Ecological Indicators xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original article

# Detecting indicator species: Some extensions of the IndVal measure

János Podani[a,*], Béla Csányi[b]

[a] Department of Plant Taxonomy and Ecology, Institute of Biology, Eötvös University, Pazmany P. S. 1, Budapest, Hungary
[b] VITUKI, Environmental Protection and Water Management Research Institute, Budapest, Hungary

ARTICLE INFO

*Article history:*
Received 30 June 2009
Received in revised form 10 March 2010
Accepted 19 March 2010

*Keywords:*
Classification
Fidelity
Indicator species
Specificity

ABSTRACT

The indicator value (IndVal) of a species has long been the most popular measure to express species importance in community classifications. Nevertheless, a few problems concerning the original definition of IndVal still require clarification and some modifications are also in order so as to exploit the capabilities of the method more fully. In particular, we propose novel component terms (specificity, concentration and fidelity) that may be incorporated in the calculation of IndVal and also suggest some minor, although important terminological amendments. We argue that the choice among these terms should largely depend on whether the target classification is based on abundance or presence–absence data. The expanded capabilities of the approach and the sensitivity of IndVal variants to the sharpness of classifications are illustrated by actual examples coming from a benthic macroinvertebrate survey along the Danube River and a study of dolomite grassland communities in the Buda Hills, Hungary. We found that analyses by the original IndVal plus the new variants may give a more complete picture on any classification than a particular selection among component terms of IndVal. The use of several indices simultaneously is particularly recommended when selection of indicator species is the primary objective of the study.

## 1. Introduction

The detection of species that best characterize some set of sites (sampling units, stands, relevés) is an important step in evaluating classifications in community ecology. In addition to purely descriptive studies, in which measuring the explanatory power of species may be the only goal, the application of an appropriate procedure is essential in biodiversity surveys, conservation biology and environmental sciences whenever priority must be given to species that best reflect environmental quality (McGeoch and Chown, 1998). The most widely used approach for finding indicator species is due to Dufrêne and Legendre (1997) who proposed a composite index called IndVal (indicator value). Their paper has been highly cited (over 1090 in January, 2010; a figure not including technical reports, books, theses, and articles from journals without impact factor). Not surprisingly, therefore, the taxonomic coverage of study objects in published applications of the IndVal methodology is very wide, ranging from bacteria (Shawkey et al., 2009) and diatoms (Potapova and Charles, 2007) through forest plants (Bataineh et al., 2007) to birds (Mikusinski et al., 2001). The high demand for a relatively simple and easily interpretable coefficient is thus obvious. However, we argue that a few issues concerning the original definition of IndVal require clarification and some modifications are also in order to exploit the capabilities of the method more fully. We first give a short account of the original definition of the index, and then propose novel component terms of IndVal together with minor, albeit essential terminological amendments. The expanded capabilities of the indicator value method will be illustrated by actual examples coming from a benthic macroinvertebrate survey along the Danube River and a study of rock grassland communities in the Buda Hills, Hungary.

## 2. The basic definition of IndVal

Let the abundance data be presented in data matrix $\mathbf{X} = \{x_{ik}\}$, with species in rows ($i = 1, \ldots, p$) and sites in columns ($k = 1, \ldots, m$). The presence/absence data matrix corresponding to $\mathbf{X}$ is denoted by $\mathbf{Y}$ in which $y_{ik} = 1$ if $x_{ik} > 0$ and $y_{ik} = 0$ otherwise. We assume that each species occurs in at least one site, and that every site has at least one species in it. Let the number of groups (or clusters) in the classification (site typology) be $g > 1$, and the number of sites in the classes be specified by vector $\mathbf{n} = \{n_h\}$, $h = 1, \ldots, g$.

The value of IndVal$_{ij}$, as originally suggested, is the product of two terms, the first referring to the performance of species $i$ in terms of abundance over all groups and the other referring to the performance of the same species in terms of presence–absence within site group $j$. The first part, termed by Dufrêne and Legendre (1997)

[*] Corresponding author.
*E-mail addresses:* podani@ludens.elte.hu (J. Podani), bela.csanyi@gmail.com (B. Csányi).

1470-160X/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ecolind.2010.03.010

as *specificity* is obtained as

$$A_{ij} = \frac{\sum_{k\in j} x_{ik}/n_j}{\sum_{h=1}^{g}\sum_{k\in h} x_{ik}/n_h} = \frac{\bar{x}_{ij}}{\sum_{h=1}^{g}\bar{x}_{ih}} \tag{1}$$

which is the ratio of the mean abundance of species *i* in site group *j* and the sum of means of the same species over all groups. It is maximum, i.e., $A_{ij} = 1$, if the species appears only in group *j*, no matter how abundant it is. $A_{ij}$ is zero for species entirely absent from cluster *j*. The second part of the index is another ratio, called *fidelity*, and is given by the following formula:

$$B_{ij} = \sum_{k\in j}\frac{y_{ik}}{n_j} = \bar{y}_{ij} \tag{2}$$

which is the proportion of sites in which species *i* is present within group *j*. Its range is also [0,1], the minimum obtained when the species is absent from group *j*, and the maximum resulting when the species occurs in every site in that cluster. These two terms are multiplied and then scaled to 100 to express the indicator value of species *i* with respect to cluster *j* in terms of percentages:

$$\mathrm{IndVal}_{ij} = 100A_{ij}B_{ij} \tag{3}$$

Finally, for each species the largest value is found in order to express its indicator value with respect to the given typology, that is

$$\mathrm{IndVal}_{i} = \max_j\ \{\mathrm{IndVal}_{ij}\}. \tag{4}$$

In the present paper, we shall consider formula (3) in a general framework so that *A* and *B* can have different meanings from those defined above, thus providing opportunities to express the overall and local behavior of the species in alternative, albeit equally if not more meaningful ways. In particular, we show that *A* can be redefined such that its value will also be influenced by the number of clusters. We suggest to distinguish between two ways in which the distribution of mean abundances over the clusters is handled, namely specificity and concentration. Furthermore, while the original formula of IndVal is a composite measure using both abundances and presence–absence data, we argue that indicator values are more logically defined based on either data type exclusively, thus making the analysis fully compatible with the numerical classification method by which the typology was derived.

## 3. Materials and methods

### 3.1. Material

The analysed *benthic macroinvertebrate* data set originates from the sampling program of the Second Joint Danube Survey in 2007. A total of 47 locations out of 74 Danubian sites were selected which are not exposed to high levels of anthropogenic impacts. Samples were collected from the shallow bank zone of the Danube using the “kick and sweep” method (Armitage et al., 1983), between 2415 and 865 river km, and dredging at sites between 865 river km and the confluence of the Black Sea (0 rkm), respectively. Dredging was the only accessible sampling method, due to the coincidental flooding and elevated water level after the 2nd Iron Gate reservoir. The material was divided according to major taxonomic groups for further identification by experts to the best level possible. In the illustrative analysis presented here, only molluscs, malacostracans, and insects were included, oligochaetes and chironomids were not taken into consideration. Rare species (with no more than 1–2 occurrences) were removed from the data set because they do not influence IndVal analysis anyway, so that the total number of species retained was 44.

The second actual example derives from a detailed survey of *dolomite grassland communities* in Sas-hegy (Buda Hills, Budapest, Hungary). Eighty 4 m × 4 m quadrats were located in relatively undisturbed parts of the vegetation, and percentage cover of vascular plants was recorded for each species. A total of 123 species were found, all of them retained in subsequent analyses. A more detailed description of the study sites and results obtained by different multivariate methods are available (see Podani and Miklós, 2002, and references therein).

### 3.2. Specificity vs. concentration for abundances

First, we start with a minor criticism of the term *specificity*, as coined to formula (1) by Dufrêne and Legendre (1997). The problem is that *A* does not depend on the number of groups in the classification. If species *i* appears, say, only in two clusters, then its specificity to any of these two groups remains the same value no matter how many additional clusters devoid of species *i* are in the collection, because the sum in the denominator is unaffected. The intuitive meaning of specificity, however, would imply that the higher the number of clusters from which a species is absent the more specific it is to a cluster in which it does occur. Indeed, *A* as defined in Eq. (1) is rather a measure of *concentration* of abundances of species *i* into cluster *j* (McCune and Grace, 2002). Thus, in our interpretation, its maximum indicates that all abundances are concentrated into a single cluster, 0 indicates absence and thus zero concentration in that cluster, and intermediate values depend on how much of the total abundance of the species *i* is confined to group *j*.

We suggest that a true specificity measure should be sensitive to the distribution of mean abundances over all clusters in the classification. Therefore, we look for an expression which takes unity as above (i.e., the species occurs in one group only, so it is *both concentrated* and *specific to that cluster*), and zero when the mean abundances are the same over all groups. Specificity as understood here is related to the evenness of means in the *other clusters* of the typology and the deviation of the mean in the given cluster from the mean of all other means. We derived the following formula:

$$A^{E}_{ij} = \frac{\bar{x}_{ij} - \bar{x}_{i-j}}{\max_h\{\bar{x}_{ih}\}}, \qquad \bar{x}_{i-j} = \sum_{h\neq j}^{g}\frac{\bar{x}_{ih}}{g-1} \tag{5}$$

where $\bar{x}_{ij}$ and *g* are as defined in Eq. (1), and $\bar{x}_{i-j}$ is the mean of mean abundances of species *i* over all groups except group *j*. The numerator is the amount that the mean abundance of species *i* in group *j* should change in order to reach perfect evenness, and the denominator is the maximum of means over all groups which is used to standardize the above difference. If the species appears only in cluster *j*, and is absent from all other groups, the value of $A^{E}_{ij}$ will be 1. Eq. (5) returns zero when all means are identical. The extra property of this definition, which is fully compatible with the theory of bioindication, is that function (5) returns the value of −1 when species *i* is *absent* from group *j* while its means take identical nonzero values in all other groups. This is in agreement with Juhász-Nagy (1964) who suggested that a species can be totally indicative of a group of sites not only if it is present just there, but also if it is absent only from that group and is equally abundant everywhere else. That is, negative indication (i.e., absence) is also an indication and can be interpreted ecologically/biologically just as well as the positive indication (presence). Dufrêne and Legendre (1997) also recognized this by raising the possibility of an unsymmetric version of *A*, obtained as the ratio of the difference between the maximum mean abundance and the mean abundance in the given group to the sum of such differences over all clusters; they called it the “relative mean nonrealized abundance”. Our symmetric measures(s) represent this term in the revision.

To clarify the issue further, we compare the behavior of $A_{ij}$ and $A^{E}_{ij}$ using a series of simple artificial examples. Fig. 1 shows differ-

Please cite this article in press as: Podani, J., Csányi, B., Detecting indicator species: Some extensions of the IndVal measure. Ecol. Indicat. (2010), doi:10.1016/j.ecolind.2010.03.010

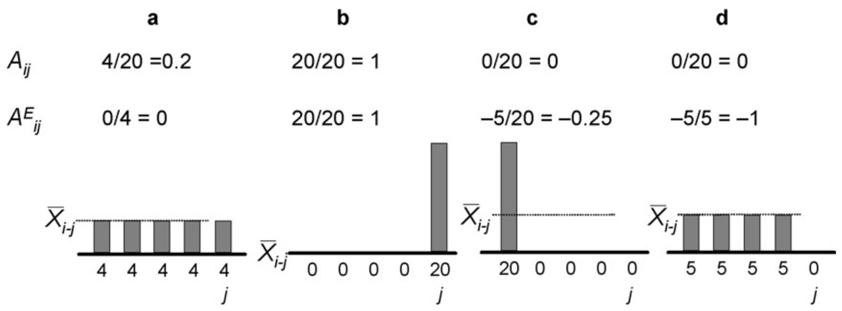

**Fig. 1.** Artificial examples for the comparison of concentration ($A_{ij}$) and specificity ($A^E_{ij}$) values calculated for different distributions with the same total. (a) The mean abundances are equal for all groups, (b) the species appears only in group *j*, (c) the species appears only in one group other than *j*, and (d) the species is absent from group *j* and has equal mean abundances everywhere else. Numbers under the axes show mean abundances for the five groups in each case, the dotted horizontal lines are the means of group means excluding cluster *j*.

ent hypothetical distributions of the mean abundances of a species for five clusters and includes the values of the two variants of *A*. The two functions gave identical results only in case b (=full concentration, =maximum specificity), whereas they differ slightly for perfectly even mean abundances (a). In the latter case, $A_{ij} = 0.2$ which is counter-intuitive and demonstrates that this is not specificity indeed, but concentration. Cases c and d are treated as being identical by $A_{ij}$, although the distributions are very different. The species is most specific to (i.e., entirely absent from) group *j* in case d (−1), and only slightly specific thanks to its absence in cluster *j* and three others in c (−0.25).

### 3.3. *Abundance vs. presence/absence*

As suggested by Dufrêne and Legendre (1997), the original formulation of IndVal (i.e., Eqs. (1) and (2) combined in Eq. (3)) is useful for typologies “obtained by any hierarchical or non-hierarchical classification procedure; its use is independent of the classification method”. Whereas independence in the statistical sense is true, there are two difficulties: the first practical the second theoretical.

The practical problem is that very often presence/absence data are available only, so that the original formula loses its power. It is because the “specificity” value (Eq. (1)), if applied to presence/absence data, will become the *fidelity* value (Eq. (2)) as divided by the sum of fidelity values for all groups, so that the two parts of IndVal will express practically the same thing. As a remedy, Dufrêne and Legendre (1997) suggested to replace $A_{ij}$ by the following formula:

$$A^P_{ij} = \frac{\sum_{k\in j} y_{ik}}{\sum_{k=1}^{m} y_{ik}}, \qquad (6)$$

which is the number of sites in group *j* in which species *i* is present divided by the number of all sites where this species is present. However, as with $A_{ij}$, this is not influenced by the total number of groups in the classification and thus $A^P_{ij}$ does not measure how specific species *i* is in selecting among the groups. Indeed, it is a measure of *concentration* because the more presences are within group *j* as compared to the entire set of sites, the more concentrated is its presence into that group. True *specificity* should depend on the number of groups, *g*, and should increase when the number of clusters from which the species is absent increases. There are several possibilities to define specificity for p/a data which satisfy this requirement; we propose the presence/absence version of Eq. (5):

$$A^S_{ij} = \frac{\bar{y}_{ij} - \bar{y}_{i-j}}{\max_h\{\bar{y}_{ih}\}}, \qquad (7)$$

in which $\bar{y}_{ij}$ is defined by Eq. (2), and $\bar{y}_{i-j}$ is the mean of means for all clusters except *j*. Its minima and maxima can be interpreted in the same way as for Eq. (5). In particular, its value is 1 if species *i* is confined to this group and −1 when the species is absent here and is present with the same nonzero average in all other groups.

The theoretical problem with the original formula of IndVal is that one component uses presence/absence information while the other rests on abundances. This is somewhat incompatible with the classification procedure which is *either* presence/absence based *or* abundance based, as mainly determined by the nature of the dissimilarity coefficient used in cluster analysis (forget now about subjective classifications in which “data type” is undefined). That is, the evaluation part is logically separated from the analytical part—which may not be desirable for a simple reason: one cannot expect that finding indicator species is *always* optimal by the original, “mixed” measure while the clustering method itself utilizes only one type of information. For the presence/absence case, the problem has been solved as shown above; both *A* and *B* can be expressed in terms of binary data. But then, how to express fidelity for abundance data while (1) we have a choice between concentration and specificity to express overall behavior by *A*, and (2) formula (2) defines fidelity in terms of p/a data only? Now, we feel that fidelity within a cluster in the p/a case is analogous to evenness in case of abundances: the more even the distribution of abundances, the higher its fidelity and in turn its indicative power. Thus, for species *i* within group *j* we suggest the use of the following measure:

$$B^E_{ij} = 1 - \frac{0.5\sum_{k\in j}\left|x_{ik} - \bar{x}_{ij}\right|}{\sum_{h\in j} x_{ih}} = 1 - 0.5\sum_{k\in j}\left|\frac{x_{ik}}{\sum_{h\in j} x_{ih}} - \frac{1}{n_j}\right| \qquad (8)$$

(evenness, see Podani, 2006). It has the maximum value of 1 if the abundances are equal for all sites (perfect evenness) within the group. For total absence, of course, we cannot speak of evenness at all so that the second term is defined to be 1 and thus fidelity is zero. Fig. 2 facilitates understanding the behavior of this equation

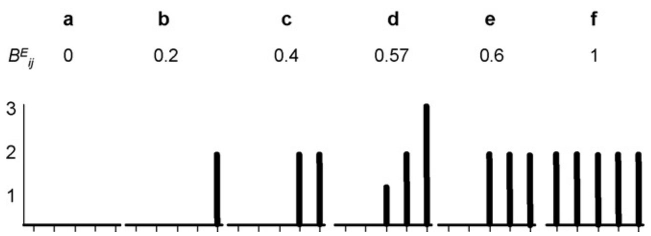

**Fig. 2.** Artificial examples illustrating fidelity ($B^E_{ij}$, Eq. (8)) of a species in a group of five sites (horizontal axis), based on abundance data (vertical axis). (a) The species is absent from the group, (b) the species is present in one site, (c) the species is present in two sites with equal abundances, (d) the species is present in three sites with unequal abundances, (e) the species is present in three sites with equal abundances, and (f) the species is present in all sites with equal abundances.

Please cite this article in press as: Podani, J., Csányi, B., Detecting indicator species: Some extensions of the IndVal measure. Ecol. Indicat. (2010), doi:10.1016/j.ecolind.2010.03.010

**Table 1**
Combinations of *A* and *B* in calculating IndVal. The original formulation (APCF) and PCF were proposed by Dufrêne and Legendre (1997), the other three are newly suggested here.

| | *A*—between groups | *B*—within group | Abbreviation |
|---|---|---|---|
| Abundance and presence/absence | Concentration (Eq. (1)) × | Fidelity (Eq. (2)) | APCF |
| Abundance | Concentration (Eq. (1)) ×<br>Specificity (Eq. (5)) × | Fidelity (Eq. (8)) | ACF<br>ASF |
| Presence/absence | Concentration (Eq. (6)) ×<br>Specificity (Eq. (7)) × | Fidelity (Eq. (2)) | PCF<br>PSF |

in other situations as well. For presence/absence data, functions (2) and (8) return the same value, showing the close relationship between the notions of evenness and fidelity.

### 3.4. *Sharpness of classifications*

Indicator values do not merely reflect species importance but also measure how the species support a given classification. Obviously, the larger the support, the sharper is the classification. We suggest therefore that the sum of all indicator values be used to characterize the overall sharpness of the typology of sites:

$$S = \sum_{i=1}^{p} |\mathrm{IndVal}_i|. \tag{9}$$

This is different from Dufrêne and Legendre's (1997) proposition which implies calculating the total of *significant* indicator values only (but see next subsection, for randomization tests suggested here). Eqs. (5) and (7) can provide negative results; this is why the absolute values are summed up in Eq. (9). The maximum of *S* is 100*p*, obtained in the perfect situation when every species characterizes unequivocally a single group of sites. The *S* values can be useful for several purposes:

1. to compare the sensitiveness of the different versions of *A* and *B* in measuring the sharpness of the same classification, and
2. to evaluate the relative merits of different classification methods at a fixed *g*.

Yet another possibility is to find the species with the highest contribution to *S*, that is the value:

$$T = \max_i |\mathrm{IndVal}_i| \tag{10}$$

is useful for characterizing the typology in terms of species with the highest indicative power.

### 3.5. *Randomization*

The evaluation of the performance of different versions of $\mathrm{IndVal}_i$ as applied to a given classification and the comparison of different classifications assume that the *S* values are comparable. However, although the theoretical range of each variant is the same (i.e., 0–100), the different formulations do not utilize this range in the same manner: a given value may be relatively low for one coefficient and relatively high for another. The underlying distributions for the indices are likely to be different, so that randomization is necessary to facilitate a meaningful comparison. We suggest that randomization is required especially for the *S* statistic, and propose the use of the following, well-known method of standardization:

$$S' = \frac{S - \bar{S}}{\mathrm{std}(S)} \tag{11}$$

in which $\bar{S}$ is the mean calculated from all simulated values obtained by a large number of randomizations (we used 99) plus the actual value, and std(*S*) is the standard deviation of the these (99 + 1) values. The larger the newly obtained *S*′ the more significant is its deviation from random expectation (i.e., zero). The proposed randomization method involves reassigning each site randomly to the classification, by retaining the number of clusters and cluster sizes of the original typology.

### 3.6. *Classifications*

Three different classifications of the Danubian benthic sample sites into three groups were evaluated to illustrate the extended possibilities of calculating IndVal. The first typology is based on the subjective segmentation of the Danube River into three sections according to the field experience of the second author. Classification 2 is based on the 3-cluster level solution in the Jaccard index/UPGMA dendrogram obtained from presence/absence data. Classification 3 was derived by cutting at the 3-cluster level the Similarity ratio/UPGMA dendrogram computed from abundance data (see Legendre and Legendre, 1998; Podani, 2000; for methodological details). The grassland quadrats were also classified using the Jaccard index and Similarity ratio/UPGMA based on presence–absence and percentage cover, respectively, and the dendrograms were pruned to three clusters in both cases, which is in complete agreement with previous results and the field experience of the first author (Podani, 1998).

The original formula of IndVal and the four recommended combinations of *A* and *B* (for abbreviations used here, see Table 1) were calculated for each classification. Cluster analyses were performed by the SYN-TAX 2000 package (Podani, 2001), whereas the IndVal scores were calculated and standardized using a new FORTRAN routine INDVALCOM available from the first author's web site (http://ramet.elte.hu/~podani).

**Table 2**
Sharpness of three classifications of the macroinvertebrate assemblages into three clusters, as measured by the *S* statistic and its randomization. Maxima of *S*′ are in bold for columns and in italics for rows.

| IndVal | Subjective classification | | | Jaccard/UPGMA | | | Similarity ratio/UPGMA | | |
|---|---|---|---|---|---|---|---|---|---|
| | *S* | Mean | *S*′ | *S* | Mean | *S*′ | *S* | Mean | *S*′ |
| APCF | 1918 | 1002 | *5.97* | 1871 | 1008 | 5.77 | 1673 | 1038 | 4.54 |
| ACF | 1245 | 700 | *5.32* | 1232 | 705 | 5.26 | 1157 | 731 | 4.45 |
| ASF | 1327 | 740 | *5.08* | 1317 | 746 | 5.07 | 1232 | 784 | 4.26 |
| PCF | 1615 | 906 | 5.59 | 1634 | 918 | *5.61* | 1434 | 997 | 3.66 |
| PSF | 1512 | 609 | ***6.71*** | 1547 | 627 | **6.58** | 1301 | 667 | **4.90** |

Please cite this article in press as: Podani, J., Csányi, B., Detecting indicator species: Some extensions of the IndVal measure. Ecol. Indicat. (2010), doi:10.1016/j.ecolind.2010.03.010

**Table 3**
Sharpness of two classifications of the grassland quadrats into three clusters, as evaluated by the $S$ statistic and its randomization. Maxima of $S'$ are in bold for columns and in italics for rows.

| IndVal | Jaccard/UPGMA | | | Similarity ratio/UPGMA | | |
|---|---|---|---|---|---|---|
| | $S$ | Mean | $S'$ | $S$ | Mean | $S'$ |
| APCF | 3892 | 2228 | *5.22* | 3264 | 2248 | 3.53 |
| ACF | 3468 | 2003 | *5.17* | 2933 | 2017 | 3.56 |
| ASF | 3526 | 1677 | *5.96* | 2845 | 1689 | **4.35** |
| PCF | 3906 | 2511 | *4.62* | 3284 | 2724 | 1.91 |
| PSF | 3515 | 1470 | ***6.41*** | 2546 | 1472 | 4.21 |

## 4. Results and their discussion

### 4.1. Comparative evaluation of indices

The sharpness scores ($S$), the simulated means ($\bar{S}$) and the standardized values ($S'$) calculated for the three classifications of macroinvertebrate assemblages strongly differ with variants of the indicator value (Table 2). The means are the highest for the original IndVal (APCF), followed by PCF, and then by ACF and ASF. The lowest averages are provided by the combination of specificity and fidelity for presence/absence data (PSF). The rank order of simulated mean values for the two grassland classifications is similar, albeit with interchanges in positions 1–2 and 3–4 (Table 3). These findings indicate fairly consistent behavior of different variants of IndVal and underlie the importance of randomization in comparative studies. Thus, only the standardized $S$ scores are interpreted, because these are fully compatible and comparable by rows as well as by columns. The deviation of these standardized values from zero ranges from 3.66 to 6.71 (benthic macroinvertebrates) and from 1.91 to 6.41 std units (grassland), indicating significant group structure compared to the randomized classifications.

Comparison of *typologies* according to each of the five IndVal versions is done first to test how the indices perform under different circumstances (values in italics are row maxima in Tables 2 and 3). The subjective classification of macroinvertebrate assemblages is found to be the sharpest by four of the five indices, that is, APCF, ACF, ASF and PSF, but differences are sometimes very small (Table 2). The only IndVal version for which the Jaccard/UPGMA classification is the best is PCF, whereas the classification obtained by the Similarity ratio/UPGMA is not found to be the sharpest by any of these coefficients. In fact, the standardized $S$ scores are consistently lower for this classification than for the other two, indicating weaker support by species and thus a less clear group structure in the abundance data space. The results are similar in several respects for the grassland data (Table 3). The presence–absence based classification proved to be sharper than the abundance based typology for all IndVal variants.

Finding the column maxima reveals that all the classifications of macroinvertebrate data (Table 2) and the Jaccard/UPGMA classification of grassland quadrats (Table 3) were shown to be sharpest by PSF, suggesting that the combination of Eqs. (2) and (8) is the most sensitive indicator of group sharpness for, interestingly enough, all types of classifications. The Similarity ratio/UPGMA classification of the plant cover data is the only exception: for this ASF yielded the highest score, i.e., an abundance-based measure as expected.

### 4.2. Most indicative species

The highest $T$ values and the associated species are listed in Table 4 for the macroinvertebrate assemblages. A most remarkable observation is that the presence of *Dreissena bugensis*, an invasive species, is a very strong ecological signal for the lower section of the Danube, as well as its full absence from the Middle and Upper Danube. It was chosen as the most indicative species by two p/a variants of IndVal for all the three classifications (PCF, PSF, last two rows in Table 4) and by the original IndVal for the Jaccard/UPGMA case. In terms of abundance, *Corbicula fluminea* and *Unio tumidus* appear to be the best indicators along the Danube River. *C. fluminea*, another invasive species, is extremely abundant in the lowest section of the Danube (100–1000 individuals per sampling unit), whereas its abundance rarely exceeds 100 elsewhere. *U. tumidus* reaches 30–60 individuals per sampling unit in the Middle Danube, it is almost entirely absent from the upper part, and is only occasional on the lowest stretch. A fourth species, *Potamopyrgus antipodarum* was detected as the "best" species only once, by the original version of IndVal. It is an invasive hydrobiid snail introduced from the Rhine system, occurring principally in the Upper Danube in the study area. In conclusion, in the typification of the Danubian macroinvertebrate assemblages invasive species play a principal role, which seems to be a temporary situation and may change considerably within a short time depending on how fast these species invade the river in its full length.

**Table 4**
Species with the highest indicator values for three 3-cluster classifications of macroinvertebrate assemblages, expressed as $T$ values derived from five variants of IndVal.

| IndVal | Subjective classification | Jaccard/UPGMA | Similarity ratio/UPGMA |
|---|---|---|---|
| APCF | Potant 88 | Drebug 89 | Corfla 88 |
| ACF | Unitum 64 | Unitum 65 | Corfla 60 |
| ASF | Unitum 67 | Unitum 68 | Corfla 64 |
| PCF | Drebug 83 | Drebug 89 | Drebug 60 |
| PSF | Drebug 86 | Drebug 89 | Drebug 72 |

*Abbreviations*: Corfla: *Corbicula fluminea*; Drebug: *Dreissena bugensis*; Potant: *Potamopyrgus antipodarum*; Unitum: *Unio tumidus*.

**Table 5**
Species with the highest indicator values for two 3-cluster classifications of grassland data, as evaluated by five variants of IndVal and expressed as $T$ values.

| IndVal | Jaccard/UPGMA | Similarity ratio/UPGMA |
|---|---|---|
| APCF | Bupfal 82 | SesSad 97 |
| ACF | Bupfal 82 | SesSad 67 |
| ASF | Dralas 83 | SesSad 68 |
| PCF | Bupfal 82 | Fumpro 69 |
| PSF | Dralas 83 | SesSad 93 |

*Abbreviations*: Bupfal: *Bupleurum falcatum*; Dralas: *Draba lasiocarpa*; Fumpro: *Fumana procumbens*; Sessad: *Sesleria sadleriana.*

In the rock grassland classification, species characteristic of the closed stands have the highest indicator values, irrespective of the data type (Table 5). For abundances, *Sesleria sadleriana*, a very highly dominant grass (when present, it has a cover of more than 50% in a quadrat) in closed grassland stands has the strongest signal in four cases, and only once it is replaced by *Fumana procumbens*. For presence–absence, *S. sadleriana* loses its discriminatory power and *Bupleurum falcatum* and *Draba lasiocarpa*, two other highly associated species of closed grasslands become the most important. Whereas in the macroinvertebrate survey mostly invasive species were detected as being the most indicative, in the rock grassland all the four species mentioned are native, two of them (*S. sadleriana* and *D. lasiocarpa*) with distribution restricted to the Carpathian basin and the Balkans.

## 5. Concluding remarks

The present paper introduced new alternatives for measuring the indicator value of species in community classification. We showed that the results depend on how the component terms of IndVal are chosen, and pointed out that the originally suggested formulae (APCF and PCF) are not always the most sensitive ones. Although the differences among the five possibilities discussed are usually not enormous, we suggest that the choice of the IndVal coefficient should be governed primarily by the data type used. Then, we can decide whether the fidelity measure should be multiplied by concentration or specificity (Table 1). The choice between the latter two is facilitated by Fig. 1 which illustrates that concentration focuses mainly on the accumulation of nonzero data within vs. between clusters. Specificity, as understood here, is a symmetric measure because it responds equally to the accumulation of nonzero and zero values within a cluster.

IndVal versions theoretically less suited to a given data type may nevertheless provide high indicator values, and therefore high sharpness scores, which is probably a reflection of the high similarity between classifications derived from different data types and of strong group structure in the data. We feel, therefore, that analyses by the original IndVal plus the new variants may give a more complete picture on any classification than a particular selection among component terms of IndVal. Nevertheless, one of the five possibilities examined appears to have a more restricted utility: PCF originally suggested by Dufrêne and Legendre (1997) for the presence–absence case almost always produced much lower $S'$ and $T$ values than PSF (last two rows in Tables 2 and 3), and its performance was particularly poor when applied to the “wrong” data type. In fact, PSF was in many cases a more sensitive indicator of sharpness than the other four versions, even for abundance data. ACF and ASF were more case dependent in this study: for the macroinvertebrates ACF gave consistently higher scores than ASF, which was in the other way around for the grassland classifications. As obvious from our results, the original IndVal combination (abbreviated here as APCF) did perform relatively well. Notwithstanding its mixed nature, it seems to be a fairly useful combination of a concentration measure for abundances and a fidelity measure for presences. This is illustrated well with the $T$ values for the macroinvertebrate assemblages: for the presence/absence based classification it has selected the same species as PCF and PSF, whereas for abundances its result was identical to that of ACF and ASF (Table 4).

To our knowledge, our paper represents the most detailed study of the properties of IndVal coefficients since the original propositions were made by Dufrêne and Legendre (1997). Except for a few minor amendments proposed in a web site (Dufrêne, 2004), no new variants were published either. Based on two actual ecological data sets, we were able to illuminate the relative merits of different variants of IndVal and to give some recommendations for future use of these coefficients. We believe that the extended methodology of indicator value calculation provides a widely applicable tool for ecological applications. A more detailed survey, including the study of suboptimal species combinations based on a wider variety of field data is required to obtain a more general picture on the performance of the different IndVal coefficients. Further possibilities include the examination of the utility of IndVal variants in the detection of the optimum number of clusters and to find the resemblance coefficient and clustering method whose combination yields the sharpest classification of a given collection of objects. Another possibility for future research is to develop a classification method which involves direct optimization of IndVal, rather than its use as an *a posteriori* evaluator of a typology.

## Acknowledgements

We are grateful to the anonymous referees for their constructive comments on previous drafts of the manuscript. J.P.'s research was supported by Hungarian Research Grant OTKA No. NI68218. B.Cs. is most grateful to the International Commission for the Protection of the Danube River (ICPDR) for organizing the Second Joint Danube Survey in 2007.

## References

Armitage, P.D., Moss, D., Wright, J.F., Furse, M.T., 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Res. 17, 333–347.

Bataineh, M.M., Oswald, B.P., Bataineh, A.L., Farrish, K.W., Coble, D.W., Edminster, C.B., 2007. Plant communities associated with *Pinus ponderosa* forests in the Sky Islands of the Davis Mountains, Texas. J. Torrey Bot. Soc. 134, 468–478.

Dufrêne, M., 2004. IndVal or How to Identify Indicator Species of a Sample Typology?, http://biodiversite.wallonie.be/outils/indval/home.html.

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

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

Legendre, P., Legendre, L., 1998. Numerical Ecology, 2nd ed. Elsevier, Amsterdam.

McCune, B., Grace, J.B., 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, OR.

McGeoch, M.A., Chown, S.L., 1998. Scaling up the value of bioindicators. Trends Ecol. Evol. 13, 46–47.

Mikusinski, G., Gromadzki, M., Chylarecki, P., 2001. Woodpeckers as indicators of forest bird diversity. Conserv. Biol. 15, 208–217.

Podani, J., 1998. A complex numerical analysis of dolomite rock grasslands of the Sas-hegy Nature Reserve, Budapest, Hungary. In: Csontos, P. (Ed.), Synbotanical Exploration of Rock Grasslands. Scientia, Budapest, pp. 213–229 (in Hungarian with English abstract).

Podani, J., 2000. Introduction to the Exploration of Multivariate Biological Data. Backhuys, Leiden.

Podani, J., 2001. SYN-TAX 2000. Computer programs for data analysis in ecology and systematics. In: User's Manual. Scientia, Budapest.

Podani, J., 2006. With a machete through the jungle: some thoughts on community diversity. Acta Biotheor. 54, 125–131.

Podani, J., Miklós, I., 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. Ecology 83, 3331–3343.

Potapova, M., Charles, D.F., 2007. Diatom metrics for monitoring eutrophication in rivers of the United States. Ecol. Indic. 7, 48–70.

Shawkey, M.D., Firestone, M.K., Brodie, E.L., Beissinger, S.R., 2009. Avian incubation inhibits growth and diversification of bacterial assemblages on eggs. PLoS ONE 4 (2), e4522, doi:10.1371/journal.pone.0004522.

Please cite this article in press as: Podani, J., Csányi, B., Detecting indicator species: Some extensions of the IndVal measure. Ecol. Indicat. (2010), doi:10.1016/j.ecolind.2010.03.010