Indicator Species: Computation

Pierre Legendre, Université de Montréal, Montréal, QC, Canada

Glossary

Fidelity The degree to which a species is present at all sites of a group.

Indicator species A biological species that has a high indicator value for a group of sites. In the Indictor value method, species j is an indicator of group of sites k if the indicator value of species j is the highest, among all groups, for that group of sites, and is statistically significant at a preselected significance level.

Indicator value The degree to which a species is indicator of (the conditions found in) a group of sites.

Pseudospecies To model the concept of differential species (i.e., species with clear ecological preferences), which is qualitative, TWINSPAN creates pseudospecies. Each species is recorded into a set of dummy variables, called pseudospecies, corresponding to relative abundance levels (or percentage cover for plants) at each site; these classes are cumulative. If the pseudospecies cutting levels are 1%, 11%, 26%, 51%, and 76%, for instance, a relative abundance of 18% at a site will occupy the first and second dummy pseudospecies vectors with "1" (=presence). Cutting levels are arbitrarily decided by users. A (sites × pseudospecies) data table is created.

Specificity The degree to which a species is found only in a given group of sites.

Ecological Interest for Indicator Species

How to identify characteristic or indicator species is a traditional problem in community ecology and biogeography. It is grounded in the paradigm that species are the best indicators available for particular environmental conditions. Field studies describing habitats or groups of sites often mention one or several species that characterize each habitat. Because indicator species add ecological meaning to site groups discovered by clustering, they provide criteria to compare typologies derived from data analysis, to identify where to stop dividing clusters into subsets, and to point out the main levels in a hierarchical classification of sites. Indicator species differ from species associations in that they are indicative of particular, predetermined groups of sites. Good indicator species should be found mostly in a single group or a limited number of groups of a typology, and be present at most of the sites belonging to that or these groups. This duality is of ecological interest; yet it is not always used in indicator species studies.

There is clearly a need for the identification of characteristic or indicator species in the fields of monitoring, conservation, and management. For example,

- Species may be used as indicators of the state of an ecosystem and of human-induced changes to the environment and biodiversity.
- In long-term environmental follow-up studies for conservation or ecological management, researchers are looking for biological indicators of habitat types or combinations of habitat types they want to preserve or rehabilitate (McGeoch and Chown, 1998).

As for many other ecological concepts, the empirical method of early ecologists who walked through ecosystems and examined data tables to identify indicator species has gradually been replaced by statistical methods that attempt to reproduce and formalize the reasoning of the traditional ecologists, with computer programs facilitating the calculations. In this article, the author will focus on statistical methods for the identification of indicator species, comparing the merits of the Two-way indicator species analysis (TWINSPAN) and Indicator value (IndVal) methods.

Statistical Methods: TWINSPAN and IndVal

TWINSPAN: A Brief View

Two-way indicator species analysis (TWINSPAN) (Hill, 1979) is fundamentally a method for hierarchical divisive classification of communities, based on progressive refinement of a single ordination axis obtained by correspondence analysis (CA) or detrended correspondence analysis (DCA) of a community composition (sites × species) data matrix. The algorithm, which is rather complex, is detailed by Kent and Coker (1992). An attractive feature of the computer output is a two-way table with the sites (columns) sorted according to the splits of the hierarchical classification. The species (rows) are also sorted so as to form blocks corresponding to the groups of sites of the classification. The body of the table contains the highest pseudospecies scores (*see* Glossary) found at the sites.

An additional feature of TWINSPAN is that it computes an indicator values index (I) for the species for every split of the hierarchical classification of the sites. According to Kent and Coker (1992), the index is computed as follows using the pseudospecies data (*see* Glossary):

$$I_j = \frac{n_j^+}{n^+} - \frac{n_j^-}{n^-}$$

where n^+ and n^- are, respectively, the number of sites on the (arbitrarily chosen) positive and negative sides of the split, whereas n_j^+ and n_j^- are the number of sites on the positive and negative sides, respectively, that contain pseudospecies j.

 http://dx.doi.org/10.1016/B978-0-12-384719-5.00430-5

A pseudospecies present in every site on the positive side and in none of the sites on the negative side obtains $I_j = 1$, and -1 if it is found in every site on the negative side and in none on the positive side. A pseudospecies that occurs in all sites on both sides of the split obtains $I_j = 0$. In TWINSPAN, only one pseudospecies of a single species is declared an indicator of a split, and that is the pseudospecies that has the highest absolute value of I. n_j^+/n^+ is the measure of fidelity to a group used in the INDVAL method (*see* The INDVAL Method).

The major disadvantage of TWINSPAN is that it can only compute the indicator value of species for the hierarchical classification produced by itself. Another critique is that the importance of a species in the analysis depends on the abundances of the other species because the pseudospecies, which form the data basis of the method, are based on species relative abundances. The method has also been criticized by Belbin and McDonald (1993) because it assumes the existence of a single, strong gradient dominating the data structure, and because the cutting points for the whole group, and then for subgroups, are always chosen to be the centroid of the group to be split instead of a point where a large gap occurs in the data. Because of that, sites that are very similar in species composition may end up in separate groups. The last problem has been alleviated by a modification to the method proposed by Roleček *et al.* (2009).

The INDVAL Method

Dufrêne and Legendre (1997) presented an alternative to TWINSPAN in the search for indicator species and species assemblages characterizing groups of sites. Like TWINSPAN, the indicator value (IndVal) method analyses the species with reference to a prior partition of the sites. The first novelty of IndVal is that it derives indicator species from any hierarchical or nonhierarchical classification of the objects (sampling sites), contrary to TWINSPAN where indicator species can only be derived for a classification obtained by splitting sites along a CA axis. The second novelty lies in the way the indicator value of a species is measured for a group of sites. The indicator value index (INDVAL) is based only on within-species abundance and occurrence comparisons; its value is not affected by the abundances of other species. The significance of the indicator value of each species is assessed by a randomization procedure.

The *IndVal* index is defined as follows. For each species j in each cluster of sites k, one computes the product of two values, A_{kj} and B_{kj}. A_{kj} is a measure of specificity based on abundance values whereas B_{kj} is a measure of fidelity computed from presence data:

$$A_{kj} = N\text{individuals}_{kj}/N\text{individuals}_{+k}$$

$$B_{kj} = N\text{sites}_{kj}/N\text{sites}_{k+}$$

$$INDVAL_{kj} = A_{kj}B_{kj}$$

n the formula for specificity (A_{kj}), N individuals$_{kj}$ is the mean abundance of species j across the sites pertaining to cluster k and N individuals$_{+k}$ is the sum of the mean abundances of species j within the various clusters. The mean number of individuals in each cluster is used, instead of summing the individuals across all sites of a cluster, because this removes any effect of variations in the number of sites belonging to the various clusters. Differences in abundance among sites of a cluster are not taken into account. A_{kj} is maximum when species j is present in cluster k only. In the formula for fidelity (B_{kj}), N sites$_{kj}$ is the number of sites in cluster k where species j is present and N sites$_{k+}$ is the total number of sites in that cluster, as in index I of TWINSPAN. B_{kj} is maximum when species j is present at all sites of cluster k. Quantities A and B must be combined by multiplication because they represent independent information (i.e., specificity and fidelity) about the distribution of species j.

The indicator value of species j for a partition of sites is the largest value of $IndVal_{kj}$ observed over all clusters k of that partition:

$$IndVal_j = \max[IndVal_{kj}]$$

The index is maximum (its value is 1) when the individuals of species j are observed at all sites belonging to a single cluster. A random permutation procedure of the sites among the site groups is used to test the significance of $IndVal_j$. A correction for multiple testing is necessary before reporting the results when multiple tests (for several species) have been conducted. The index can be computed for a given partition of the sites, or for all levels of a hierarchical classification of the sites.

Numerical Example

Table 1 describes the example given by Dufrêne and Legendre (1997), slightly modified, to illustrate the computation of the *IndVal* index. The data represent three species observed at 25 sites, which are divided into five groups. To facilitate comparisons, the sums of the mean group abundances are 20 for all three species. For species 1, $IndVal_{k1}$ has the highest value (0.30) for group 2, so $IndVal_1 = 0.30$. Following similar reasoning, $IndVal_2 = 0.40$ and $IndVal_3 = 0.90$. The permutational p-values computed by functions **IndVal()** of LABDSV and ***multipatt*()** of INDICSPECIES in R are significant in all three cases.

Statistical Method: *INDVAL* Expanded

De Cáceres and Legendre (2009) described other statistics that can be used to assess the indicator value of species. The statistics are divided into (1) correlation indices, which are used for determining the ecological preferences of species among a set of alternative site groups or site group combinations, and (2) indicator value indices, including *IndVal*, which are used for assessing the predictive values of species as indicators of the conditions found in site groups, for example for field determination of community types or ecological monitoring. Each of these families of indices comes in different types: there are indices for presence–absence and for quantitative species data; there are also nonequalized indices that give equal weights to individual sites and group-equalized indices that give equal weights to all groups whatever the number of sites they contain. For studies involving several groups of sites, De Cáceres *et al.* (2010) showed that the interpretation of indicator value analysis could be improved by computing the statistics for all possible

Table 1 Numerical example: abundance of three species at 25 sites divided into five groups

Groups	*Group 1*					*Group 2*					*Group 3*					*Group 4*					*Group 5*				
Sites	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Species 1	4	4	4	4	4	6	6	6	6	6	5	5	5	5	5	3	3	3	3	3	2	2	2	2	2
Species 2	8	8	8	8	8	4	4	4	4	4	6	6	6	6	6	4	4	2	0	0	0	0	0	0	0
Species 3	18	18	18	18	18	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Species 1																									
A_{k1}	4/20 = 0.20					6/20 = 0.30					5/20 = 0.25					3/20 = 0.15					2/20 = 0.10				
B_{k1}	5/5 = 1					5/5 = 1					5/5 = 1					5/5 = 1					5/5 = 1				
$IndVal_{k1}$	0.20					**0.30**					0.25					0.15					0.10				
Species 2																									
A_{k2}	8/20 = 0.40					4/20 = 0.20					6/20 = 0.30					2/20 = 0.10					0/20 = 0.00				
B_{k2}	5/5 = 1					5/5 = 1					5/5 = 1					3/5 = 0.6					0/5 = 0				
$IndVal_{k2}$	**0.40**					0.20					0.30					0.06					0.00				
Species 3																									
A_{k3}	18/20 = 0.90					2/20 = 0.10					0/20 = 0.00					0/20 = 0.00					0/20 = 0.00				
B_{k3}	5/5 = 1					5/5 = 1					0/5 = 0					0/5 = 0					0/5 = 0				
$IndVal_{k3}$	**0.90**					0.10					0.00					0.00					0.00				

Top panel: species abundance data. Bottom: calculation of the specificity (A_{kj}), fidelity (B_{kj}) and $IndVal_{kj}$ index for each species (j) in each group of sites (k). The maximum value of $IndVal_{kj}$ for each species is in bold.

Source: Modified from Dufrêne M and Legendre P (1997) Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs* 67, 345–366.

combinations of site groups. Moretti *et al.* (2010) published an application of that method.

A detailed discussion of the limitations of indicator value analysis is found in De Cáceres *et al.* (2010). They point out that more indicator species will be found than expected by chance when the classification of sites was obtained from the same species composition data that are used for IndVal analysis. In that case, p-values must be interpreted with caution: they do not result from a genuine test of significance where the classification of sites has to be independent of the species data used in the test.

Variants of the *IndVal* index have been proposed by Podani and Csányi (2010). Instead of using specificity and fidelity alone, these authors proposed to define the indicator value of a species as the product of two among three quantities: specificity A_{kj} (that they called concentration), specificity (new equation, with allowance for positive or negative species preferences), and fidelity B_{kj}. They provided formulas based on presence–absence or abundance data for each of these three quantities.

According to McGeoch and Chown (1998), IndVal is important for conservation biology because it is conceptually straightforward and allows researchers to identify biological indicators for any combination of habitat types or areas of interest – existing conservation areas, or groups of sites based on the outcome of a classification procedure. The method is sometimes used to identify biological indicators for groups of sites classified using the target taxa, subject to the caveat about p-values mentioned two paragraphs up, or in other instances using nontarget taxa, for example finding insect biological indicators for a classification of sites based on plant community data.

An IndVal index for a species is calculated independently of the other species in the assemblage. Because of that, comparisons of indicator values can be made between taxonomically unrelated taxa, or taxa in different functional groups or in different communities. Comparisons across taxa are robust to differences in abundance that may or may not be due to differences in capturability or sampling methods. The method is also robust to differences in the numbers of sites between site groups, in abundance among sites within a particular group, and in the absolute abundances of very different taxa that may show similar trends.

When a group of sites for which indicator species are sought corresponds to a well-delimited geographic area, superposition of distribution maps for the indicator species of that group should help identify the core conservation areas for these species, even when little other biological information is available.

Taxa described as biological indicators in the literature are often merely the favorite taxa of their proponents; ornithologists prefer birds, lepidopterists butterflies, and coleopterists beetles. IndVal provides an objective method for addressing this problem by enabling assessment of the relative merits of different taxa for a given study area (McGeoch and Chown, 1998). The species that emerge from this procedure as the best indicators for a group of sites should prove useful for monitoring site alterations, natural or man-induced.

Ecological Applications

Carabid Beetles in Belgium

Dufrêne and Legendre (1997) analyzed a large data set of Carabid beetle distributions in open habitats of Belgium

(189 species collected in pitfall traps at 69 sites). Classification of the sites was obtained by distance-based *K*-means partitioning computed as follows: first, a distance matrix (percentage difference, also called Bray-Curtis distance) was computed from the log-transformed species abundance data; this distance matrix was subjected to principal coordinate analysis (PCoA), also called metric multidimensional scaling; all ordination axes (i.e., the principal coordinates) produced by PCoA were used as input data into *K*-means partitioning. Although the clusters produced by *K*-means were not forced to be hierarchically nested, they showed a strong hierarchical structure for $K =$ two to ten groups. This allowed the authors to represent the relationships among partitions as a dendrogram with $K = 10$ groups, which corresponded to the main types of habitat recognized *a priori* by surveying the sites.

Indicator values were computed for each species and partitioning level. Some species were found to be stenotopic (narrow niches) whereas others were eurytopic (species with wide niches, present in a variety of habitats). Other species characterized intermediate levels of the hierarchy. The best indicator species ($IndVal > 0.25$) were assembled into a two-way indicator table; this tabular representation displayed hierarchical relationships among the species.

The IndVal results were compared to TWINSPAN. The partitions of sites used in the two methods were not the same; the TWINSPAN typology was obtained by partitioning CA ordination axes. TWINSPAN identified, as indicators, pseudospecies pertaining to very low cut-off levels. These species were not particularly useful for prediction because they were simply known to be present at all sites of a group. Several species identified by TWINSPAN as indicators also received a high indicator value from the IndVal procedure for the same or a closely related habitat class. IndVal identified several other indicator species, with rather high indicator values, that also contributed to the specificity of the groups of sites but had been missed by TWINSPAN. So, the IndVal method appeared to be more sensitive than TWINSPAN to the fidelity and specificity of species.

More Application Examples

Here are more examples of the many applications of indicator species analysis found in the literature.

- Borcard (1996) and Borcard and Vaucher-von Ballmoos (1997) present applications of the indicator value method to the identification of the Oribatid mite species that characterize well-defined zones in a peat bog of the Swiss Jura.
- The indicator values of beetle species characterizing different types of forests have been studied by Barbalat and Borcard (1997).
- Tuomisto *et al.* (2003) used spatially-constrained clustering to group 86 sampling units, each 500 m in length, forming a 43-km long transect in the Amazonian rain forest in Peru, into spatial clusters on the basis of satellite image pixel values. They also surveyed the ferns and Melastomaceae in the 86 sampling units in the field. Then they used the IndVal method to determine the species of ferns and Melastomaceae that were good indicators of the spatial clusters.
- Legendre *et al.* (2009) used multivariate regression tree analysis (De'ath, 2002) to identify habitat types that were similar in topographic conditions and in species composition in a Chinese permanent forest plot divided in 20 m × 20 m quadrats; then they used the IndVal method to identify, among the 159 tree species, the nine species that were statistically significant indicators of the five main habitat types. In this paper, the species used for IndVal analysis had been used to obtain the classification of the sites; as a consequence, the p-values had to be interpreted with caution.
- De Cáceres *et al.* (2010) carried out indicator species analysis of the vegetation of the Barro Colorado Island (BCI) permanent forest plot in Panama, also divided in 20 m × 20 m quadrats, grouped into seven habitat types identified in the literature. Among 307 tree species, they identified 44 indicator species of individual habitat types and 64 species for habitat type combinations. In this paper, the classification of the sites was independent of the species analyzed for indicator value.

Conclusion

Indicator species are biological indicators of groups of sites representing habitat types or combinations of habitat types; they are of prime interest for ecosystem conservation and management. Statistical methods are now available to identify indicator species in different situations (presence–absence or abundance data, group-equalized or nonequalized indices) and for a variety of purposes: correlation indices are used for determining the ecological preferences of species among a set of alternative site groups or site group combinations, whereas indicator value indices are used for assessing the predictive values of species as indicators of the conditions found at groups of sites. Both types of indices are readily available for computation in R functions. Tests of significance can be computed for these indices, producing p-values that help identify the most interesting indicator species.

Software

In the R statistical language, indicator value indices (IndVal) can be computed by functions ***strassoc()*** and ***multipatt()*** of INDICSPECIES and by function ***indval()*** of LABDSV. The functions in INDICSPECIES offer a choice of the several different indicator statistics described in De Cáceres and Legendre (2009). The IndVal index is also available in the computer package PC-ORD. The TWINSPAN program, in FORTRAN, can be obtained from several Web pages.

See also: Indicator Species

References

Barbalat S and Borcard D (1997) Distribution of four beetle families (Coleoptera: Buprestidae, Cerambycidae, phytophagous Scarabaeidae and Lucanidae) in different forest ecotones in the Areuse Gorges (Neuchâtel, Switzerland). *Écologie* 28: 199–208.

Belbin L and McDonald C (1993) Comparing three classification strategies for use in ecology. *Journal of Vegetation Science* 4: 341–348.

Borcard D (1996) Typologie des assemblages d'espèces d'Oribates (Acari, Oribatei) de la tourbière du Cachot (Jura suisse): espèces indicatrices ou groupements caractéristiques? *Bulletin de la Societe neuchâteloise des Sciences naturelles* 119: 63–73.

Borcard D and Vaucher-von Ballmoos C (1997) Oribatid mites (Acari, Oribatida) of a primary peat bog-pasture transition in the Swiss Jura mountains. *Écoscience* 4: 470–479.

De'ath G (2002) Multivariate regression trees: A new technique for modeling species–environment relationships. *Ecology* 83: 1105–1117.

De Cáceres M and Legendre P (2009) Associations between species and groups of sites: Indices and statistical inference. *Ecology* 90: 3566–3574.

De Cáceres M, Legendre P, and Moretti M (2010) Improving indicator species analysis by combining groups of sites. *Oikos* 119: 1674–1684.

Dufrêne M and Legendre P (1997) Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs* 67: 345–366.

Hill MO (1979) *TWINSPAN – A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-way Table by Classification of the Individuals and Attributes*. Ithaca, New York: Section of Ecology and Systematics, Cornell University.

Kent M and Coker P (1992) *Vegetation Description and Analysis – A Practical Approach*. New York: John Wiley & Sons.

Legendre P, Mi X, Ren H, *et al.* (2009) Partitioning beta diversity in a subtropical broad-leaved forest of China. *Ecology* 90: 663–674.

McGeoch MA and Chown SL (1998) Scaling up the value of bioindicators. *Trends in Ecology and Evolution* 13: 46–47.

Moretti M, De Cáceres M, Pradella C, *et al.* (2010) Fire-induced taxonomic and functional changes in saproxylic beetle communities in fire sensitive regions. *Ecography* 33: 760–771.

Podani J and Csányi B (2010) Detecting indicator species: some extensions of the IndVal measure. *Ecological Indicators* 10: 1119–1124.

Roleček J, Tichy L, Zeleny D, and Chytry M (2009) Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. *Journal of Vegetation Science* 20: 596–602.

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