Journal of Vegetation Science 17: 809-818, 2006

Statistical determination of diagnostic species for site groups of unequal size

Tichý, Lubomír[1,2] **& Chytrý, Milan**[1*]

[1]*Institute of Botany and Zoology, Masaryk University, Kotlářská 2, CZ-61137, Brno, Czech Republic;*
[2]*E-mail tichy@sci.muni.cz; *Corresponding author; Fax +420 532146213; E-mail chytry@sci.muni.cz*

Abstract

Aim: Concentration of species occurrences in groups of classified sites can be quantified with statistical measures of fidelity, which can be used for the determination of diagnostic species. However, for most available measures fidelity depends on the number of sites within individual groups. As the classified data sets typically contain site groups of unequal size, such measures do not enable a comparison of numerical fidelity values of species between different site groups. We therefore propose a new method of measuring fidelity with presence/absence data after equalization of the size of the site groups. We compare the properties of this new method with other measures of statistical fidelity, in particular with the Dufrêne-Legendre Indicator Value (*IndVal*) index.

Methods: The size of site groups in the data set is equalized, while relative frequencies of species occurrence within and outside of these groups are kept constant. Then fidelity is calculated using the phi coefficient of association.

Results: Fidelity values after equalization are independent of site group size, but their numerical values vary independently of the statistical significance of fidelity. By changing the size of the target site group relative to the size of the entire data set, the fidelity measure can be made more sensitive to either common or rare species. We show that there are two modifications of the *IndVal* index for presence/absence data, one of which is also independent of the size of site groups.

Conclusion: The phi coefficient applied to site groups of equalized size has advantages over other statistical measures of fidelity based on presence/absence data. Its properties are close to an intuitive understanding of fidelity and diagnostic species in vegetation science. Statistical significance can be checked by calculation of another fidelity measure that is a function of statistical significance, or by direct calculation of the probability of observed species concentrations by Fisher's exact test. An advantage of the new method over *IndVal* is its ability to distinguish between positive and negative fidelity. One can also weight the relative importance of common and rare species by changing the equalized size of the site groups.

Keywords: Community classification; Dufrêne-Legendre Indicator Value index; Fidelity; phi coefficient; Presence/absence data; Vegetation database.

Nomenclature: Ehrendorfer (1973).

Abbreviation: *IndVal* = Dufrêne-Legendre Indicator Value index.

Introduction

Diagnostic species are important for ecological interpretation of community classifications or habitat typologies (Whittaker 1962; Barkman 1989). Determination of diagnostic species is related to the concept of fidelity (Szafer & Pawłowski 1927), which is a measure of concentration of species occurrence or abundance within the target site group (cluster, community type) relative to other site groups or to the complementary part of the data set. Species with a high fidelity to the target site group are considered as its diagnostic species.

In vegetation science, the term *site*, i.e. the basic sampling unit, is usually identical to *relevé* and *site group* is identical to *relevé cluster, vegetation unit* or *syntaxon*. Although the concept of fidelity is popular especially in vegetation science, it is not exclusive to this field: it can be applied in all branches of community ecology whenever records of species composition of plant or animal assemblages are classified into groups. Therefore we will use the more general terms *site* and *site group* throughout the present paper.

Several statistical measures of species fidelity have been proposed recently, both for presence/absence data (Bruelheide 1995, 2000; Botta-Dukát & Borhidi 1999; Chytrý et al. 2002a) and quantitative data such as species abundance (Dufrêne & Legendre 1997). Chytrý et al. (2002a) reviewed several fidelity measures for presence/absence data, which can be divided into two types: (1) fidelity values increase with the size of the data set, i.e. with the number of sites; (2) fidelity values are independent of the size of the data set.

The former type includes the chi-square statistic, *G*-statistic of the likelihood ratio test, *u*-values and Fisher's exact test. Their increasing or decreasing values directly reflect an increase or decrease in statistical significance of fidelity. In small data sets containing few sites there is higher uncertainty that the observed patterns of species concentration in community types represent the real patterns existing in the statistical population. Therefore, species in small data sets are generally given lower values of fidelity to particular

community types, and comparisons of numerical values of these fidelity measures are not possible between data sets with different total numbers of sites.

The latter type, which includes the phi coefficient of association and the Dufrêne-Legendre Indicator Value (*IndVal*), is not influenced by the size of the data set. This is a suitable property for practical use, because in most applications users do not expect species fidelity to community types to depend on the number of available field records. However, the values of these fidelity measures are not functions of statistical significance of fidelity. Therefore there is a danger that species with non-significant occurrence concentration in some site groups will be given a high fidelity value and will be erroneously considered as diagnostic species. This is a disadvantage, but it can be easily removed by performing an additional independent calculation of statistical significance, e.g. by permutation test (Dufrêne & Legendre 1997) or Fisher's exact test (Chytrý et al. 2002a).

Except for the *IndVal* (in the sense of Eq. 16 in Chytrý et al. 2002a, i.e. the categorical form of the equation on p. 350 in Dufrêne & Legendre 1997), all the statistical fidelity measures for presence/absence data reviewed by Chytrý et al. (2002a) are dependent on the relative size of site groups within the data set, i.e. on the number of sites assigned to the target group divided by the total number of sites in the data set. This fact may result in situations where in two community types X and Y, X being smaller than Y, some species have higher relative frequency but lower fidelity in X than in Y (Chytrý et al. 2002a: Tables 5 and 6). Such results are correct from the statistical point of view, because estimates based on a smaller number of replicates (sites) are less reliable. Fidelity is therefore given a low numerical value in order to avoid the invalid conclusion that some species are diagnostic. From the practical point of view, however, this property of fidelity measures is not desirable. Classifications of site records of species composition nearly always produce partitions with groups of unequal size. In such classified data sets, numerical values of fidelity can be directly compared between different species within the same site group; however, comparisons of fidelity values of the same species between different site groups do not match intuitive expectations. Most users would expect fidelity to be stable for each pair of species and site group no matter how many site records are currently available.

The aim of this paper is to propose and test a method that would remove the effect of unequal size of site groups on fidelity calculations. Generally, the new method can be applied in combination with any statistical measure of fidelity, but if combined with the chi-square statistic, *G*-statistic, *u*-values or Fisher's exact test, the resulting numerical values are no longer functions of statistical significance, so the important property of these measures is lost. We will therefore describe this method in combination with the phi coefficient of association. We will compare it with *IndVal*, because this index also measures fidelity and one of its two forms is independent of the relative size of site groups.

The new method: phi coefficient applied to site groups of equalized size

phi coefficient of association

This coefficient is defined in terms of a 2 × 2 contingency table:

Number of sites ...	in the target site group	not in the target site group
containing the species	a	b
not containing the species	c	d

by the formula (Sokal & Rohlf 1995: 741, 743):

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+c)}}. \qquad (1)$$

Bruelheide (1995, 2000) introduced an alternative notation for defining statistical measures of fidelity, which was also used by Chytrý et al. (2002a):
N = number of sites in the data set,
N_p = number of sites in the target site group,
n = number of occurrences of the species in the data set,
n_p = number of occurrences of the species in the target site group.

The contingency table filled with this notation is as follows:

Number of sites ...	in the target site group	not in the target site group
containing the species	n_p	$n - n_p$
not containing the species	$N_p - n_p$	$N - N_p - n + n_p$

and the corresponding formula for the phi coefficient of association, derived by substitution of variables in Eq. 1 and subsequent simplification is (Chytrý et al. 2002a: eq. 9):

$$\Phi = \frac{N \cdot n_p - n \cdot N_p}{\sqrt{n \cdot N_p \cdot (N-n) \cdot (N-N_p)}} \qquad (2)$$

The phi coefficient ranges from –1 to 1. It is equal or close to zero when the species occurrence in the data set does not show any preference or avoidance of the target site group. Higher values indicate that species occurrences are concentrated in the target site group, and lower values indicate that they are under-represented in the target site group. A value of 1 indicates that the species occurs in all sites of the target site group and is

absent in all sites not belonging to the target site group; a value of –1 indicates the reverse pattern. For identification of diagnostic species, positive Φ-values are of particular importance, although negative Φ-values can be also used for negative differentiation of community types, especially if there are not too many site groups in the given typology.

Equalization of the size of site groups

The phi coefficient is independent of the data set size (N), but depends on the size of the target site group N_p, which may vary from 1 to (N – 1). To remove this dependence, we suggest here equalizing the size of all site groups in a data set (N_p) to a new value N_p'

$$N_p' = s \cdot N \qquad 3)$$

where s is a number ranging from 1/N to (N – 1)/N, which indicates the proportional size of the target site group after equalization relative to the size of the whole data set. In this equalization, relative frequencies of species within and outside the target site group, i.e. the quantities n_p / N_p and $(n - n_p) / (N - N_p)$, respectively, remain constant. Thus the number of species occurrences in the target site group after equalization is

$$n_p' = N_p' \cdot (n_p / N_p) = s \cdot N \cdot (n_p / N_p), \qquad (4)$$

the number of species occurrences outside the target site group is

$$n_n' = (N - N_p') \cdot [(n - n_p) / (N - N_p)], \qquad (5)$$

and the total number of species occurrences in the whole data set is

$$n' = n_p' + n_n' = s \cdot N \cdot (n_p / N_p) + (1 - s) \cdot N \cdot [(n - n_p) / (N - N_p)]. \quad (6)$$

The total number of sites in the data set (N) remains unchanged. If the new values N_p', n_p' and n' are substituted in turn for N_p, n_p and n in Eq. 2 and Φ is calculated for site groups that were of different size before equalization, the resulting Φ-values are directly comparable between site groups, provided the same value of s is used for each of these groups.

The value of s is selected subjectively. The most obvious choice is putting s equal to the reciprocal number of site groups contained in the data set (e.g. $s = 0.5$ for a data set with two groups and 0.25 for a data set with four groups). However, the choice of s need not be dependent on the number of site groups in the data set. For example, in a data set of four site groups such as that in Fig. 1a, we can set $s = 0.5$ and compute the phi coefficient for the first site group as if this group made up 50% of the entire data set size and the other three groups altogether were also 50%, then compute the phi coefficient for the second site group again as if this group was 50% of the entire data set size, etc. (Fig. 1b).

Table 1 shows a simple artificial data set with site groups of unequal size. If the phi coefficient is applied to a data set with site groups of non-equalized size, it yields different values even though the same relative frequencies of species occurrences within and outside the target site group are compared, i.e. 100% vs. 33.3% for species 1-3, and 100% vs. 66.7% for species 4-6. The above described equalization of the size of the target site groups (in this case, equalized size of 10% of the entire data set was used for each site group; $s = 0.1$), removes the dependence of the phi coefficient value on the size of site groups for species 1-6 but not for species 7 and 8. For species 7, the relative frequencies compared would be 100% vs. 66.7% for group A and 100% vs. 33.3% for group B. Therefore the phi coefficients of species 7 for group A and B differ, even though both compare relative frequencies of 100% with 100% and 0%. In this case, and also in the case of species 8, fidelity of species to the target site group depends on the unequal size of other site groups. This shows that fidelity calculations involving more than two site groups can be made entirely independent of the unequal size of site groups only if they are preceded by equalization of the size of all site groups of interest. The equalized size of the target site group can differ from the equalized size of the other site

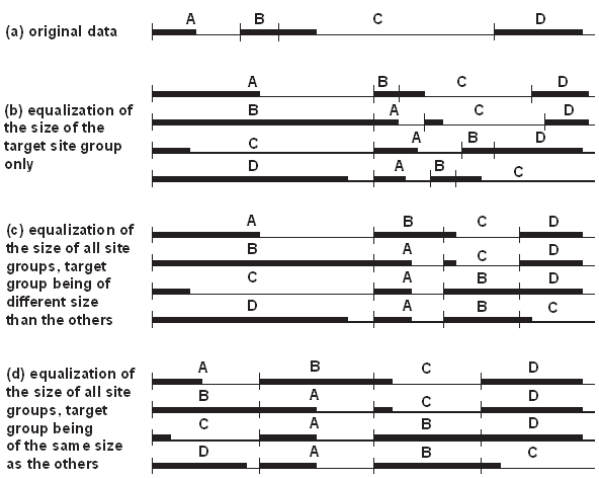

Fig. 1. A scheme of different equalizations of the size of site groups. Each line represents distribution of a species in the original (**a**) and equalized (**b**-**d**) data sets, each data set with four site groups. Segments are site groups labelled A, B, C and D, and segment lengths correspond to the number of sites in each group. The thick parts of each segment represent sites with occurrence of the given species, and the thin parts represent sites where the species is absent. In (**b**) to (**d**), the four lines represent in turn equalizations used for calculation of species fidelity to the target site groups A, B, C and D.

Table 1. Artificial species-by-sites data set with three site groups A, B and C. Left part shows individual sites with presence (x) and absence (.) of species; right part shows the Φ-values for these three site groups in three variants: (1) without equalization; (2) with the size of the target site group equalized to 10% of the entire data set size and the size of the other site groups non-equalized – equalization type corresponding to Fig. 1b; (3) with the size of the target site group equalized to 10% of the entire data set size and the other site groups equalized to the same size – equalization type corresponding to Fig. 1c. Only positive Φ-values are shown.

	Original data set			(1) non-equalized size of site groups			(2) equalized size of the target site group			(3) equalized size of all site groups		
	A	B	C	A	B	C	A	B	C	A	B	C
Species 1	xxx	xx....	xxx.....	0.500	-	-	0.408	-	-	0.408	-	-
Species 2	x..	xxxxxx	xxx.....	-	0.632	-	-	0.408	-	-	0.408	-
Species 3	x..	xx....	xxxxxxxxx	-	-	0.707	-	-	0.408	-	-	0.408
Species 4	xxx	xxxx..	xxxxxx...	0.277	-	-	0.218	-	-	0.218	-	-
Species 5	xx.	xxxxxx	xxxxxx...	-	0.378	-	-	0.218	-	-	0.218	-
Species 6	xx.	xxxx..	xxxxxxxxx	-	-	0.447	-	-	0.218	-	-	0.218
Species 7	xxx	xxxxxx	x.....	0.400	0.632	-	0.320	0.408	-	0.272	0.272	-
Species 8	x..	xxxxxx	xxxxxxxxx	-	0.250	0.354	-	0.140	0.167	-	0.218	0.218

groups (Fig. 1c) or can be the same as the equalized size of the other site groups (Fig. 1d). The Φ-values after such equalization are entirely independent of the size of site groups and can be directly compared across different site groups.

In some cases, however, it may not be desirable to equalize the size of all site groups in the data set. If diagnostic species for a few similar community types are determined using a large database, it may be useful to perform the fidelity calculation with a data set that includes not only the sites belonging to the community types of interest, but also sites of other, unrelated community types (Chytrý et al. 2002a). This approach enables the determination of diagnostic species of more general validity, because they are compared against the background of other community types in the same geographic area. Usually the group of sites of other community types is much larger than the aggregate size of the site groups of interest. An equalization of this large group to the size of the target site groups would strongly and undesirably reduce its effect. As there is hardly ever a reason for determination of diagnostic species of such a large and heterogeneous site group, its size can be held constant and the size of the other groups can be equalized in such a way that the sum of their equalized sizes is the same as the sum of their sizes before equalization.

Weighting the importance of common and rare species

The phi coefficient applied to the data set with site groups of equalized size is independent of the actual differences in size of individual site groups. However, it depends on the equalized relative size of the target site group (s), which may be either equal to the size of the other site groups or set to any arbitrary value between 1 and (N – 1). Fig. 2 shows the dependence of the Φ-values on the differences between relative frequency of species within and outside the target site group, i.e. n_p / N_p vs. (n – n_p) / (N – N_p). For the equalized size of the target site group (N_p') set to 10% of the entire data set (s = 0.1); (Fig. 2a), Φ is relatively high also for the species that are not very common within the target site group, provided the difference between relative frequency within and outside the tar-

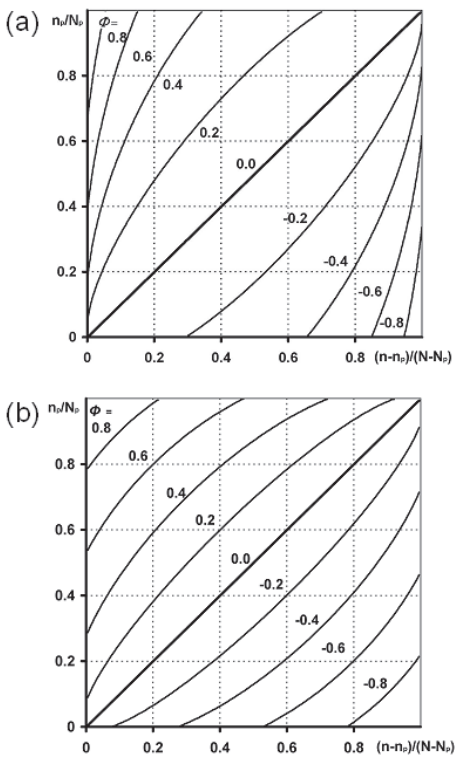

Fig. 2. Dependence of Φ on the relative frequency of species occurrences within (vertical axis) and outside (horizontal axis) the target site group, shown for site groups equal to (**a**) 10% and (**b**) 50% of the size of the entire data set.

get site group is large. However, if the difference in relative frequency is small, even the species with high relative frequency within the target site group are given low Φ-values. In contrast, if N_p' is set to a higher value, such as 50% of the size of the entire data set (s = 0.5; Fig. 2b), species must generally have higher relative frequencies within the target site groups in order to have high Φ-values; if a species has a high relative frequency within the target site group, a high Φ-value can be attained even with a smaller difference in relative frequency between the target and other site groups. These two graphs illustrate a general trend that setting the equalized relative size of the target site group (s) to a higher value gives a higher weight to common species and their relative frequency in the target site group. By contrast, setting s to a lower value gives a higher weight to rare species and to the differences in relative frequency of species within and outside the target site group. Changing of s-value can thus be used as a tool for modifying the properties of the phi coefficient with respect to weighting of common or rare species.

The effect of data set heterogeneity

It is important to note that the ratio of relative frequency of species within and outside the target site group depends on the heterogeneity of the data set. Fig. 3 shows the patterns of these relative frequencies for two selected examples from the grassland data sets used by Chytrý et al. (2002a). Fig. 3a compares site group A, which represents a type of rock-outcrop dry grassland and contains 204 relevés (= sites), within a data set of 502 relevés. All of these 502 relevés belong to various types of rock-outcrop dry grasslands from the Czech Republic. The size of both the target site group and the other site groups is equalized according to the scheme shown in Fig. 1c. Curves representing the value of $\Phi = 0.5$ are inserted in the graph for different sizes of the target site group equalized to 50%, 10% and 2% of the entire data set (s = 0.5, 0.1 and 0.02, respectively). Due to rather high similarity between the target and other site groups in this relatively homogeneous data set, most species have a similar relative frequency within and outside the target site group, and few species attain high Φ-values. Fig. 3b shows a similar comparison of the same site group, but within a data set of 15 989 sites of various types of Czech grasslands. Due to the heterogeneity of this data set, relative frequencies of most species inside and outside the target site group are very different. Therefore several species attain high Φ-values. If only the size of the target site group was equalized but not the sizes of other groups, dots representing individual species would be slightly shifted to the left or right in Fig. 3b.

Comparison of the new method with the Dufrêne-Legendre Indicator Value index

Standard Indicator Value index: IndVal$_1$

Dufrêne & Legendre (1997) proposed an index called *Indicator Value* (*IndVal*), which is also suitable for determination of diagnostic species of site groups. This index measures concentration of species abundance or occurrence within the target site group, weighted by the relative frequency of species occurrence within this group. It should be noted that Dufrêne & Legendre (1997) used the term *fidelity* for the relative frequency of species occurrence within the target site group (n_p/N_p), i.e. the variable that is usually called *constancy* in vegetation science. They define *IndVal* as a product of *specificity* (measure of abundance or

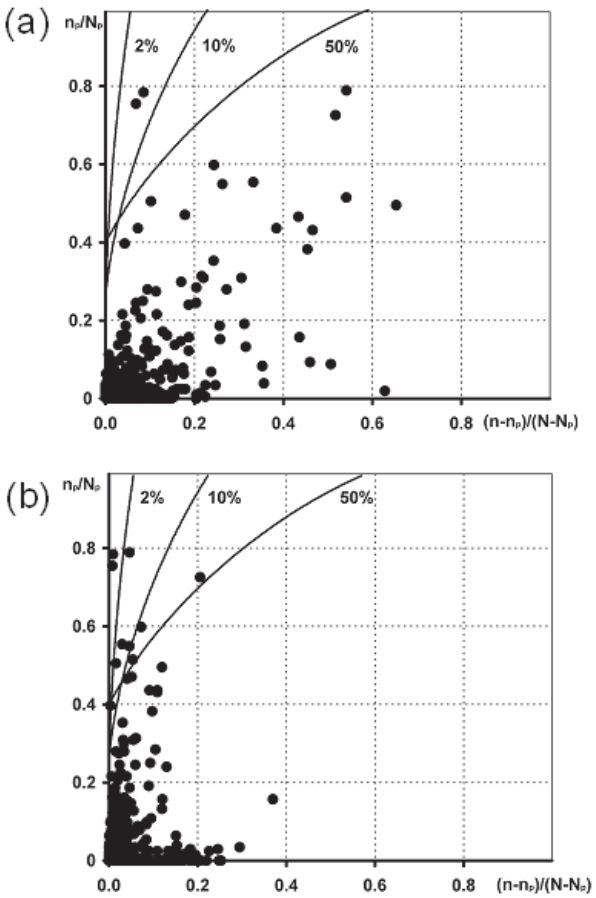

Fig. 3. Relative frequencies of species occurrences within (vertical axis) and outside (horizontal axis) the target group of 204 rock-outcrop dry grassland sites, compared with (a) 298 sites of other types of rock-outcrop dry grasslands and (b) 15 487 sites of all other types of grasslands. Each dot is a species. Curves represent the value of $\Phi = 0.5$ for cases when the size of the target group is equalized to 50%, 10% and 2% of the entire data set, respectively. Φ-values > 0.5 are above these curves,

occurrence concentration of a species within a site group) and *fidelity* (in their terms). According to the prevailing terminology of community ecology and vegetation science, however, the whole *IndVal* index (not only its part) should be considered a measure of *fidelity*. This terminology is also used throughout the present paper.

Dufrêne & Legendre (1997: 350) proposed the Indicator Value (*IndVal*) index for quantitative data (species abundances). It is defined as a product of *specificity* and relative frequency of species occurrence (*fidelity sensu* Dufrêne & Legendre 1997) within the target site group. Specificity is the mean abundance of the species in the target site group divided by its mean abundance in all site groups of the data set. The sum of mean abundances within each group is used instead of the sum of actual abundances over all groups in order to remove the effect of unequal size of the site groups. Hereafter we will call this measure $IndVal_1$ in order to distinguish it from a modification of the Indicator Value index proposed by Dufrêne & Legendre (1997: 363; see below). Although the $IndVal_1$ index was originally proposed for quantitative data, it can also be used with presence/absence data: in that case mean abundances automatically become equal to relative frequencies of species occurrence. For a comparison of two site groups based on the presence/absence data and with the notation used in this paper, $IndVal_1$ can be expressed as:

$$IndVal_1 = \frac{n_p/N_p}{\left[n_p/N_p\right]+\left[\left(n-n_p\right)/\left(N-N_p\right)\right]} \cdot \frac{n_p}{N_p}. \qquad (7)$$

This formula is identical to Eq. 16 in Chytrý et al. (2002a). In this version, it compares the target site groups with all the other sites within the data set as if they formed a single group. However, it can be easily modified for a comparison of the target site group with several site groups of the given typology by replacing the species relative frequency outside the target site group, i.e. $(n - n_p) / (N - N_p)$ in the denominator of the first fraction by the sum of relative frequencies in all the other site groups.

Like the phi coefficient applied to site groups of equalized size, $IndVal_1$ is independent of the relative size of the target site group. It implicitly equalizes the size of all site groups in the data set, including the target site group (as in Fig. 1d). The size of the target site group (N_p) is treated as equal to N divided by the number of site groups in the data set, and changing N_p to other arbitrary values does not influence the resulting numerical value. Fig. 4a shows that in its standard form, corresponding to Eq. 7, the $IndVal_1$ gives a very high weight to common species and a lower weight to the differences in relative frequency of species, a property which was already demonstrated by Chytrý et al. (2002a) in their trials with real data. Therefore the $IndVal_1$ probably does not fit most researchers' intuitive expectations of the properties of a suitable fidelity measure. However, it is possible to modify the relative importance of specificity and relative frequency of species occurrence by incorporating weighting coefficients in the Indicator Value formula (available in advanced options of the INDVAL program by M. Dufrêne; http://mrw.wallonie.be/dgrne/sibw/outils/indval/home.html).

Modified Indicator Value index: $IndVal_2$

Dufrêne & Legendre (1997: 363) proposed that with presence/absence data, Indicator Value should be calculated with a modified index. In this index (hereafter referred to as $IndVal_2$) they replaced the specificity with the fraction n_p/n, i. e., the number of occurrences of the species in the target site group divided by the number of its occurrences in the whole data set. Note that there is an error in the explanation of this index on p. 363 (second column, second line) of the Dufrêne & Legendre (1997) paper, which is corrected on the web site by M. Dufrêne (http://mrw.wallonie.be/dgrne/sibw/outils/indval/home.html): 'total number of sites in cluster *i*' should be read as 'total number of sites occupied by species *i*'. Using our notation, $IndVal_2$ is defined as:

$$Indval_2 = \frac{n_p}{n} \cdot \frac{n_p}{N_p}. \qquad (8)$$

In fact, this index is quite different from $IndVal_1$ for presence/absence data; rather it is a quite different index. Its numerical results are different from those obtained when the standard formula for $IndVal_1$ (Dufrêne & Legendre 1997: 350) is applied to presence/absence data. For example, if presence/absence data are used as input for the PC-ORD 4 program (McCune & Mefford 1999), Indicator Values are calculated according to the $IndVal_1$ formula which corresponds to Eq. 7 in the present paper. The same holds true when the INDVAL program by M. Dufrêne is run in the default mode. Advanced options of this program must be used in order to obtain results corresponding to Eq. 8.

In the context of the present paper, $IndVal_2$, unlike $IndVal_1$, is dependent on the relative size of the target site group. If the relative frequency of species occurrence (n_p/N_p) in all site groups is held constant and the relative size of the target site group (N_p/N) is decreasing, then the value of n_p/n is also decreasing, which causes the value of $IndVal_2$ to decrease too. $IndVal_2$ is therefore not suitable for measuring fidelity for site groups of unequal size, but it can be applied to the data in which the size of site groups is equalized in the same way as suggested above for the phi coefficient.

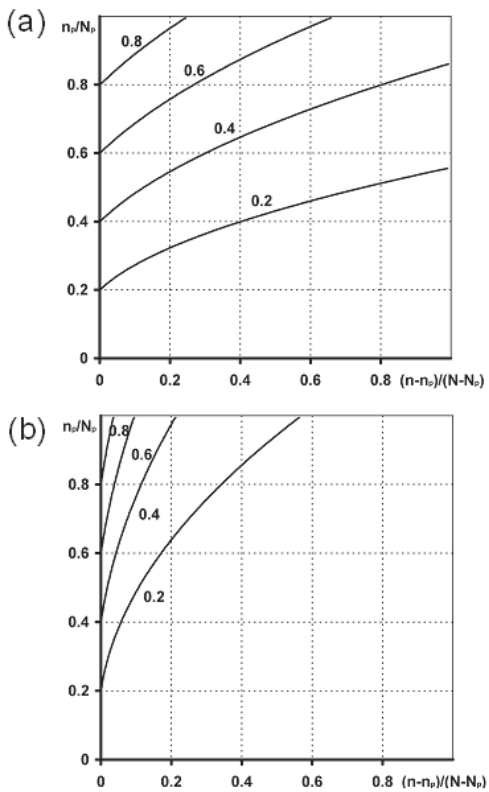

Fig. 4. Dependence of the Dufrêne-Legendre Indicator Value (*IndVal*) on the relative frequency of species occurrences within (vertical axis) and outside (horizontal axis) the target site group. **a.** $IndVal_1$, i.e. the categorical form of the equation on p. 350 in Dufrêne & Legendre (1997), corresponding to Eq. 7 in this paper; **b**. $IndVal_2$, i.e. the equation on page 363 in Dufrêne & Legendre (1997), corresponding to Eq. 8 in this paper.

Inability to distinguish positive and negative fidelity

An undesirable property of the *IndVal* indices, both of $IndVal_1$ and $IndVal_2$, is that their value is not zero (nor any other constant value) when the relative frequency of a species within the target site group is equal to the relative frequency of that species in the complementary part of the data set. For such species, the values of these indices increase with their frequency in the data set (n). Therefore neither of the two *IndVal* indices enables the distinguishing of positive fidelity (i.e., cases when relative frequency of species in the target site group is larger than in the rest of the data set) from negative fidelity (i.e. cases when relative frequency of species in the target site group is smaller than in the rest of the data set). By contrast, phi coefficient and some other statistical measures of fidelity reviewed by Chytrý et al. (2002a) do distinguish positive and negative fidelity and take on a zero value when the relative frequency is the same within and outside the target site group (Fig. 2, Table 3).

Application of the new method to real data: an example

To demonstrate the effect of equalization of the size of site groups on fidelity calculation, we computed the phi coefficients after equalization for the same data set as Chytrý et al. (2002a) used for testing performance of the phi coefficient. This data set is a part of the Czech National Phytosociological Database (Chytrý & Rafajová 2003) and includes 15 989 relevés (sites) of various grassland types. Of these, 502 relevés representing rock-outcrop dry grasslands were classified into eight vegetation units (site groups) and the phi coefficient was computed to quantify the association between each vascular plant species and each of these eight vegetation units. 15 487 relevés of other grassland types were retained in the data set while computing fidelity and became a part of the negative group of relevés, i.e. those not belonging to the target vegetation unit. On average these eight vegetation units contained 62.75 relevés, i.e. 0.395% of the entire data set, therefore each of them was equalized to that size ($s = 0.00395$). The group of other vegetation types retained the size of 96.84%, i.e. 100% – 8 × 0.395%. In order to obtain a table with 50 diagnostic species (the same number as in the corresponding Table 6 in Chytrý et al. 2002a), the numerical threshold of Φ was set to 0.222, which yielded this required number of diagnostic species. In order to exclude from the groups of diagnostic species those with non-significant fidelity (occurrence concentration in the target vegetation unit), Fisher's exact test was computed additionally to the calculation of Φ, using the actual (non-equalized) size of each vegetation unit. However, fidelity of all species with $\Phi >$ 0.222 appeared to be significantly different from random at $P < 0.001$, thus no species was excluded. Calculations were done using the JUICE 6.3 program (Tichý 2002; www.sci.muni.cz/botany/juice.htm).

The results are summarized in Table 2. Generally the groups of diagnostic species are not much different from those computed by applying the phi coefficient to vegetation units of non-equalized size (compare Chytrý et al. 2002a: Table 6). Of 50 diagnostic species contained in the tables, 41 occur in both tables. However, larger vegetation units lost some diagnostic species and smaller vegetation units got some additional diagnostic species when the phi coefficient was computed with equalized group size. Evaluated by an expert judgement, most of these changes contributed to a better characterization of the particular vegetation units. The biggest advantage of the table prepared with the equalized size of vegetation units is that for each species, the order of its relative frequencies within different vegetation units is the same as the order of its fidelities to those vegetation units.

Table 2. Synoptic table of 502 relevés of Czech rock-outcrop dry grasslands, based on fidelity comparison with 15 487 additional relevés from the Czech Republic. Fidelity was computed using the phi coefficient applied to a data set with the size of vegetation units A-H equalized to 0.395% of the size of the entire data set (i.e. to the average relative size of each of these vegetation units). Diagnostic species (grey-shaded values) are those with $\Phi > 0.222$; they are ranked by a decreasing value of Φ. Dots in part (**a**) of the table indicate species absence, dashes in part (**b**) indicate negative fidelity or positive but non-significant fidelity at $P < 0.001$. Asterisks before species names indicate species that would be among 50 species with the highest diagnostic value if the phi coefficient was calculated without equalization of the size of site groups (compare Table 6 in Chytrý et al. 2002a).

	(a) percentage frequency (constancy)								(b) phi coefficient (Φ × 1000)							
Vegetation unit	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H
No. of relevés	204	78	25	30	66	73	11	15	204	78	25	30	66	73	11	15
No. of diagnostic species	4	2	11	15	3	6	11	6	4	2	11	15	3	6	11	6
Diagnostic species of vegetation unit A																
**Asplenium septentrionale*	75	23	.	.	3	22	.	.	460	136	-	-	-	129	-	-
**Aurinia saxatilis*	78	36	.	.	9	15	.	.	449	202	-	-	-	81	-	-
**Hieracium pallidum*	40	5	4	.	8	14	.	.	358	-	-	-	65	121	-	-
**Sedum reflexum*	50	13	24	.	6	29	.	.	231	52	105	-	-	128	-	-
Diagnostic species of vegetation unit B																
**Allium montanum*	55	100	.	30	56	47	.	.	191	354	-	98	194	159	-	-
Diagnostic species of vegetation unit C																
**Hieracium echioides*	6	3	56	.	5	.	.	.	25	-	301	-	-	-	-	-
**Gagea bohemica*	.	.	24	.	.	.	.	.	-	-	271	-	-	-	-	-
**Scleranthus perennis*	13	.	64	.	.	3	.	.	47	-	262	-	-	-	-	-
**Helichrysum arenarium*	1	.	40	.	.	.	.	.	-	-	253	-	-	-	-	-
Rumex acetosella agg.	25	6	100	.	.	.	36	.	47	-	242	-	-	-	-	-
**Achillea setacea*	.	3	48	.	.	.	.	.	-	-	236	-	-	-	-	-
Poa bulbosa	2	13	44	3	3	3	.	.	-	63	235	-	-	-	-	-
Agrostis stricta	2	1	56	.	.	1	.	.	-	-	220	-	-	-	-	-
Diagnostic species of vegetation unit D																
**Teucrium montanum*	.	.	.	97	5	3	.	.	-	-	-	815	-	-	-	-
**Scorzonera austriaca*	.	.	.	73	6	1	.	.	-	-	-	671	51	-	-	-
**Fumana procumbens*	.	.	.	53	2	1	.	.	-	-	-	613	-	-	-	-
**Poa badensis*	.	4	.	80	8	4	.	.	-	-	-	587	51	-	-	-
**Minuartia setacea*	1	4	4	100	18	23	.	.	-	-	-	570	98	127	-	-
**Melica ciliata*	1	5	.	77	8	15	.	.	-	-	-	449	38	83	-	-
**Jovibarba sobolifera*	28	33	4	87	38	29	.	.	124	149	-	401	171	127	-	-
**Campanula sibirica*	.	.	.	63	5	3	.	.	-	-	-	360	-	-	-	-
**Allium flavum*	3	3	8	77	14	11	45	.	-	-	-	332	52	40	193	-
**Alyssum montanum*	7	14	8	87	24	15	18	.	16	45	-	328	84	48	-	-
**Anthericum ramosum*	8	10	8	90	56	49	.	33	-	-	-	259	156	136	-	87
**Seseli osseum*	51	62	52	100	83	82	.	.	124	151	125	255	210	207	-	-
**Dorycnium germanicum*	.	.	4	70	8	7	55	.	-	-	-	227	-	-	174	-
Diagnostic species of vegetation unit E																
**Stachys recta*	31	53	4	47	86	18	.	7	69	129	-	112	221	33	-	-
Diagnostic species of vegetation unit F																
**Saxifraga paniculata*	.	.	.	.	18	41	.	.	-	-	-	-	175	401	-	-
**Asplenium trichomanes*	16	14	.	.	17	48	18	.	105	91	-	-	109	323	-	-
**Cardaminopsis petraea*	.	.	.	.	3	11	.	.	-	-	-	-	-	272	-	-
**Vincetoxicum hirundinaria*	31	21	4	7	55	68	.	.	98	61	-	-	182	232	-	-
Diagnostic species of vegetation unit G																
**Asplenium cuneifolium*	.	.	.	.	.	.	100	.	-	-	-	-	-	-	948	-
**Thlaspi montanum*	.	.	.	.	3	.	73	.	-	-	-	-	-	-	746	-
**Biscutella laevigata*	1	3	12	3	14	19	100	.	-	-	72	-	83	119	648	-
**Stellaria holostea*	1	.	.	.	2	1	36	.	-	-	-	-	-	-	400	-
Valeriana wallrothii	.	.	.	.	2	11	27	.	-	-	-	-	-	105	267	-
**Sedum maximum*	35	58	4	.	17	19	73	.	111	190	-	-	46	55	242	-
Viola tricolor	5	9	.	.	2	5	36	.	30	53	-	-	-	-	234	-
Thymus praecox	9	8	84	27	42	27	100	67	-	-	187	48	86	49	226	145
Diagnostic species of vegetation unit H																
**Coronilla vaginalis*	.	.	.	.	.	.	.	40	-	-	-	-	-	-	-	387
**Cirsium acaule*	.	.	.	.	.	.	.	87	-	-	-	-	-	-	-	327
Gentianella ciliata	.	.	.	.	.	.	.	33	-	-	-	-	-	-	-	271
Scorzonera hispanica	.	.	.	.	.	.	.	33	-	-	-	-	-	-	-	261
Ononis spinosa	.	.	.	.	3	1	.	80	-	-	-	-	-	-	-	251

→

Table 2, cont.

	(a) percentage frequency (constancy)								(b) phi coefficient (Φ × 1000)							
Vegetation unit	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H
No. of relevés	204	78	25	30	66	73	11	15	204	78	25	30	66	73	11	15
No. of diagnostic species	4	2	11	15	3	6	11	6	4	2	11	15	3	6	11	6
Common diagnostic species of two vegetation units																
**Sedum album*	47	88	.	90	64	62	.	.	132	263	-	267	185	179	-	-
**Festuca pallens*	79	44	100	100	53	64	18	.	207	107	266	266	134	166	-	-
**Genista pilosa*	.	.	60	.	2	14	82	.	-	-	285	-	-	59	391	-
**Armeria elongata*	1	.	52	.	.	.	91	.	-	-	232	-	-	-	412	-
**Asplenium ruta-muraria*	14	18	.	7	38	47	.	.	79	106	-	-	230	285	-	-
Common diagnostic species of more than two vegetation units																
**Sesleria varia*	2	1	.	40	98	100	100	100	-	-	-	137	354	360	360	360

Notes on practical application of the new method

Using the phi coefficient for site groups of equalized size has several advantages over the other methods of statistical measuring of fidelity. First, the measure is independent of the size of the entire data set and of the size of the target site group. Therefore it does not depend on the number of sampled sites, which usually strongly varies among site groups in classification studies. Second, it is flexible in weighting the importance of common and rare species in fidelity calculation by changing the size of the target site group relative to the size of the entire data set: larger relative size gives higher weight to common species and vice versa. A disadvantage of the phi coefficient is that its numerical values vary independently of statistical significance of fidelity. Therefore this coefficient may attain high values even for those species whose occurrence concentration in the target site group is not different from random, especially if the original size of the target site group before equalization is small. Therefore it is necessary to do either a randomization test or a parallel calculation of some other fidelity measure that can be directly interpreted in terms of statistical significance. In the JUICE program (Tichý 2002), there is an option to compute the Fisher's exact test simultaneously with the phi coefficient and to exclude species with non-significant fidelity from the groups of diagnostic species. Table 3 summarizes the properties of different statistical measures of fidelity reviewed by Chytrý et al. (2002a) in comparison with the phi coefficient applied to data sets with equalized size of site groups.

The relative size to which the target site groups are equalized (s) must be selected arbitrarily. If there is some reason for giving a higher weight in fidelity calculation to common species, a size of the target group equal to 50% ($s = 0.5$) of the entire data set size seems to be a reasonable choice, because it corresponds to a comparison of the target site group with the equally sized group of sites not belonging to this site group. By contrast, if rare species and differences in relative frequency within and outside the target site group should be emphasized, which is perhaps more in accordance with the intuitive concept of diagnostic species, at least

Table 3. Comparison of some properties of different fidelity measures with properties of the phi coefficient applied to the site groups of equalized size. P/A = presence/absence. *Chi-square enables distinguishing between positive and negative fidelity if supplemented by additional information on whether the observed occurrence frequency of the species in the target site group is higher or lower, respectively, than the expected frequency.

	Input data	Dependence on the size of the entire data set	Dependence on the relative size of the site groups	Numerical value is a function of statistical significance	Weight given to common species	Enables to distinguish positive and negative fidelity
Binomial u-value u_{binB}	P/A	yes	yes	yes	low	yes
Hypergeometric u-value u_{hyp}	P/A	yes	yes	yes	average	yes
chi-square	P/A	yes	yes	yes	average	(yes)*
G statistic	P/A	yes	yes	yes	high	yes
Fisher's exact test	P/A	yes	yes	yes	high	no
phi coefficient	P/A	no	yes	no	average	yes
Dufrêne-Legendre $Indval_1$	P/A or abundances	no	no	no	very high(flexible)	no
Dufrêne-Legendre $Indval_2$	P/A	no	yes	no	high(flexible)	no
phi coefficient with equalized size of site groups	P/A	no	no	no	flexible	yes

in phytosociology, the size of the target site group can be set to the average size of all site groups present in the data set (e.g. 20% for five groups or 5% for twenty groups). This approach was followed in fidelity calculations presented in Table 2. However, the equalized size of the target site group can be arbitrarily set to any value, depending on the intended weighting of common/rare species. When determining diagnostic species for site groups in large databases, where the size of site groups is very small relative to the number of sites in the entire database, the equalized sizes of the site groups can be small fractions of the entire database size.

Equalization of the site groups proposed in this paper is meant to be used for determination of diagnostic species for site groups in the previously established classifications. As the phi coefficient may also be used for quantifying the rate of species co-occurrences in the Cocktail classification algorithm (Bruelheide 2000; Kočí et al. 2003), it is to be pointed out that the proposed equalization method should be avoided within the Cocktail algorithm, because it might produce biased estimations of the species co-occurrence rates.

Conclusions

The recently emerging large databases of phytosociological relevés (Ewald 2001; Hennekens & Schaminée 2001) and of other kinds of ecological data stimulate the development of formalized, repeatable procedures of numerical classification and subsequent parametrization of the resulting community types. Determination of diagnostic species with statistical measures of fidelity contributes to the formalization at the stage of parametrization. Applying the phi coefficient to data sets with equalized size of site groups, as proposed in this paper, seems to be a very promising approach, but the methods and procedures for measuring fidelity still need further development. In particular, it is necessary to develop more flexible methods that would consider abundance data of a quantitative or ordinal nature (Dufrêne & Legendre 1997), to design procedures that would control for different size of basic sampling units (e.g. plot size of relevés) and for resulting differences in species richness (Dengler 2003), and to explore the effect of the context in which fidelity is measured (Chytrý et al. 2002b).

Acknowledgements. We thank Zoltán Botta-Dukát, Helge Bruelheide, Marc Dufrêne and an anonymous referee for helpful comments on the previous version of this paper. This study was funded by the grants GAČR 206/05/0020 and MSM 0021622416.

References

Barkman, J.J. 1989. Fidelity and character-species, a critical evaluation. *Vegetatio* 85: 105-116.

Botta-Dukát, Z. & Borhidi, A. 1999. New objective method for calculating fidelity. Example: The Illyrian beechwoods. *Ann. Bot.* (*Roma*) 57: 73-90.

Bruelheide, H. 1995. Die Grünlandgesellschaften des Harzes und ihre Standortsbedingungen. Mit einem Beitrag zum Gliederungsprinzip auf der Basis von statistisch ermittelten Artengruppen. *Diss. Bot.* 244: 1-338.

Bruelheide, H. 2000. A new measure of fidelity and its application to defining species groups. *J. Veg. Sci.* 11: 167-178.

Chytrý, M. & Rafajová, M. 2003. Czech National Phytosociological Database: basic statistics of the available vegetation-plot data. *Preslia* 75: 1-15.

Chytrý, M., Tichý, L., Holt, J. & Botta-Dukát, Z. 2002a. Determination of diagnostic species with statistical fidelity measures. *J. Veg. Sci.* 13: 79-90.

Chytrý, M., Exner, A., Hrivnák, R., Ujházy, K., Valachovič, M. & Willner, W. 2002b. Context-dependence of diagnostic species: A case study of the Central European spruce forests. *Folia Geobot.* 37: 403-417.

Dengler, J. 2003. *Entwicklung und Bewertung neuer Ansätze in der Pflanzensoziologie unter besonderer Berücksichtigung der Vegetationsklassifikation*. Martina Galunder-Verlag, Nümbrecht, DE.

Dufrêne, M. & Legendre, P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol. Monogr.* 67: 345-366.

Ehrendorfer, F. (ed.) 1973. *Liste der Gefäßpflanzen Mitteleuropas*. 2nd ed. G. Fischer, Stuttgart, DE.

Ewald, J. 2001. Der Beitrag pflanzensoziologischer Datenbanken zur vegetationsökologischen Forschung. *Ber. R.-Tüxen-Ges.* 13: 53-69.

Hennekens, S.M. & Schaminée, J.H.J. 2001. TURBOVEG, a comprehensive data base management system for vegetation data. *J. Veg. Sci.* 12: 589-591.

Kočí, M., Chytrý, M. & Tichý, L. 2003. Formalized reproduction of an expert-based phytosociological classification: A case study of subalpine tall-forb vegetation. *J. Veg. Sci.* 14: 601-610.

McCune, B. & Mefford, M.J. 1999. *PC-ORD. Multivariate analysis of ecological data. Version 4*. MjM Software Design, Gleneden Beach, OR, US.

Sokal, R.R. & Rohlf, F.J. 1995. *Biometry*. 3rd ed. W.H. Freeman and Company, New York, NY, US.

Szafer, W. & Pawłowski, B. 1927. Die Pflanzenassoziationen des Tatra-Gebirges. Bemerkungen über die angewandte Arbeitstechnik. *Bull. Int. Acad. Pol. Sci. Lett.* B 3, Suppl. 2: 1-12.

Tichý, L. 2002. JUICE, software for vegetation classification. *J. Veg. Sci.* 13: 451-453.

Whittaker, R.H. 1962. Classification of natural communities. *Bot. Rev.* 28: 1-239.

Received 20 October 2005;
Accepted 23 July 2006;
Co-ordinating Editor: H. Bruelheide.