj} - 2\omega_{ij})$$
$$+ (c-a^2)(\omega_{oo} - 2\omega_{io} - 2\omega_{oj} + 2\omega_{ij})] \quad \text{eqn A20}$$

where, in standard statistical notation, a circle indicates summation across that subscript:

$$\omega_{io} = \Sigma_{j(\neq i)}\omega_{ij},\ \omega_{oj} = \Sigma_{i(\neq j)}\omega_{ij},\ \omega_{oo} = \Sigma_i\Sigma_{j(\neq i)}\omega_{ij}.$$
$$\text{eqn A21}$$

Substituting equation A20 into equation A13, the summation over the two outer subscripts (i, j) is:

$$(c-a^2)\omega_{oo}^2 + 4(b-c)\Sigma_i\omega_{io}^2$$
$$+ 2(a-2b+c)\Sigma_i\Sigma_{j(\neq i)}\omega_{ij}^2 \quad \text{eqn A22}$$

and using the definitions of ω $(\equiv\Delta^+)$ and the 'variance-like' properties σ_ω^2 and $\sigma_{\bar{\omega}}^2$ in equations 6–9 of the main text, equation A13 becomes:

$$\mathrm{var}(\Delta_m^+) = m^{-2}(m-1)^{-2}s(s-1)\{2(a-2b+c)\sigma_\omega^2$$
$$+ 4(b-c)(s-1)\sigma_{\bar{\omega}}^2 + [2(a-2b+c)$$
$$+ 4(b-c)(s-1) + (c-a^2)s(s-1)]\bar{\omega}^2\}$$
$$\text{eqn A23}$$

On substituting for a, b and c from equations A15, A17 and A19, the coefficient of $\bar{\omega}^2$ disappears, and the desired variance formula is obtained:

$$\mathrm{var}(\Delta_m^+) = 2(s-m)[m(m-1)(s-2)(s-3)]^{-1}$$
$$[(s-m-1)\sigma^2_\omega + 2(s-1)(m-2)\sigma^2_{\bar{\omega}}]. \quad \text{eqn A24}$$