ABSTRACT

Most ecological diversity indices summarize the information about the relative abundances of species without reflecting taxonomic differences between species. Nevertheless, in environmental conservation practice, data on species abundances are mostly irrelevant and generally unknown. In such cases, to summarize the conservation value of a given site, so-called ‘taxonomic diversity’ measures can be used. Such measures are based on taxonomic relations among species and ignore species relative abundances. In this paper, bridging the gap between traditional biodiversity measures and taxonomic diversity measures, I introduce a parametric diversity index that combines species relative abundances with their taxonomic distinctiveness. Due to the parametric nature of the proposed index, the contribution of rare and abundant species to each diversity measure is explicit.

Keywords
Expected species diversity, taxonomic distinctness, taxonomic trees.

INTRODUCTION

Traditional diversity indices such as the Shannon entropy or the Simpson index summarize the information about the relative abundances of species within a community or sample without regard to differences between species. Nevertheless, for large-scale environmental protection, data on species abundances are often unknown or are of little use for conservation purposes. The usual treatment of species abundances is largely meaningless for systematically remote organisms such as oaks and orchids, which are commonly the focus of conservation efforts (Izsák & Papp, 2000). Vane-Wright et al. (1991) were the first to suggest that, for conservation purposes, we should quantify the taxonomic values we assign to different species ignoring their relative abundances. In this view, they proposed a measure of taxonomic distinctiveness of individual species based on phylogenetic relationships among species. Vane-Wright et al. (1991) further developed an index of community taxonomic distinctness simply by adding together the values of taxonomic distinctiveness of the species in a sample. Notice that in this paper, I use the terms ‘taxonomic distinctness’ and ‘taxonomic distinctiveness’ sensu Warwick & Clarke (2001). They take ‘taxonomic distinctiveness’ to refer to the distinctiveness of a particular species in relation to the rest of the community. By contrast, ‘taxonomic distinctiveness’ is taken to be a property of the community as a whole, e.g. an index of whether one community has a high taxonomic ‘breadth’ in relation to another community in a different habitat, biogeographical location, etc.

The proposal of Vane-Wright et al. (1991) is based only on the topology of cladistic classifications and is appropriate when branch lengths are unknown, whereas Faith (1992) suggested measuring taxonomic distinctness based on known branch lengths. The resulting measure of ‘phylogenetic diversity’ (PD) is simply the cumulative branch length of the full phylogenetic tree. In this view, Crozier (1992) pointed out that phylogenetic measures rely excessively on taxonomic knowledge and judgments concerning speciation. To overcome this drawback, he proposed summarizing taxonomic diversity from genetic distances between species, as measured, for example, by substitution rates for DNA sequences.

Also, detailed fully resolved cladograms at the species level are not available for most groups of organisms. This drawback is particularly important in environmental monitoring research, where the emphasis is on assessing whether sampled communities exhibit some changes in biodiversity following environmental degradation or remediation efforts. Here, the basic information on species relatedness is often just the set of pairwise distances between species based on a Linnaean taxonomy (Izsák & Papp, 1995; Clarke & Warwick, 1998; Rogers et al., 1999). These distances can also be based on morphological or functional differences (Izsák & Papp, 1995), or on more refined molecular biological methods (Solow et al., 1993; Shimatani, 2001).

Rao (1982) proposed an index termed quadratic entropy Q that incorporates both species relative abundances and a measure of the pairwise taxonomic distances between species.
Consider an N-species community characterized by the relative abundance vector \( p = (p_1, p_2, \ldots, p_N) \) such that \( 0 \leq p_i \leq 1 \) and \( \sum_{i=1}^{N} p_i = 1 \). Rao (1982) defined quadratic entropy as

\[
Q = \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} p_i p_j
\]

(1)

where \( d_{ij} \) is the difference between the i-th and the j-th species \( (d_{ij} = d_{ji}, \text{and} d_{ii} = 0) \) and \( Q \) expresses the average difference between two randomly selected individuals with replacement. It is easily shown that if \( d_{ij} = 1 \) for all \( i \neq j \), \( Q \) reduces to the Simpson index \( 1 - \sum_{i=1}^{N} p_i^2 \).

Using pairwise species distances obtained from the structure of Linnaean classifications, Warwick & Clarke (1995) demonstrated a continuous decrease in the quadratic entropy of a marine assemblage along a gradient of increasing environmental contamination, whereas Shimatani (2001), examining the effects of different forest management techniques in south-west Michigan (USA), showed that the thinning operations for promoting survival of specific species increased the taxonomic diversity of the vegetation.

However, since the ultimate aim of any summary statistics is to provide a manageable tool for characterizing and comparing different multivariate sets based on distinct objectives and motivations, it is generally understood that different indices may inconsistently rank a given pair of sets. The main reason for this confusion is that, by mapping the structure of a multidimensional set such as a biological community with scalars, information is necessarily lost, and there is no ideal function capable of uniquely characterizing all aspects of taxonomic diversity. A paradigmatic example within the context of conservation biology is Faith’s (1992) criticism on the measure of taxonomic distinctness proposed by Vane-Wright et al. (1991). Faith (1992) notes that in re-examining reserve-selection scenarios based on a phylogeny of bumble bees (Apidae), the index PD produces different priorities for species conservation relative to the measure proposed by Vane-Wright et al. (1991). Therefore, some would argue that diversity indices ‘conceal more than they reveal’ (Rousseau et al., 1999). But ecological data are often multivariate of high dimension so there is a need for summarization.

As suggested by a number of researchers (Hurlbert, 1971; Hill, 1973; Patil & Taillie, 1982; Tóthmérész, 1995), a more complete summarization of diversity requires a parametric family of indices whose members have varying sensitivities to the presence of rare and abundant species. In this paper, expanding on Hurlbert (1971) and Grassle & Smith (1976), I propose a parametric measure of taxonomic diversity that incorporates both species relative abundances and a measure of species taxonomic distinctiveness.

THE EXPECTED TAXONOMIC DISTINCTNESS

The index of Hurlbert (1971) is one of the most valuable measures for summarizing parametric species diversity given the proportions \( p_i \) of individual species in a sample. The expected number of species present when a random sample of size \( m \) is drawn with replacement from the community is given by (Hurlbert, 1971; Grassle & Smith, 1976):

\[
N(m) = \sum_{i=1}^{N} (1 - (1 - p_i)^m)
\]

(2)

The index \( N(m) \) is generally termed ‘expected species diversity’. Conceptually, the parameter \( m \) is an integer \( (m \geq 1) \), but non-negative real values make (also) mathematical sense (Tóthmérész, 1995). An essential advantage of using the \( N(m) \) index for summarizing community diversity is that one can set the index sensitivity to differences in the domain of rare species. By increasing the value of the parameter \( m \), one can enhance the sensitivity of this measure to the small species abundances and the expected species diversity will be more and more sensitive to the abundances from a wider range of species (not only the dominant and subdominant ones). For \( m = 1, N(1) = 1 \), whereas for \( m \to \infty \), the expected species diversity converges to the actual number of species in the data set. Therefore, changing \( m \) allows for vector description of community diversity by means of its diversity profile of \( N(m) \) vs. \( m \) (Patil & Taillie, 1982).

Similarly, imagine a non-negative measure \( w_i \) that quantifies the taxonomic distinctiveness of the i-th community species. In this case, the ‘expected taxonomic distinctness’ \( T(m) \) when a random sample of size \( m \) is drawn with replacement from the community is given by:

\[
T(m) = \frac{\sum_{i=1}^{N} w_i (1 - (1 - p_i)^m)}{N(m)}
\]

(3)

Eq. 3 represents a parametric measure of taxonomic distinctness in which the analysed taxonomic properties are associated with the single species instead of species pairs. For \( m = 1, T(1) \) equals the weighted mean of the single species taxonomic distinctiveness \( w_i \), 

\[
T(1) = \sum_{i=1}^{N} p_i w_i.
\]

Conversely, for \( m \to \infty \), the expected taxonomic distinctness converges to the algebraic mean of the single species taxonomic distinctiveness, \( T(\infty) = (1/N) \sum_{i=1}^{N} w_i \). Between these extremes, for \( m > 1 \), due to the complex interactions among species relative abundances \( p_i \) and their taxonomic distinctiveness \( w_i \), the values of the index \( T(m) \) are going to depend nonlinearly from the values of the parameter \( m \). This nonlinear behaviour should not be a cause for undue pessimism. To the contrary, it captures the intrinsic complexity that governs species assemblages as regards species relative abundances and their taxonomic relations in response to environmental constraints.

To apply the index \( T(m) \) in practice, the values of \( w_i \) need to be specified. For the purposes of this paper, the information available for computing species taxonomic distinctiveness \( w_i \) is the set of pairwise species distances \( d_{ij} \) (Solow & Polasky, 1994). Hence, I generated species distances based on the topology of the taxonomic tree (see Izsák & Papp, 1995; Ricotta, 2002): let \( d_{ij} \) stand for the topological distance (i.e. the number of edges) between the i-th and the j-th species in the Linnaean taxonomic tree. An illustration of an artificial taxonomic tree along with its species distance matrix is given in Fig. 1. For example, in Fig. 1, the taxonomic distance \( d_{AC} \) from species A to species C is 4.
Based on the species distance matrix $\Delta$ of Fig. 1, a simple way to compute a per-species measure of taxonomic distinctiveness $w_i$ associated to the $i$-th species in the data set is to add all $d_{ij}$ elements along row $i$ (or column $i$) of $\Delta$ and then dividing the result by the number of nonzero distances $N - 1$. This results in a vector $w = (w_1, w_2, \ldots, w_n)$ whose elements $w_i$ represent the taxonomic distinctiveness of the $i$-th species measured as the average distance between the $i$-th species and all other species in the taxonomic tree. For example, in Fig. 1, the taxonomic weight $w_A$, associated to species $A$ is $18/4 = 4.5$.

Regarding index properties, since $T(m)$ is not a traditional diversity measure computed uniquely from species relative abundances, it violates the usual diversity axiom that for a given number of species $N$ the maximal diversity arise for an equiprobable species distribution (i.e. a distribution where $p_i = p$ for all $i \neq j$). Another diversity axiom that is violated by $T(m)$ is the permutation invariance (Pielou, 1975). This postulates that the diversity values corresponding to the relative abundances $p_1, p_2, \ldots, p_n$ and to a $p'_1, p'_2, \ldots, p'_n$ permutation of those are identical. In a previous paper (Ricotta, 2002), I defined the diversity indices such as $T(m)$ that have their greatest value for noncompletely even communities ‘weak diversity indices’.

**CONCLUSIONS**

To the best of my knowledge, a parametric diversity index that combines species relative abundances and species taxonomic distinctiveness is proposed here for the first time. As shown in the previous paragraphs, when the parameter $m$ is a continuous variable, $T(m)$ describes taxonomic diversity as a scaling process that takes place not in the real, but in the data space.

However, it should be clear that the problem of measuring taxonomic diversity remains very much open and the proposed measure is complementary rather than competitive to more conventional diversity indices (Vane-Wright et al., 1991; Crozier, 1992; Faith, 1992; Solow & Polasky, 1994; Clarke & Warwick, 1998; Rogers et al., 1999; Izsák & Papp, 2000; Ricotta, 2002).

In the mean time, Eq. (3) might be useful if imperfect to summarize parametric taxonomic distinctness. For instance, the taxonomic distinctiveness $w_i$ of a given species depends on the presence (and taxonomic affiliation) of the other species in the community. This means that the presence of a single taxonomically distant species even if found in negligible abundance, would increase the taxonomic distinctiveness of all other species, thus increasing the taxonomic distinctness of the whole community.

This is a necessary feature of any measure concerning the taxonomic distinctness of single species. Whether it is or is not a problem depends on the specific questions we are asking. For instance, for interspecific comparisons it can be a problem (as the measured species features should not be affected by the momentary context). To the contrary, it decreases in effect for intercommunity comparisons, as differences in the overall taxonomic structure of a collection of interacting organisms is exactly what $T(m)$ tries to measure.

Alternatively, to fix the problem of the current formulation of the taxonomic distinctiveness, Webb (2000) suggests that the $w_i$ values are computed from a taxonomic tree that includes all species found in all communities to be compared. This would make the $w_i$ invariant within the bound of the specific analysis.

Finally, it is worth noticing that the same formulation can also be used to compute functional diversity (i.e. the extent of functional differences among species in a given community). Starting from a matrix of species functional traits (see for example Golluscio & Sala, 1993; Chapin et al., 1996), a functional distance matrix can be easily obtained by standard multivariate methods as proposed by Petchev & Gaston (2002). Next, combining the pairwise functional distances between species with their relative abundances, a parametric measure similar to $T(m)$ can be obtained.

**REFERENCES**


