Ecology, 90(12), 2009, pp. 3566–3574

Associations between species and groups of sites: indices and statistical inference

Miquel De Cáceres[1] and Pierre Legendre

Département de Sciences Biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7 Canada

Abstract. Ecologists often face the task of studying the association between single species and one or several groups of sites representing habitat types, community types, or other categories. Besides characterizing the ecological preference of the species, the strength of the association usually presents a lot of interest for conservation biology, landscape mapping and management, and natural reserve design, among other applications. The indices most frequently employed to assess these relationships are the phi coefficient of association and the indicator value index (IndVal). We compare these two approaches by putting them into a broader framework of related measures, which includes several new indices. We present permutation tests to assess the statistical significance of species–site group associations and bootstrap methods for obtaining confidence intervals. Correlation measures, such as the phi coefficient, are more context-dependent than indicator values but allow focusing on the preference of the species. In contrast, the two components of an indicator value index directly assess the value of the species as a bioindicator because they can be interpreted as its positive predictive value and sensitivity. Ecologists should select the most appropriate index of association strength according to their objective and then compute confidence intervals to determine the precision of the estimate.

Key words: *community types; fidelity measure; indicator value; phi coefficient of association.*

Introduction

In many ecological studies, researchers analyze the relationship between single species and one or more groups of sites. The classification of sites into groups (hereafter called "site groups") may have been derived from the similarities in environmental conditions among sites (e.g., habitat types, disturbance states), or in species composition (i.e., community types); it may also have been given by the study design (e.g., comparison of geographic regions or sampling years) or obtained using other criteria. With respect to the species, the analysis of its strength of association with site groups provides a characterization of its ecological preferences (e.g., Harms et al. 2001). With respect to the site group, the list of species strongly associated to the site group has a lot of interest for predictive purposes. For example, in vegetation studies, plant species that preferably occur in a single or a few vegetation types are generally called "diagnostic species," and are useful for the identification of vegetation types in field surveys (Chytrý et al. 2002*b*). A species restricted to one or a few habitat types potentially represents a better ecological indicator of environmental change than a habitat generalist, owing to the greater susceptibility of the specialist to local or regional extinction. Species of this kind are called "indicator species" by ecologists, and are used to monitor environmental changes and assess the impacts of disturbances on an ecosystem (Carignan and Villard 2002). McGeoch (1998) suggested that indicator species could be used in three distinct ways: (1) to reflect the biotic or abiotic state of the environment; (2) to reveal evidence for the impact of environmental changes; and (3) to indicate the diversity of other species, taxa, or communities within an area. Both "diagnostic" and "indicator" species essentially refer to the same concept: using the preference of the species for predicting purposes. This concept is frequently applied in conservation, land management, and reserve design.

In European phytosociology, the study of species–site group associations has a long tradition. The strength of the association was called "fidelity," which was defined as a measure of species concentration in vegetation units. The determination of diagnostic species is still an active research topic in vegetation science (e.g., Bruelheide 1995, 2000, Chytrý et al. 2002*a*, *b*, Tichý and Chytrý 2006, De Cáceres et al. 2008, Willner et al. 2009), where the most widely used index is the phi coefficient of association or modified forms of it (Chytrý et al. 2002*b*, Tichý and Chytrý 2006, Willner et al. 2009). In contrast, many ecologists prefer to determine indicator species using the IndVal index (Dufrêne and Legendre 1997). Although some comparisons between the two approaches have been made (Tichý and Chytrý 2006, De Cáceres et al. 2008), ecologists still lack a thorough framework to help them understand the differences between the available indices and determine when to apply one or

Manuscript received 1 October 2008; revised 9 March 2009; accepted 10 March 2009. Corresponding Editor: H. H. Wagner.
[1] E-mail: miquelcaceres@gmail.com

TABLE 1. Non-equalized indices of association between a species and a group of sites.

Index type	Correlation indices	Indicator value indices
Presence–absence	$r_\Phi = \dfrac{N \times n_p - n \times N_p}{\sqrt{(N \times n - n^2) \times (N \times N_p - N_p^2)}}$†	$\sqrt{\text{IndVal}_{pa}} = \sqrt{A_{pa} \times B_{pa}} = \sqrt{\dfrac{n_p}{n} \times \dfrac{n_p}{N_p}}$†
General abundance	$r_{pb} = \dfrac{N \times a_p - a \times N_p}{\sqrt{(N \times l^2 - a^2) \times (N \times N_p - N_p^2)}}$†	$s_{cos} = \sqrt{\dfrac{a_p}{l^2} \times \dfrac{a_p}{N_p}}$
Individual-based (partial)		$\sqrt{\text{IndVal}_{ind}} = \sqrt{A_{ind} \times B_{pa}} = \sqrt{\dfrac{a_p}{a} \times \dfrac{n_p}{N_p}}$†
Individual-based (full)	$r_{ind} = \dfrac{N \times a_p - a \times N_p}{\sqrt{(N \times c \times a - a^2) \times (N \times N_p - N_p^2)}}$	$s_{ind} = \sqrt{A_{ind} \times B_{ind}} = \sqrt{\dfrac{a_p}{a} \times \dfrac{a_p}{c \times N_p}}$

Notes: IndVal is the indicator value of a species, which is the product of two quantities, A and B, and r is the correlation index. See *Indices for assessing species–site group associations* for details. For the presence–absence case, we follow the notation used in previous works (Bruelheide 2000, Chytrý et al. 2002*b*, Tichý and Chýtrý 2006, De Cáceres et al. 2008): N, total number of sites; N_p, number of sites belonging to the target site group; n, number of occurrences of the species among all sites; n_p, number of occurrences of the species within the target site group. Indices for species abundance data require the following values: a_p, sum of the abundance values of the species within the target site group; a, sum of the abundance values of the species over all sites; l, norm of the vector abundances of the species; c, constant representing the total number of individuals or the total biomass per site.

† Recommended indices.

another. With the aim of improving the statistical practices of ecologists regarding species–site group associations, in the present work we compare the phi coefficient and indicator value approaches by integrating them into a broader algebraic framework of indices. We also address methods of inference for this kind of ecological relationships, specifically covering permutation tests of hypotheses and confidence intervals derived from bootstrap methods.

INDICES FOR ASSESSING SPECIES–SITE GROUP ASSOCIATIONS

In this section, we present several species–site group association indices and elaborate their mathematical relationships. Some of these indices are well known to ecologists, but others are less known or have never been used before. All of them share the following properties: (1) the association with a site group is calculated for each species independently; unlike the TWINSPAN method (Hill 1979), the pattern observed for a given species does not influence the association value of another species to a site group; (2) the strength (not the statistical significance) of the association is independent of the number of observations; (3) the strength of the association is bounded, i.e., there is a maximum and a minimum value; (4) in the case of indices for species abundance data, the strength of the association is invariant to multiplication of species abundance values by a constant. To follow the description of the indices, readers should refer to Tables 1 and 2, which present the formulas of all indices, as well as Fig. 1, which displays the mathematical relationships between indices.

TABLE 2. Group-equalized indices of association between a species and a group of sites.

Index type	Correlation indices	Indicator value indices
Presence–absence	$r_\Phi^g = \dfrac{N \times n_p^g - n^g \times N_p^g}{\sqrt{(N \times n^g - n^{g2}) \times (N \times N_p^g - N_p^{g2})}}$†	$\sqrt{\text{IndVal}_{pa}^g} = \sqrt{A_{pa}^g \times B_{pa}} = \sqrt{\dfrac{n_p/N_p}{\sum_{k=1}^{K} n_k/N_k} \times \dfrac{n_p}{N_p}}$†
General abundance	$r_{pb}^g = \dfrac{N \times a_p^g - a^g \times N_p^g}{\sqrt{(N \times l^{g2} - a^{g2}) \times (N \times N_p^g - N_p^{g2})}}$†	$s_{cos}^g = \sqrt{\dfrac{a_p/N_p}{\sum_{k=1}^{K} l_k^2/N_k} \times \dfrac{a_p}{N_p}}$
Individual-based (partial)		$\sqrt{\text{IndVal}_{ind}^g} = \sqrt{A_{ind}^g \times B_{pa}} = \sqrt{\dfrac{a_p/N_p}{\sum_{k=1}^{K} a_k/N_k} \times \dfrac{n_p}{N_p}}$†
Individual-based (full)	$r_{ind}^g = \dfrac{N \times a_p^g - a^g \times N_p^g}{\sqrt{(N \times c \times a^g - a^{g2}) \times (N \times N_p^g - N_p^{g2})}}$	$s_{ind}^g = \sqrt{A_{ind}^g \times B_{ind}} = \sqrt{\dfrac{a_p/N_p}{\sum_{k=1}^{K} a_k/N_k} \times \dfrac{a_p}{c \times N_p}}$

Notes: In addition to the quantities defined in Table 1, the following symbols are used: K, number of site groups; N_k, number of sites belonging to the kth site group; n_k, number of occurrences of the species in the kth site group; a_p, sum of the abundance values of the species in the kth site group. Quantities to use as substitutes in correlation indices are computed as $N_p^g = N/K$, $n_p^g = N_p^g(n_p/N_p)$, $n^g = N_p^g \times \Sigma_{k=1}^{K} (n_k/N_k)$, $a_p^g = N_p^g(a_p/N_p)$, $a^g = N_p^g \times \Sigma_{k=1}^{K}(a_k/N_k)$, and $l^{g2} = N_p^g \times \Sigma_{k=1}^{K}(l_k^2/N)$.

† Recommended indices.

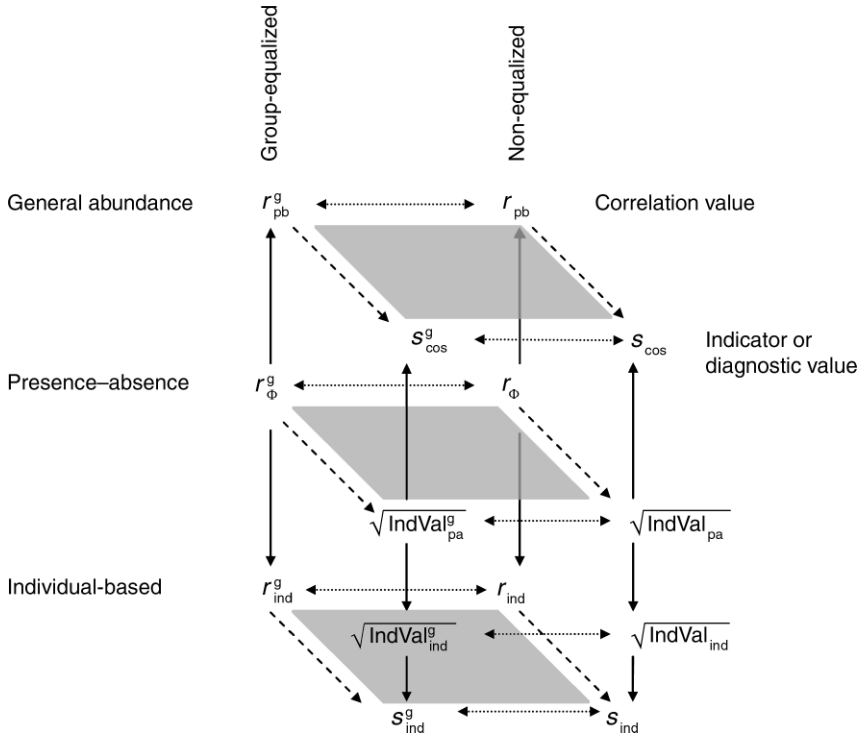

FIG. 1. Representation of the relationships among the 14 association indices considered in this study. Dotted horizontal arrows indicate equal values when groups have the same relative size. Downward/upward arrows indicate generalizations from presence/absence to abundance data. Dashed diagonal arrows indicate the limit value on data sets encompassing a large ecological range. Notation is as in Tables 1 and 2.

Correlation indices

We begin with species presence–absence data and the phi coefficient of association (r_Φ). This well-known index is the Pearson correlation computed using two binary vectors: a vector containing "1" for sites where the species is present and "0" for those where it is absent, and a vector with "1" for sites belonging to the group of interest and "0" for sites belonging to other groups. There are at least two possible generalizations of r_Φ to species abundance data. One is the point-biserial correlation coefficient (r_{pb}), which is the Pearson correlation computed between a quantitative vector (i.e., the vector containing the species abundance values at the various sites) and a binary vector (i.e., the vector of site membership values). The second generalization is intended to be used with individual count data, and it is equivalent to the one proposed by Chao et al. (2006) for site similarity measures. This latter generalization consists in computing r_Φ on the presence-absence of individuals instead of the presence–absence of the species (see Legendre and Legendre 1998:463 for more information). Unfortunately, this approach requires knowing the total number of sampled individuals in a site, and hence it uses information from species other than the species of interest. In order to fulfill the first above-mentioned requirement, we need to assume that the total number of individuals per site is constant for all sites. The resulting individual-based correlation index (r_{ind}) is rather similar to r_{pb} (see Table 1). In order to avoid setting the total number of individuals to an arbitrary value one can transform the species data table by dividing each abundance value by the sum of the abundances at the corresponding site, and then set the constant c equal to one. Willner et al. (2009) suggested a very similar generalization of the phi coefficient, in which they used percentages of cover as abundance values; they did not explicitly assume that all sites should have a sum of covers equal to 100%.

A disadvantage of the three correlation measures is that they are dependent on the relative size of the target site group (i.e., the proportion of sites in the data set that belong to the target site group, N_p/N), and this precludes comparisons between values corresponding to site groups of different relative sizes. Tichý and Chytrý (2006) recognized this limitation in r_Φ and suggested a modification to allow such comparisons. The modification consists in equalizing the relative sizes of all site groups. Several of the quantities used to compute r_Φ are affected by this modification (see Table 2 for details). The group-equalized phi coefficient (r^g_Φ) is obtained by substituting the original quantities in r_Φ by the modified ones. We can similarly define group-equalized counterparts for r_{pb} and r_{ind}, obtaining two new indices r^g_{pb} and r^g_{ind} (see Table 2). If all site groups originally have the same size (i.e., the same number of sites) then the following equalities hold: $r_\Phi = r^g_\Phi$, $r_{pb} = r^g_{pb}$ and $r_{ind} = r^g_{ind}$.

An alternative modification of r_Φ, which we will not include in our framework, consists in fixing the relative size of the target group to an arbitrary user-defined parameter (Knollová et al. 2005, Tichý and Chytrý 2006).

Indicator value indices

Dufrêne and Legendre (1997:350) defined the indicator value (IndVal) of a species as the product of two quantities, called A and B. For species abundance data, (IndVal^g_{ind} in Table 2), quantity A was defined as the mean abundance of the species in the target site group divided by the sum of the mean abundance values over all groups (A^g_{ind}). The sum of the mean abundances over all groups was used, instead of the sum of the actual abundances over all groups, in order to control for the effect of unequal sizes of the site groups. Quantity B was defined as the relative frequency of occurrence (presence–absence) of the species inside the target site group (B_{pa}). Note that the original definition of IndVal was based on the number of individuals to assess A but on presence–absence data to assess B. If A was also based on presence–absence data, we would have A^g_{pa}, the relative frequency of the species in the target site group divided by the sum of relative frequencies over all groups, and the resulting indicator value index would be IndVal^g_{pa} (Table 2). However, instead of using this latter index, Dufrêne and Legendre (1997:363) proposed that for species presence-absence data, indicator values should be calculated using IndVal_{pa} (Table 1), where quantity A is the absolute frequency of the species inside the target group divided by the sum of absolute frequencies (A_{pa}). Hence, Dufrêne and Legendre did not control for the effect of unequal sizes of the site groups when defining their IndVal index for presence–absence data. The abundance-based counterpart of IndVal_{pa} is IndVal_{ind} (Table 1), which is obtained by defining A as the sum of abundances of the species within the target site group divided by the sum of abundances of the species over all sites (A_{ind}). Analogously to correlation measures, we can see that A^g_{pa} and A^g_{ind} are the group-equalized counterparts of A_{pa} and A_{ind}, respectively. If all site groups originally have the same size then the following two equalities hold: $\text{IndVal}_{pa} = \text{IndVal}^g_{pa}$, $\text{IndVal}_{ind} = \text{IndVal}^g_{ind}$.

Relationships between correlation and indicator values for presence-absence data

With species presence–absence data, the correlation and indicator value approaches may yield similar results when computed on a data set with very high beta diversity. This happens because in that case there are a large number of zeros for many species and site groups (i.e., many double zeros). The limit of r_Φ when adding an infinite amount of double zeros is the Ochiai (1957) similarity coefficient, which is the square root of IndVal_{pa} (Janson and Vegelius 1981, De Cáceres et al. 2008). The Ochiai coefficient is closely related to two distance measures commonly used in community ecology, namely the chord (Orlóci 1967) and Hellinger (Rao 1995) distances (Legendre and Legendre 1998). Moreover, suppose that the first groups in a data set contain occurrences of the species of interest while the remaining groups do not. The limit of r^g_Φ when adding an infinite amount of site groups with no occurrences of the species is the square root of IndVal^g_{pa}. These relationships are the reasons why indicator value indices appear as the square roots of the original indices in Tables 1 and 2. We will hereafter follow this convention.

Indicator value alternatives for abundance data

At this point we may wonder what the limits of correlation indices in the case of species abundance data are. The limit of r_{pb} when adding double zeros is the cosine of the angle formed by the two vectors (s_{cos}). The cosine cannot take negative values because the species abundance values are always nonnegative. When computed with presence–absence data, s_{cos} is equal to the Ochiai index, and like its presence–absence counterpart it is closely related to the chord distance. Accordingly, the limit of r^g_{pb} when adding an infinite number of site groups with no occurrence of the species is s^g_{cos}, the group-equalized counterpart of s_{cos}. If we conduct the same limit operations on r_{ind} and r^g_{ind} we obtain two new indicator value indices, s_{ind} and s^g_{ind}, which allow us to understand how the indicator value measures for species abundance data fit into the general framework. Unlike $\sqrt{\text{IndVal}^g_{ind}}$ and $\sqrt{\text{IndVal}_{ind}}$, which are partial generalizations to species abundances, s^g_{ind} and s_{ind} are full generalizations following the individual-based approach (see Tables 1 and 2).

Synthetic data examples

We will use synthetic data to illustrate the differences between the four indices for presence–absence data described above (see Table 3). In all scenarios the species occurs in 50% of the sites belonging to Group 1, which is the target site group. In the basic scenario (scenario a, to which the other scenarios will be compared) the species is found in 25% of the sites belonging to Group 2. As the two site groups do not have the same relative sizes, $\sqrt{\text{IndVal}_{pa}}$ and $\sqrt{\text{IndVal}^g_{pa}}$ do not produce the same value; the same happens with the (r_Φ, r^g_Φ) pair. Equalizing the relative sizes of the groups while keeping the same frequency of the species within each group solves the disagreement within index pairs (scenario b). If one modifies the classification of the sites not belonging to the target site group (scenarios c and d) $\sqrt{\text{IndVal}_{pa}}$ and r_Φ remain unaltered compared to scenario a, but $\sqrt{\text{IndVal}^g_{pa}}$ and r^g_Φ both change their values. If sites with the target species absent are added to the nontarget group of sites (scenario e) r_Φ, $\sqrt{\text{IndVal}^g_{pa}}$ and r^g_Φ increase their values but not $\sqrt{\text{IndVal}_{pa}}$. Both indicator value indices are unaffected by the consider-

TABLE 3. Values obtained with presence–absence association indices for a set of synthetic scenarios (a–g).

Scenario	Site A	B	C	D	E	F	G	H	$\sqrt{\text{IndVal}^{g}_{pa}}$	$\sqrt{\text{IndVal}_{pa}}$	r^{g}_{Φ}	r_{Φ}
a												
Group	1	1	2	2	2	2			0.577	0.500	0.258	0.250
Species	1	0	0	0	1	0						
b												
Group	1	1	1	1	2	2	2	2	0.577	0.577	0.258	0.258
Species	1	1	0	0	0	1	0	0				
c												
Group	1	1	2	2	2	3			0.548	0.500	0.351	0.250
Species	1	0	0	0	1	0						
d												
Group	1	1	2	2	2	3			0.408	0.500	0.000	0.250
Species	1	0	0	0	0	1						
e												
Group	1	1	2	2	2	2	2		0.598	0.500	0.314	0.300
Species	1	0	0	0	1	0	0					
f												
Group	1	1	2	2	2	2	3	3	0.577	0.500	0.408	0.333
Species	1	0	0	0	1	0	0	0				
g												
Group	1	1	2	2	2	2			0.447	0.354	−0.258	−0.250
Species	1	0	1	1	0	1						

Notes: The group to which each site A–H belongs is indicated by a nominal variable "Group," whereas "Species" is a binary variable indicating the presence or absence of the species in each site. The index values measure the association of the species to target group 1. Scenario a is the scenario to which all others are compared. Differences between scenarios are described in *Interpretation of association indices: Correlation vs. indicator value indices*.

ation of an additional group of sites where the species is absent (scenario f), whereas the r_{Φ} and r^{g}_{Φ} values change. Finally, in scenario g, the species occurs in 75% of the Group 2 sites. Two correlation indices return negative values whereas the two indicator values are lower but remain positive.

Interpretation of Association Indices

In this section we address the ecological interpretation of the indices introduced in the previous section.

Correlation vs. indicator value indices

A careful look at the first column of Tables 1 and 2 reveals that the numerator of correlation indices is essentially a difference between the observed occurrence (or abundance) of species in the target site group and the expected occurrence (or abundance) under the hypothesis of no association. Negative correlation values tell us when a species "avoids" the target site group (scenario g in Table 3), a situation where traditional phytosociologists would say that the species shows "negative fidelity." Such ecological behavior could also be tested with IndVal by interchanging the role of presence and absence, but at the expense of having to recompute the association value (Dufrêne and Legendre 1997, McGeoch and Chown 1998).

Another important difference between the correlation and indicator value approaches is their treatment of absences of the species outside the target site group (scenarios e and f in Table 3). Correlation indices take into account absences outside the target site group, and these absences contribute to increase the strength of the association as much as the presences inside that group. In contrast, indicator value indices do not treat species absences in the same way in the target site group and in its complementary set. Indicator value indices assume that the species can be absent from many conceivable habitats, and having fewer or more absences in the complementary set of sites is irrelevant and not taken into account. This makes correlation indices more context dependent than indicator value indices. Nevertheless, the ecological and geographical context surrounding the assessment of species–site group relationships is always crucial and should be explicitly stated in all applications (e.g., Chytrý et al. 2002*a*, Willner et al. 2009). In short, correlation indices indicate the degree of preference for the target site group compared to the other groups, whereas indicator value indices assess how much the target site-group matches the set of sites where the species is found.

Components of the indicator value

Dufrêne and Legendre (1997) used the terms "specificity" and "fidelity" for quantities A and B, respectively, but we avoid this terminology here because it differs from the traditional phytosociological concept of fidelity (Tichý and Chytrý 2006, Willner et al. 2009). Instead, we prefer to use terms that relate to the contribution of A

and B to bioindication. The interpretation of A depends on whether species presence-absence or abundance data are considered: A_{pa} is the probability that a site belongs to the target site group, given the fact that the species has been found. It is therefore the positive predicting value of the species as indicator of the site group (Murtaugh 1996). A_{ind} is the probability of the target site group given the fact that an individual of the species has been found, so it has nearly the same interpretation as A_{pa}, but based on individuals. Group-equalized versions of these quantities do not change their basic interpretation. Quantity B_{pa} is the probability of finding the species when the site belongs to the target site group, thus expressing the sensitivity of the species as a bioindicator (Murtaugh 1996). A very high B_{pa} value for a site group means that if the species is not found at a given site the probability that the site belongs to that site group is low. The practical usefulness of B_{pa} comes from the fact that high values indicate species that are easier to detect in habitats belonging to the target site group. Note that B_{pa} could be adapted to quantify how easily the species is detected not only in space but also in time. Species whose phenology and/or behavior makes them visible during a short portion of the year may be present at all sites belonging to a site group when one pools survey data for the whole year. However, they will not be good indicators of that site group if seasons are treated as different observations because they cannot be observed during most of the year. One of the reasons why the indicator value approach is not very popular among vegetation scientists is that the B component gives too much weight to common species compared to rare species (e.g., Chytrý et al. 2002*b*, Tichý and Chytrý 2006). Rare plant species are often considered good indicators of vegetation types. That is, a species may be difficult to detect, but once found it is its positive predictive value that is of interest. Alternatively, Wagner and Edwards (2001) use A, and not IndVal, in order to assess the indicator value of a species without taking its rarity into account. Other combinations of A and B may be of interest (see McGeoch et al. 2002). A species with high values in both quantities has a good positive predictive value and is easy to find, so it becomes a very good species to be used as a bioindicator. Dufrêne and Legendre (1997) referred to species with either high A or high B values, but not both, as asymmetrical indicator species. This terminology conceptually matches the traditional phytosociology concept of asymmetrical fidelity (Juhász-Nagy 1964, De Cáceres et al. 2008).

The alternative quantitative indicator value indices present some problems regarding the interpretation of their components. The components in cosine measures (s_{cos} and s^{g}_{cos}) cannot be interpreted as positive predictive power or sensitivity (see Tables 1 and 2). Full generalizations following the individual-based approach (s_{ind} and s^{g}_{ind}) may not be recommendable either, because the interpretation of quantity B changes when computed with abundances instead of presence-absence data. B_{ind} is the average relative abundance of individuals of the species at a site that belongs to the target site group. It is hence measuring the dominance of the species in the community rather than its frequency, and it is less useful as a sensitivity measure. As an example, consider a species that is always found at sites belonging to the target site group, but occurs in low abundance compared to other species in the community. In that case, B_{pa} is one whereas B_{ind} has a low value.

Statistical Inference for Species–Site Group Associations

The study of the relationship between a species and a site group is not complete after computing the value of an index. It is also important to assess (1) whether the species is significantly preferring (or avoiding) the environmental conditions of the sites in the target site group and (2) the precision of the estimated association strength. Since many of the indices compared here have neither exact nor approximate known statistical distributions, statistical methods that randomize the observed data should be used to answer these questions. In this section we discuss the application of permutation tests to answer the first question and bootstrap confidence intervals for the second.

Significance tests for the association between a species and a site group

In order to determine if a species is associated with a site group, we must conduct a statistical test of the null hypothesis that there is no such relationship. Under this null hypothesis, the fact that the species was observed at a site belonging to the target site group is due to chance only. A permutation test is a procedure that involves comparing an observed test statistic with a distribution obtained by randomly reordering (i.e., permuting) the data. If the null hypothesis of no association is true, the association value computed after randomly reassigning species occurrence or abundance values to sites will be similar or very close to that observed for unpermuted data. The P value of the permutation test of positive (negative) species preference is the proportion of permutations that yielded the same or higher (lower) association values than that observed for the unpermuted data.

Any of the 14 association indices can be used in a permutation test. We list in Table 4 four distinct one-tailed tests that have a clear ecological interpretation of the null and alternative hypotheses. The first two address differences in frequency, whereas the other two focus on differences in relative abundance. Each case is duplicated depending on whether a non-equalized measure or a group-equalized measure is used as test statistic. The difference in interpretation of the null hypothesis is that group-equalized indices make explicit reference to the relative abundance (or frequency) of the species in each of the nontarget site groups, whereas non-equalized measures do not. Sometimes different statistics can be found that yield the same P value in

TABLE 4. Summary of the recommended and equivalent statistics for testing the significance of association by means of permutation tests.

Null ecological hypothesis	Recommended statistic	Equivalent statistics
Differences in frequency		
The species is not more (less) frequently found at sites belonging to the site group than at sites not belonging to it.	n_p	r_Θ, $\sqrt{\text{IndVal}_{pa}}$
The frequency of the species in the target site group is not higher (lower) than the frequency in the other site groups	A^g_{pa}	r^g_Φ, $\sqrt{\text{IndVal}^g_{pa}}$
Differences in abundance		
The abundance of the species in sites belonging to the site group is not higher (lower) than its abundance in sites not belonging to it	a_p	r_{pb}, r_{ind}, s_{ind}, s_{cos}
The average abundance of the species in the target site group is not higher (lower) than the average abundance in the other site groups	A^g_{ind}	

Notes: Notation follows Tables 1 and 2. In addition, s_{cos} is the cosine between the species abundance vector and the vector of membership to the site–group; s_{ind} is the full generalization of $\sqrt{\text{IndVal}_{pa}}$ following the individual-based approach.

permutation tests. The statistics are then said to be equivalent for the permutation testing (Edgington 1995). For example, when using species presence-absence data, the only quantity that is affected by the permutation of the values is the number of occurrences of the species within the target site group. This makes r_Φ and $\sqrt{\text{IndVal}_{pa}}$ equivalent statistics for testing whether the species is more frequently found among sites belonging to the target site group than at other sites. The number of occurrences of the species within the target group is also an equivalent statistic to r_Φ and $\sqrt{\text{IndVal}_{pa}}$, and it is the preferred statistic because it is the fastest to calculate. We report in the Appendix some power and type I error simulation results for this specific test.

The permutation test described in Dufrêne and Legendre (1997) deserves a special comment due to its widespread use. Regarding the interpretation of the results, looking at Table 4 we can see the ecological hypothesis being tested when using $\sqrt{\text{IndVal}^g_{pa}}$ or $\sqrt{\text{IndVal}_{pa}}$ as test statistics. Since $\sqrt{\text{IndVal}_{ind}}$ and $\sqrt{\text{IndVal}^g_{ind}}$ are partial generalizations to species abundance values, their corresponding null hypotheses are more difficult to interpret ecologically because they combine frequency and abundance values (but see Mouillot et al. 2002). Significance tests can be used, to study the association of a species to any one or all of the site groups. However, the number of simultaneous tests has to be taken into account when reporting the results. Regardless of the index used, testing multiple species–site group associations will require a correction for multiple testing and is likely to result in fewer significant species–site group associations. The recommendation of Dufrêne and Legendre (1997) was to test only the significance of the association with the group for which the indicator value was the highest. Moreover, they avoided the problem of multiple testing by using the maximum indicator value among the site groups as test statistic (i.e., the maximum value was recomputed after each permutation). The same approach can be adopted with other association indices in order to avoid corrections for multiple testing.

Some additional recommendations are in order with permutation tests. First, the statistical null model assumed in an unrestricted permutation approach (such as the one described above) may be very ecologically unlikely, to the extent that testing it may become irrelevant. Falsifying null hypotheses generated from more restricted null models of association (e.g., Gotelli 2000) may provide more interesting ecological information. Second, users should be aware that autocorrelation in the data could make permutation tests too liberal. One way to solve that problem in a grid of observations of rectangular shape is by restricting the possible permutations to those allowed by a toroidal shift (Harms et al. 2001, Fortin and Dale 2005) or by simulating random autocorrelated patterns (Roxburgh and Chesson 1998). Finally, practitioners should keep in mind that if the classification of the sites has been obtained from the species composition itself (for example by *K*-means partitioning), the site groups would not be completely independent of the species data. In such a case of circularity, we can expect more significantly associated species than expected by chance only.

Confidence intervals for species–site group associations

Like all measurements that imply statistical uncertainty, it is important to complement the measured strength of association along with an indication of the precision of the assessment. In the case of indicator values, we recommend computing confidence intervals not only for the indicator value index but also for quantities *A* and *B*, as suggested by De Cáceres et al. (2008). Many of the above-mentioned association indices do not allow deriving confidence intervals using parametric methods; therefore we recommend using bootstrap confidence intervals. Bootstrapping is the practice of estimating properties of an estimator by measuring those properties when sampling from an approximating distribution, usually the empirical distribution of the observed data. Efron's simple percentile bootstrap method (Efron 1979, Manly 1997) obtains the limits of the confidence interval for an estimator from the percentiles of the distribution of values generated by resampling the observed data with replacement. Using

simulated data we studied the performance of bootstrap confidence intervals for the four presence–absence association indices. Asymptotic parametric methods were also evaluated when available. Details of the simulation methods and results are given in the Appendix. In practical applications, bootstrap confidence intervals can be trusted for indicator values between 0.1 and 0.9 when the sample size is not small (i.e., $N > 100$ sites) and at least 30–50 sites belong to the target site group. Larger samples may be needed to obtain valid confidence intervals for indicator values near 0 or 1. The same rule can be used for correlation measures, and in this case it applies analogously for negative correlation values.

Final Remarks

We discussed here the properties of several related indices measuring the association between species and groups of sites. Although both the correlation and indicator value approaches can be successful in many applications, they have been developed with slightly different purposes. (1) For determining the ecological preference of a given species among a set of alternative site groups, the correlation approach is probably more useful than the indicator value approach, because the former naturally allows the detection of negative preferences. An example would be the determination of species-habitat associations (Harms et al. 2001, Gunatilleke et al. 2006). The phytosociological concept of "differential species" (a species whose pattern of occurrence in a particular community type can be used to differentiate between that type and closely related types) also matches this purpose (De Cáceres et al. 2008). For species abundance data, we believe that the point-biserial coefficients (r_{pb} or r^g_{pb}) generally seem to be a better choice than the individual-based indices (r_{ind} or r^g_{ind}). The reason is that assuming a constant total number of sampled individuals (or biomass) for all sites may be very unrealistic, especially in animal communities. (2) For assessing species predictive values (e.g., for field determination of community types or ecological monitoring), one should preferably compute indicator values (McGeoch and Chown 1998), because they are decomposable into components that are interpretable as probabilities. In addition, we believe the indicator value approach is well suited for analyzing species-habitat associations based on large biological record databases because it provides potentially unique values (De Cáceres et al. 2008). For example, we may want to assess the strength of the association between the nesting sites of a bird species and a given forest type. If the entire geographical distributions of the bird species and the forest type are considered and the sample of observation points can be assumed to be random, the indicator value will provide a single population value regardless of the beta-diversity range of the database. This is not true for the correlation approach even if the previous two conditions are met. The symmetric treatment of absences in correlation indices has the effect that any inclusion of new habitat types into the database leads to a different population value.

In order to compare association values for a species across site groups, the association index must be independent of the relative sizes of the groups. Non-equalized indices give the same weight to individual sites, whereas group-equalized indices give equal weights to all site groups. Using group-equalized measures can be considered similar to performing a group-based stratified re-sampling with the aim of ensuring that no site groups are over-sampled with respect to others. Therefore, the decision whether to use group-equalized indices is related to how the sampling was conducted.

We presented here permutation tests to assess the statistical significance of species–site group associations and bootstrap methods for obtaining confidence intervals. Permutation tests allow discarding species with no preference but do not allow selecting species with high association values. Therefore, we encourage ecologists to use confidence intervals in order to use them to evaluate the likeliness of arbitrary population association values. The behavior of inference methods with group-equalized indices is a topic that deserves further research. The main problem is that groups with small number of sites are very influential when randomizing or bootstrapping data. In the meanwhile, we recommend using group-equalized measures when all groups should contain a reasonable number of sites (i.e., at least 10).

Some other topics could not be covered in this paper. For example, species niche preferences may involve more than one group of sites, and this ecological situation is not properly addressed with the current analysis approach. Moreover, the results of species–site group association analyses have been suggested to be used in order to choose among competing partitions of the same set of sites, either representing hierarchical levels of a nested classification, or being the result of different partitioning methods (e.g., Dufrêne and Legendre 1997:362–363, Tuomisto et al. 2003). We believe these questions need further development and deserve to be addressed in future works.

Available Software

Some of the association measures cited in this manuscript are already available in computer programs, such as Marc Dufrêne's IndVal (*available online*),[2] PC-ORD (McCune and Mefford 1999), JUICE (Tichý 2002), or GINKGO (Bouxin 2005). The original (group-equalized) IndVal can also be computed using the "duleg" function of the R language package "labdsv" (Roberts 2006). We provide a Supplement with an R package containing functions to compute all the association measures discussed in this manuscript, including options for calculating simple bootstrap

[2] ⟨http://biodiversite.wallonie.be/outils/indval/home.html⟩

confidence intervals, and a function to perform the tests listed in Table 4.

Acknowledgments

The manuscript was substantially improved using comments provided by Melodie McGeoch and an anonymous reviewer. The authors are also grateful to Francesc Oliva for his useful recommendations on bootstrap methods, to Marco Moretti for pointing out interesting applied issues, and to Sapna Sharma for the revision of the English style. This research was funded by NSERC grant no. 7738 to P. Legendre.

Literature Cited

Bruelheide, H. 1995. Die Grünlandgesellschaften des Harzes und ihre Standortbedingungen. Dissertationes Botanicae 244: 1–338.

Bruelheide, H. 2000. A new measure of fidelity and its application to defining species groups. Journal of Vegetation Science 11:167–178.

Bouxin, G. 2005. Ginkgo, a multivariate analysis package. Journal of Vegetation Science 16:355–359.

Carignan, V., and M. Villard. 2002. Selecting indicator species to monitor ecological integrity: a review. Environmental Monitoring and Assessment 78:45–61.

Chao, A., R. L. Chazdon, R. K. Colwell, and T. J. Shen. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62:361–371.

Chytrý, M., A. Exner, R. Hrivnák, K. Ujházy, M. Valachoviè, and W. Willner. 2002*a*. Context-dependence of diagnostic species: a case study of the Central European spruce forests. Folia Geobotanica 37:403–417.

Chytrý, M., L. Tichý, J. Holt, and Z. Botta-Dukát. 2002*b*. Determination of diagnostic species with statistical fidelity measures. Journal of Vegetation Science 13:79–90.

De Cáceres, M., X. Font, and F. Oliva. 2008. Assessing species diagnostic value in large data sets: a comparison between phi coefficient and Ochiai index. Journal of Vegetation Science 19:779–788.

Dufrêne, M., and P. Legendre. 1997. Species assemblages and indicator species: the need for a flexible asymetrical approach. Ecological Monographs 67:345–366.

Edgington, E. S. 1995. Randomization tests. CRC Press, Boca Raton, Florida, USA.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7:1–26.

Fortin, M. J., and M. R. T. Dale. 2005. Spatial analysis: a guide for ecologists. Cambridge University Press, Cambridge, UK.

Gotelli, N. J. 2000. Null model analysis of species co-occurrence patterns. Ecology 81:2606–2621.

Gunatilleke, C. V. S., I. Gunatilleke, S. Esufali, K. E. Harms, P. M. S. Ashton, D. Burslem, and P. S. Ashton. 2006. Species–habitat associations in a Sri Lankan dipterocarp forest. Journal of Tropical Ecology 22:371–384.

Harms, K. E., R. Condit, S. P. Hubbell, and R. B. Foster. 2001. Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. Journal of Ecology 89:947–959.

Hill, M. O. 1979. TWINSPAN: a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Cornel University, Ithaca, New York, USA.

Janson, S., and J. Vegelius. 1981. Measures of ecological association. Oecologia 49:371–376.

Juhász-Nagy, P. 1964. Some theoretical models of cenological fidelity I. Acta Botanica Debrecina 3:33–43.

Knollová, I., M. Chytry, L. Tichy, and O. Hajek. 2005. Stratified resampling of phytosociological databases: some strategies for obtaining more representative data sets for classification studies. Journal of Vegetation Science 16:479–486.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier, New York, New York, USA.

Manly, B. F. J. 1997. Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall, Boca Raton, Florida, USA.

McCune, B., and M. J. Mefford. 1999. PC-ORD. Multivariate analysis of ecological data, version 4. MjM Software Design, Gleneden Beach, Oregon, USA.

McGeogh, M. A. 1998. The selection, testing and application of terrestrial insects as bioindicators. Biological Reviews 73: 181–201.

McGeoch, M. A., and S. L. Chown. 1998. Scaling up the value of bioindicators. Trends in Ecology and Evolution 13:46–47.

McGeoch, M. A., B. J. Van Rensburg, and A. Botes. 2002. The verification and application of bioindicators: a case study of dung beetles in a savanna ecosystem. Journal of Applied Ecology 39:661–672.

Mouillot, D., J. M. Culioli, and T. D. Chi. 2002. Indicator species analysis as a test of non-random distribution of species in the context of marine protected areas. Environmental Conservation 29:385–390.

Murtaugh, P. A. 1996. The statistical evaluation of ecological indicators. Ecological Applications 6:132–139.

Ochiai, A. 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bulletin of the Japanese Society of Scientific Fisheries 22:526–530.

Orlóci, L. 1967. An agglomerative method for classification of plant comunities. Journal of Ecology 55:193–206.

Rao, C. R. 1995. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió (Quaderns d'Estadistica i Investivació Operativa) 19:23–63.

Roberts, D. W. 2006. LABDSV: laboratory for dynamic synthetic vegephenomenology. R package version 1.2–2. ⟨http://cran.r-project.org/⟩

Roxburgh, S. H., and P. Chesson. 1998. A new method for detecting species associations with spatially autocorrelated data. Ecology 79:2180–2192.

Tichý, L. 2002. JUICE, software for vegetation classification. Journal of Vegetation Science 13:451–453.

Tichý, L., and M. Chytrý. 2006. Statistical determination of diagnostic species for site groups of unequal size. Journal of Vegetation Science 17:809–818.

Tuomisto, H., K. Ruokolainen, M. Aguilar, and A. Sarmiento. 2003. Floristic patterns along a 43-km long transect in an Amazonian rain forest. Journal of Ecology 91:743–756.

Wagner, H. H., and P. J. Edwards. 2001. Quantifying habitat specificity to assess the contribution of a patch to species richness at a landscape scale. Landscape Ecology 16:121–131.

Willner, W., L. Tichý, and M. Chytrý. 2009. Effects of different fidelity measures and contexts on the determination of diagnostic species. Journal of Vegetation Science 20:130–137.

APPENDIX

Inference for species–site group association indices (*Ecological Archives* E090-250-A1).

SUPPLEMENT

R package for performing indicator species analyses (*Ecological Archives* E090-250-S1).