Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 973–982 doi: 10.1111/j.2041-210X.2012.00246.x

Using species combinations in indicator value analyses

Miquel De Cáceres[1,2]*, Pierre Legendre[3], Susan K. Wiser[4] and Lluís Brotons[1,2]

[1]*CTFC (Centre Tecnològic Forestal de Catalunya), Ctra. antiga St. Llorenç km 2, E-25280 Solsona, Catalonia, Spain;* [2]*CREAF (Centre de Recerca Ecològica i Aplicacions Forestals), Universitat Autònoma de Barcelona, Bellaterra, Catalonia, Spain;* [3]*Département de Sciences Biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, QC, Canada, H3C 3J7; and* [4]*Landcare Research, PO Box 40, Lincoln, 7640, New Zealand*

Summary

1. Indicator species are often determined using an analysis of the relationship between the species occurrence or abundance values from a set of sites and the classification of the same sites into site groups (habitat types, community types, disturbance states, etc.). It may happen, however, that a particular site group has no indicator species even if its sites have a community composition that is clearly distinct from the sites of other site groups. This motivates an exploration of the indicator value of not only individual species but also species combinations.

2. Here, we present a novel statistical approach to determine indicators of site groups using species data. Unlike traditional indicator value analysis, we allow indicators to be *species combinations* in addition to single species. We require that *all* the species forming the combination must occur in the site to use the combination as an indicator. We present a simple algorithm that identifies the set of indicators (each one being either a single species or a species combination) that show high positive predictive value for the target site group. Moreover, we demonstrate the use of the percentage of sites of the site group where at least one of its valid indicators occurs to determine whether the group can be reliably predicted throughout its range.

3. Using a simulation study, we show that if two species are not strongly correlated and their frequency in the data set is larger than the frequency of sites belonging to the site group, the *joint* occurrence of the two species has higher positive predictive value for the site group than the two species taken independently.

4. We illustrate the proposed method by determining which combinations of vascular plants can be used as indicators for 29 shrubland and forest vegetation types of New Zealand.

5. The proposed methodology extends traditional indicator value analyses and will be useful to develop multispecies ecological or environmental indicators. Further, it will allow newly surveyed sites to be reliably assigned to previously defined vegetation types.

Key-words: diagnostic species, environmental indication, indicator species, indicator value, vegetation classification, vegetation types

Introduction

Determining the occurrence or abundance of a small set of indicator species, as an alternative to sampling the entire community, has been particularly useful in long-term environmental monitoring for conservation or ecological management. Species are chosen as indicators if they (i) reflect the biotic or abiotic state of the environment; (ii) provide evidence for the impacts of environmental change; or (iii) predict the diversity of other species, taxa or communities within an area (McGeoch 1998; Niemi & McDonald 2004). Indicator species are often determined by analysing the relationship between the species occurrence or abundance values from a set of surveyed sites and the classification of these sites into groups (Dufrêne & Legendre 1997; De Cáceres & Legendre 2009). The classification of sites into groups (hereafter called 'site groups') may have been derived from the similarities in environmental conditions among sites (e.g. habitat types or disturbed/undisturbed states), or in species composition (e.g. community or vegetation types); site groups may also have been provided by the study design (e.g. when comparing across geographic regions or repeated surveys) or obtained using other criteria, such as land use classes. With respect to individual species, the analysis of the strength of its association to site groups provides a characterization of the species niche preferences and allows its degree of ecological specialization to be assessed (De Cáceres, Legendre & Moretti 2010b; Chazdon, Chao & Colwell 2011). With respect to the site group, the list of species strongly associated with it allows the determination of whether a newly surveyed site can be labelled with the concept that the site group represents. Owing to their predictive value, indicator species possess an undeniable appeal for conservationists and

*Correspondence author. E-mail: miquelcaceres@gmail.com

land managers as they provide a cost- and time-efficient mean to assess ecosystem change (McGeoch & Chown 1998; Hilty & Merenlender 2000; Carignan & Villard 2002; McGeoch, Van Rensburg & Botes 2002).

Several alternatives exist for the statistical determination of indicator species (e.g., Hill 1979; Dufrêne & Legendre 1997; Bruelheide 2000; Chytrý *et al*. 2002; Chazdon, Chao & Colwell 2011; Urban *et al*. 2012). Among them, the most frequently used approach involves the assessment of the association between species and site groups using correlation or indicator value indices (Dufrêne & Legendre 1997; Chytrý *et al*. 2002; De Cáceres & Legendre 2009). Correlation indices assess the relative positive or negative preference of the species for the site group, compared with the remaining groups (Chytrý *et al*. 2002). In contrast, indicator value indices are non-negative and assess to what extent the sites of the target site group match the sites where the species is found (De Cáceres, Legendre & Moretti 2010b).

Typically, the output of an indicator value analysis for a given site group consists of the list of species that are significantly associated with it, presented in a decreasing indicator value order. When any of the indicator species is found in a newly surveyed site, the site can be assigned to the site group. The more indicator species are found in the newly surveyed site, the higher the confidence on the assignment. To quantify this degree of confidence, however, it would be desirable to know the probability of the indicated site group given the *joint* occurrence of all the indicator species found; unfortunately, this conditional probability is not provided by the traditional indicator value analysis. Moreover, in some occasions, the list of indicator species for a particular site group is empty even if its sites have a community composition that is clearly distinct from sites of other site groups. In these cases, one could use the *joint* occurrence of several species to indicate the site group. These issues motivate the need to explore the indicator value of not only individual species, but species combinations.

Here, we present a new statistical approach to determine *indicators of site groups* using species data. In addition to (*single*) *indicator species* as produced in standard indicator species analysis, the new method also searches for *indicator species combinations*. In the latter case, it is the joint (i.e. simultaneous) occurrence of several species in a site that is used as indication of the site group. Species assemblages have been used to develop ecological or environmental indicators. For example, Butler *et al*. (2012) recently provided a protocol to determine which bird species should be considered when building population-based indices aimed at representing the status of the wider bird community. There are also examples of indicator value analyses conducted using supra-specific taxonomic entities (genera, families, etc.) or even functional guilds (e.g. Basset *et al*. 2004). In these analyses, the occurrence of a single species, among those conforming the species group, was enough for indication. To our knowledge, the simultaneous occurrence of several species has not yet been considered within the framework of indicator value analysis.

In this article, we first briefly review the original indicator value analysis developed by Dufrêne & Legendre (1997). We then generalize the method from single species to species combinations and present the steps that allow a set of valid indicators for a given site group to be determined; we hereafter use the word 'indicator' to denote an entity that is used for indication. We then present the results of a simulation study showing the potential benefits of considering species combinations as indicators. After that, we illustrate our methodological suggestions in an example where we determine indicator plant combinations for 29 types of shrublands and forests in New Zealand. Finally, we discuss the advantages and limitations of this new approach, its potential applications and relate it to other existing extensions of the method.

Indicator value analysis for single species

This section briefly reviews the concepts and calculations of traditional indicator value analyses. Imagine that a given species S has been detected in a newly surveyed site and one is interested in knowing whether the site belongs to a given site group G (Fig. 1a). After finding the species, one would like to know $A = P(G|S)$, the probability that the surveyed site belongs to the target site group G given the fact that species S has been found. This conditional probability (A) is called the *specificity* or *positive predictive value* of species S as indicator of the site group. To be useful, indicator species not only need to have a high positive predictive value, but they also need to be easy to find. The higher the probability of finding the species in sites belonging to the site group ($B = P(S|G)$), the more likely it is that the species will be found in newly surveyed sites of the same kind. This second conditional probability (B) is called the *fidelity* or *sensitivity* of the species as indicator of the target site group (Murtaugh 1996; Dufrêne & Legendre 1997; De Cáceres & Legendre 2009).

The input of an indicator value analysis consists of two elements: (i) a site-by-species community data table **X** containing the occurrence or abundance values of species at several locations or sites; and (ii) a partition of the sites into a set of nonoverlapping classes (site groups). In their work, Dufrêne & Legendre (1997) proposed that good indicator species should be at the same time ecologically restricted to the target site group and frequent within it. They defined the Indicator Value (*IndVal*) index of a species in a site group as the *product* of A and B. The sensitivity of the species, B (which they called *fidelity*), was simply estimated as the relative frequency of the species in sites belonging to the target site group (Table 1). In contrast, the positive predictive value, A (which they called *specificity*), could be calculated from either presence–absence or abundance data. In fact, as described in the study by De Cáceres & Legendre (2009), there are at least four different ways of estimating A. If only presence–absence data are used, and assuming a representative sample of sites, an appropriate estimator of A is the number of occurrences of the species within sites belonging to the target site group, divided by the number of occurrences of the species S across all sites (A_{pa} in Table 1). Alternatively, A can be estimated from abundance data as the sum of abundance values of the species within sites belonging to the site group divided by the sum of abundance

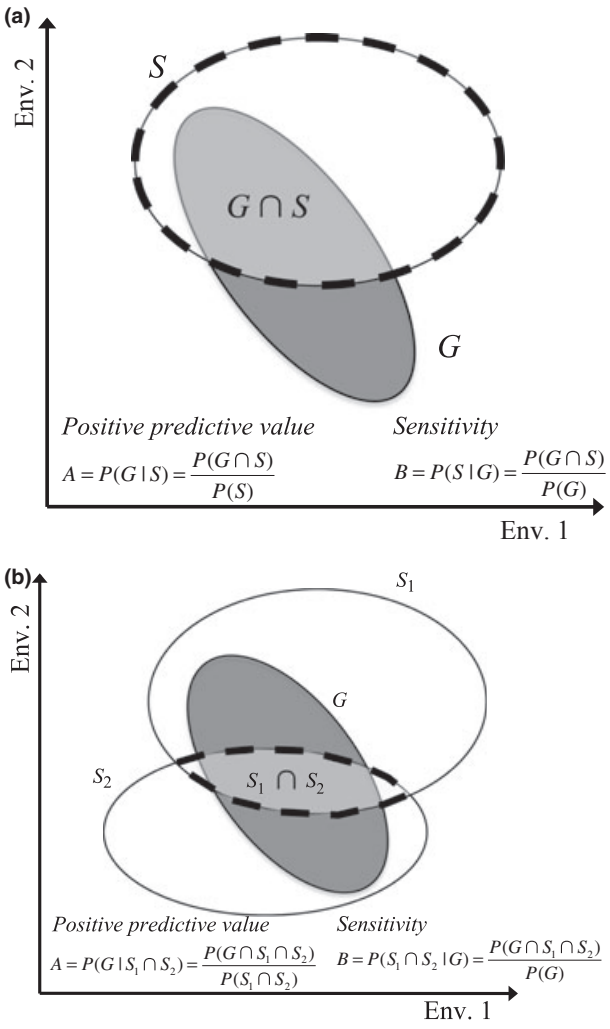

Fig. 1. Concepts of positive predictive value and sensitivity for a single species (a) and for a species pair (b). Open ovals represent the range of environmental conditions suitable for species S_i, whereas the dark grey oval represents the environmental conditions corresponding to the definition of the target site group G. Thick dashed lines delimit the occurrence of indicators (a single species or a species pair).

values across all sites (A_{ind} in Table 1). In these first two indices, we assumed that the target site group was not over- or under-sampled with respect to others in the data set. It often happens, however, that some site groups are over-represented with respect to others. In these cases, we may decide to give the same weight to all site groups in the calculation of A irrespective of the number of sites each group actually contains. For presence–absence data, we would calculate the relative frequency of the species in the target site group divided by the sum of relative frequencies over all groups (see A^g_{pa} in Table 1), whereas for abundance data, A is then defined as the mean abundance of the species in the target site group divided by the sum of mean abundance values over all groups (A^g_{ind} in Table 1; this was the index suggested by Dufrêne & Legendre 1997).

After calculating the *IndVal* value for all site groups, one looks for the site group to which the species is maximally associated. To report that the species is associated with this site

Table 1. Formulae used to estimate the positive predictive value (A) and sensitivity (B) of an indicator for a given site group. The same formulae are valid for species or species combinations, but different community data tables are used as input (either **X** or **C**, respectively; these are defined in the text) to calculate the indicator value statistics

	Nonequalized	Group equalized
Positive predictive value for presence–absence data	$A_{pa} = \frac{n_p}{n}$	$A^g_{pa} = \frac{n_p/N_p}{\sum_{k=1}^{K} n_k/N_k}$
Positive predictive value for abundance data	$A_{\text{ind}} = \frac{a_p}{a}$	$A^g_{\text{ind}} = \frac{a_p/N_p}{\sum_{k=1}^{K} a_k/N_k}$
Sensitivity	$B = \frac{n_p}{N_p}$	

Notation is as in De Cáceres & Legendre (2009): N_p, number of sites that belong to the *target* site group; n, number of occurrences of the indicator across all sites; n_p, number of occurrences of the indicator within sites that belong to the *target* site group; N_k, number of sites that belong to the site group k; n_k, number of occurrences of the indicator within sites that belong to the site group k; a_p, sum of the abundance values of the indicator within the *target* site group; a_k, sum of the abundance values of the indicator within sites of the site group k; a, sum of the abundance values of the indicator over all sites.

group, one-first needs to reject the null hypothesis that negates this association. In traditional indicator value analysis, the maximum *IndVal* value across site groups is usually tested for statistical significance using a permutation test, a procedure that involves comparing an observed test statistic with a distribution of the same statistic obtained by randomly reordering the data (Dufrêne & Legendre 1997; De Cáceres & Legendre 2009; De Cáceres, Legendre & Moretti 2010b). Alternatively, permutation tests exist to test statistical hypotheses regarding the association with each site group separately (Bakker 2008; De Cáceres & Legendre 2009). As a complement to testing null hypotheses, we can assess the precision of estimates of A, B and their product by calculating confidence intervals. Unfortunately, the exact or approximated parametric distribution is difficult to determine for many of the above indices. Alternatively, one can use the percentile bootstrap method, which involves resampling the observed data with replacement to generate an approximate distribution of the estimator, followed by taking the percentiles $\alpha/2$ and $1 - \alpha/2$ of the empirical distribution as the $(1 - \alpha)$ confidence limits (Manly 1997). Using computer simulations, De Cáceres & Legendre (2009) studied the performance of the simple percentile bootstrap method to obtain confidence intervals for indicator value indices.

Indicator value analysis for species combinations

RATIONALE OF THE METHOD

Instead of considering only one species at a time, this section defines indicators formed by *combining* the presence or abundance data of k species, S_1, S_2 … S_k. As an example, we first consider the indicator consisting of the *joint* occurrence of two species, say S_1 and S_2 (Fig. 1b). To assess the positive predictive value of this indicator, we need to estimate $P(G|S_1 \cap S_2)$, that is, the probability of the surveyed site belonging to the target site group given that S_1 and S_2 have been

found together. This conditional probability can be estimated as the number of sites within site group G where both S_1 and S_2 occur, divided by the number of sites where both S_1 and S_2 occur across all sites (note that we do not use '∩' as a set operator but to specify the set of sites where both S_1 *and* S_2 occur). As there are equal or fewer sites where two species co-occur than sites where only one of the species is present, $P(G|S_1 \cap S_2)$ will usually be estimated using a smaller sample size than $P(G|S_1)$ or $P(G|S_2)$ and hence the estimation will be less precise. For the same reason, the sensitivity of the species pair, $P(S_1 \cap S_2|G)$, will always be equal to or smaller than $P(S_1|G)$ and $P(S_2|G)$. Nevertheless, the indicator consisting of the two species jointly may have higher positive predictive value compared with the indicators of the two species considered independently because it includes information coming from two events that are generally not completely correlated. If the two species were completely correlated, then considering the pair would not allow users to obtain more information. In general, to assess the indicator value for any combination of k species (S_1, S_2 … S_k), we need to estimate $P(G|I)$ and $P(I|G)$, the positive predictive value and sensitivity of the indicator $I = S_1 \cap \ldots \cap S_k$ consisting of the joint occurrence of all species S_1 to S_k. Accounting for the joint occurrence of increasing numbers of species increases the ability to indicate specific ecological conditions (Pignatti 1980), which may lead to higher A values. However, as shown for the species pair, considering species combinations instead of species separately will often lead to lower values of B and to lower precision of A estimates. This limits the number of species that can be considered simultaneously in an indicator species combination (Pignatti 1980; Bruelheide 2000).

THE COVERAGE OF THE TARGET SITE GROUP

The target site group may sometimes have a broad geographic range (for example, if it represents a widespread vegetation type). In such case, all indicators (whether species or species combinations) may have low sensitivity because it is likely that each of them will occur within only a subset of the geographical range. This potential limitation of indicators taken individually is the reason why having a *set* of indicators can be most useful: one indicator can be used in some parts of the range of the target site group, whereas another one can be used in other parts. Thus, an important quantity that complements a given list of indicators is their pooled *coverage* of the target site group, which we define as *the percentage of sites of the site group where at least one of the indicators occurs*. The coverage of the site group provided by a single indicator is equal to its sensitivity.

ALGORITHM

In what follows we describe the steps of an indicator species analysis considering species combinations. Unlike the traditional analysis, a separate analysis is conducted for each target site group.

Step 1 – selecting candidate species

Selecting the species that are to be combined is not mandatory, but reduces the number of species combinations to be explored in the analysis. An intuitive approach for selecting candidate species involves discarding those species that appear with low frequency in sites belonging to the site group. The analyst can further restrict the list of candidate species by including additional criteria (McGeoch 1998), for example discarding the species that are difficult to identify taxonomically.

Step 2 – setting a maximum number to the species forming a combination

As species combinations involving too many species are not normally useful as indicators, limiting the maximum number of species forming a combination can substantially reduce the computational requirements of the analysis. Let J be the number of candidate species and K the maximum number of species forming a species combination (clearly $1 \leq K \leq J$). The total number of species combinations that one can form is a sum of combinatorial numbers, $\sum_{k=1}^{K} \binom{J}{k}$. For example, if from a set of ten species (i.e. $J = 10$), one wishes to generate combinations of up to three species (i.e. $K = 3$), the total number of combinations to consider will be 10 (singletons) + 45 (pairs) +120 (triplets) = 175. If all possible combinations of candidate species were to be considered (i.e. if $K = J$), the number of combinations to explore would be $2^K - 1$ (= 1023 if $J = 10$).

Step 3 – calculation of data table **C**

Estimating the two components of indicator value for species combinations is essentially the same as for single species. The only difference is that the first element of the input for the indicator value analysis is no longer a site-by-species data table **X**, but a data table **C** with sites as rows and species combinations as columns. Table **C** will contain as many columns as species combinations the analyst wants to consider. Each entry in this data table contains the 'occurrence' or 'abundance' value of the corresponding species combination in the corresponding site. Values in **C** are calculated as follows. If **X** contains abundances, the 'abundance' of the species combination in a site is calculated as the *minimum* abundance value in that site among the abundances of all the species included in the combination. For example, if the individual counts for species S_1, S_2 and S_3 in a particular site are 10, 15 and 8, respectively, the value in **C** for combination $I = S_1 \cap S_2 \cap S_3$ in that site will be 8. We use the minimum, and not other statistics like the mean, because we define the indicator as the joint presence of all the species in the combination (i.e. the indicator does not 'occur' if one of the species has zero abundance). Analogously, if **X** contains presence–absence data the 'occurrence' of the species combination in a site will be a presence (one) if *all* the species of the combination are found, and an absence (zero) otherwise.

Step 4 – calculation of indicator value components

Once data table **C** is available, the estimation of $A = P(G|I)$ and $B = P(I|G)$ for species combinations is performed straightforwardly using the formulae of Table 1 and **C**, instead of **X**, as input data.

Step 5 – selecting valid indicators

Both hypothesis testing and confidence interval calculation can be conducted for species combinations, as described in the previous section for single species. However, conducting tests of hypotheses is not a good strategy to find the best indicators because a large number of species combinations can be significantly associated with the target site group, especially if the site groups were originally defined using species composition data (De Cáceres & Legendre 2009; De Cáceres, Legendre & Moretti 2010b). A more practical approach is to determine those indicators that are strongly restricted to the target site group. To determine the set of valid indicators, we recommend choosing a threshold for minimum positive predictive value (A_t). This threshold can be interpreted as one minus the maximum false-positive rate that the user is prepared to accept in future assignments (e.g. if $A_t = 0{\cdot}6$, and then all valid indicators will erroneously indicate the target site group less than 40% of the times). Then, a given species or species combination will be a *valid indicator* if the lower bound of the 95% confidence interval for A is equal or higher than threshold A_t. Using the lower bound of the confidence interval instead of the point estimation as selection criteria is important to ensure statistical significance because, as indicated above, the precision of estimates can be very low with species combinations. Additionally, one can also set a minimum value for sensitivity (B_t) and discard those indicators that are powerful but occur too rarely (e.g. in less than 25% of sites). After determining the set of valid indicators, one should calculate the coverage of the site group that this set provides. If the coverage is small (e.g. < 40%), this means that the site group cannot be predicted with the desired predictive power in most of its range. Repeating this step with a lower threshold for positive predictive value (A_t) may increase the coverage, at the expense of increasing the rate of false positives in future assignments. Plotting the relationship between the desired positive predictive value of assignments and the coverage of the site group can be useful to guide the choice of A_t (see example application below, Fig. 3).

To increase the accessibility of our proposals, we implemented steps 3-5 of the above algorithm in a function (called 'indicators') written in R language (R Development Core Team 2012) and added it to the R library 'indicspecies' (De Cáceres & Legendre 2009).

Simulation study

We conducted a simulation study to better understand where it may be beneficial to consider species combinations rather than single species, as indicators. Let S_1 and S_2 be two potential indicator species of a target site group G. The question is whether the species pair ($S_1 \cap S_2$) will have higher positive predictive value as indicator of G than either S_1 or S_2 taken individually. This will depend on the relationship between each species and G, but also on the relationship between the two species. Furthermore, if the positive predictive value of both S_1 and S_2 is lower than A_t but that of $S_1 \cap S_2$ is higher, one should observe an increase in the coverage of the site group by considering the species pair as indicator. With these expectations in mind, we used the R package 'binarySimCLF' (Qaqish 2003) to generate three correlated random binary variables (i.e. G, S_1 and S_2) with known marginal proportions ($P(S_1)$, $P(S_2)$ and $P(G)$) and correlation structure (r_{S1S2}, r_{S1G} and r_{S2G}). To reduce the number of parameter combinations to explore, we restricted our study to $P(S_1) = P(S_2)$ (i.e. the two species occur with the same frequency in the data set) and $r_{S1G} = r_{S2G}$ (i.e. the species are equally correlated with the target site group). For each simulated data set, we evaluated (1) the difference in positive predictive value of $S_1 \cap S_2$ with respect to the maximum value observed between S_1 and S_2; (2) the difference in coverage obtained when considering $S_1 \cap S_2$ as indicator, compared with not considering it. Each simulated data set included 100 sites, and for each parameter combination, we averaged the results obtained for 100 simulated data sets.

If the two species occur more frequently than the site group ($P(S_i) = 0{\cdot}50$, $P(G) = 0{\cdot}25$), we found that the species pair had larger positive predictive value than the single species indicators only if both species had a moderate correlation with the site group ($r_{S1G} = r_{S2G} > 0{\cdot}3$; Fig. 2a). Moreover, the lower the correlation between the two species (r_{S1S2}), the better the advantage of considering the species pair. An increase in the coverage of the target site group was observed in the same situations (Fig. 2b). If the species occur with the same frequency as the site group [$P(S_i) = P(G) = 0{\cdot}50$], one could still gain positive predictive value by considering the species pair (Fig. 2c). However, in this case, the single species indicators tended to have higher positive predictive value than in the previous case; hence, including the species combination did not increase the coverage of the target site group substantially (Fig. 2d). From these results, one would expect species combinations to be most useful when the component species of the combination are common, have some preference for the site group and a low degree of correlation to each other. Such a situation could occur, for example, if there is a partial overlap between the habitat niches of two species and the target site group is a habitat in the zone of overlap.

Example application – New Zealand's woody vegetation types

New Zealand's indigenous forests and shrublands cover ca. 23% and 10% of its land surface, respectively. The forests are predominantly evergreen and dominated by different combinations of southern beeches (*Nothofagus* spp.), broadleaved angiosperms, the kauri (*Agathis australis*) and other conifers, mainly podocarps. Shrublands occur in subalpine areas and in lowland and montane regions that were presumably forested in pre-human times (Wardle 1991). Wiser *et al.* (2011) used

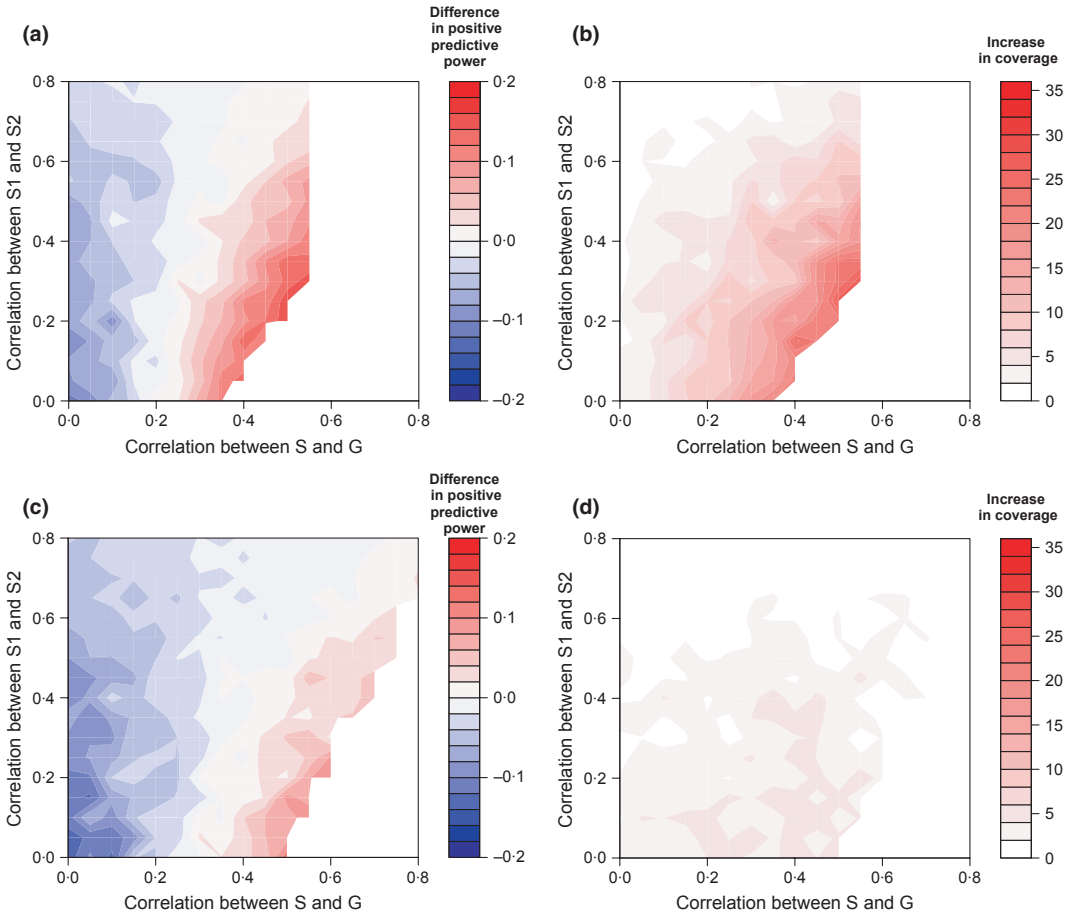

Fig. 2. Simulation study results showing the benefit of considering a species pair ($S_1 \cap S_2$) as indicator of a site group (G) instead of considering single species indicators (either S_1 or S_2). In panels (a) and (c), we show the difference in positive predictive value (A) between the species pair and the maximum of S_1 and S_2. In panels (b) and (d), we show the difference in coverage obtained when considering the indicator based on the species pair compared with not considering it (coverages were obtained using $A_t = 0{\cdot}6$). These values are plotted as a function of the correlation between the species and the site group (in the x-axis; both species are equally correlated with the site group) and the correlation between the two species (in the y-axis). Panels (a) and (b) show the results for the situation where the species occur more frequently than the target site group ($P(S_1) = P(S_2) = 0{\cdot}50$, $P(G) = 0{\cdot}25$), whereas panels (c) and (d) show the results for species being equally frequent as the target site group ($P(S_1) = P(S_2) = P(G) = 0{\cdot}50$). Empty parts of the plots indicate parameter combinations that cannot be simulated properly (see Qaqish 2003).

1177 systematically sampled vegetation plots as the basis for a quantitative classification exercise that produced 24 woody vegetation types. More recently, Wiser & De Cáceres (in press) used a fuzzy clustering framework (De Cáceres, Font & Oliva 2010a) to validate and extend this classification. For this second classification analysis, the authors merged the 1177 plots with vegetation plot records from the New Zealand's National Vegetation Survey Databank (Wiser, Bellingham & Burrows 2001; http://www.givd.info/ID/AU-NZ-001). As a result, seven of the 24 original vegetation types were discarded, whereas twelve new types were defined. The geographical extent of vegetation types varies from quite spatially restricted (< 150 000 ha) to widespread (>750 000 ha) (Wiser & De Cáceres in press).

We illustrate our method by determining whether combinations of vascular plants can be used as indicators for each of the 29 vegetation types. As input data, we took the community data table (5751 sites × 1930 vascular plant species) and the fuzzy membership table that resulted from the second classification analysis (Wiser & De Cáceres in press). We derived a hard classification from the fuzzy one by defining that a site was a member of a vegetation type if its fuzzy membership value was higher than 0·5. Using that rule, 3114 sites (plots) were members of one woody vegetation type or the other, whereas the remaining 2637 sites were deemed transitional between types or were too unique to belong to any type. The number of sites belonging to each vegetation type ranged from 15 to 366 (see Table S1 in Supporting Information). We selected as candidates those species occurring in at least 40% of the sites belonging to the target vegetation type. The number of candidate species ranged between 2 and 62 (Table S1). We considered as potential indicators from single species up to four-species combinations, and we calculated the positive predictive value, sensitivity and indicator value of all of them. Positive predictive values were calculated using cover abundance values and nonequalized indices (i.e. A_{ind} in Table 1). We estimated 95% bootstrap confidence intervals for all indices

using the simple percentile method with 10 000 bootstrap samples.

We display in Fig. 3a,b the coverage of two vegetation types (#2 and #15) as a function of the A_t value used as a threshold to determine valid indicators. These figures can be helpful to decide on which A_t threshold value to use. That is, they allow finding, by interpolation, the maximum A_t value that can be used while ensuring a given coverage of the target site group. These figures also show that the indicator value analysis often provides indicators with larger positive predictive values A when species combinations are considered, compared with an analysis with single species only. In some cases (like in Fig. 3a), this advantage allows the target site group to be predicted with moderate or high reliability in a large fraction of its range. In others (like in Fig. 3b), the positive predictive value of species combinations is not very high, despite being higher than single species, and hence the site group cannot be reliably predicted.

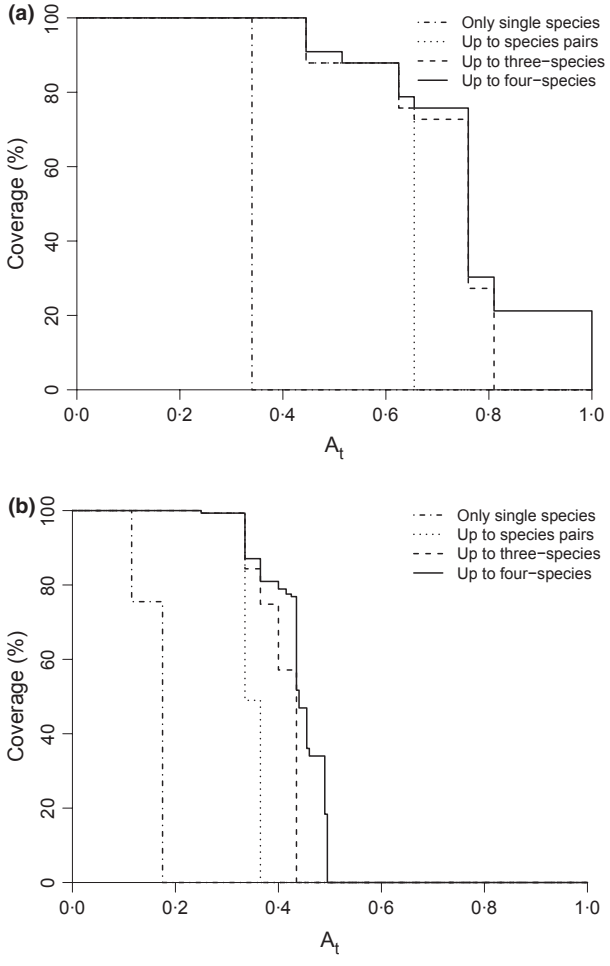

Fig. 3. Relationship between the minimum positive predictive value required for valid indicators (A_t) and the resulting coverage of the target site group. Coverage plots are shown for two site groups: (a) the montane shrubland type #2; and (b) the forest type #15. For each A_t value, valid indicators were determined comparing the lower boundary of the 95% confidence interval for A with A_t. Plotted lines correspond to coverage values obtained when considering single species indicators only, or when considering species combinations of increasing order, up to four species.

We selected the valid indicators using an $A_t = 0{\cdot}6$ threshold for positive predictive value and a $B_t = 0{\cdot}25$ threshold for sensitivity. Considering up to four-species combinations, we obtained valid indicators for 16 of 29 vegetation types only. Nevertheless, our results confirmed that species combinations often had higher predictive value than species taken individually (Fig. 4). For the remaining 13 vegetation types, reliable assignments cannot be made using indicator species or species combinations. In some cases, the list of valid indicators was quite long. Because many subsets of valid indicators can have the same coverage as the complete set, for the 16 vegetation types with valid indicators, we reduced the list of valid indicators as follows (note that many procedures would be possible for this last task). First, we removed those indicators whose occurrence pattern was nested within the occurrence pattern of others (because a nested indicator will have lower sensitivity). Second, we determined the coverage of subsets of indicators by progressively increasing numbers of indicators until a subset with the same coverage as the complete set was found, or until a maximum of four indicators was reached. The final sets of valid indicators are shown in Table S2 in Supporting Information. We use vegetation type #2 to illustrate how to interpret the indicator value results for a given target site group. Newly surveyed sites can be assigned to this montane shrubland type if *Dracophyllum uniflorum* is found along with *Festuca novae-zelandiae* (probability of being wrong is between 0·1 and 0·4), or if *D. uniflorum* is found with *Celmisia spectabilis* (probability of being wrong is between 0·05 and 0·34). One or both combinations will be found in about 88% of the range of the vegetation type.

Discussion

ADVANTAGES AND LIMITATIONS OF THE METHOD

Questions that can be addressed using indicator species analyses include the following: 'Is species X strongly and significantly restricted to a particular site group, compared to the others?', 'What is the degree of habitat specialization of species Y?' or 'Can we use the occurrence of species Z to determine whether a given site belongs to a particular site group?'. Several statistical alternatives already exist to address these questions (e.g. Dufrêne & Legendre 1997; Chytrý *et al.* 2002; De Cáceres & Legendre 2009; De Cáceres, Legendre & Moretti 2010b; Chazdon, Chao & Colwell 2011; Urban *et al.* 2012). To further advance the indicator species approach, our goal was to obtain easy-to-apply rules to assign newly surveyed sites to target site groups based on species combinations. The kind of rules we were interested in is: 'If you find species X, Y and Z simultaneously occurring at a given site, you can assign the site to the target site group with a known probability of making a mistake'. Our method represents an important improvement over the well-known Indicator Value method (Dufrêne & Legendre 1997), because we have shown that species combinations may have higher positive predictive values – and hence be less prone to false-positive error – than species taken individually. Another advantage of considering species combinations is that the information about species that are

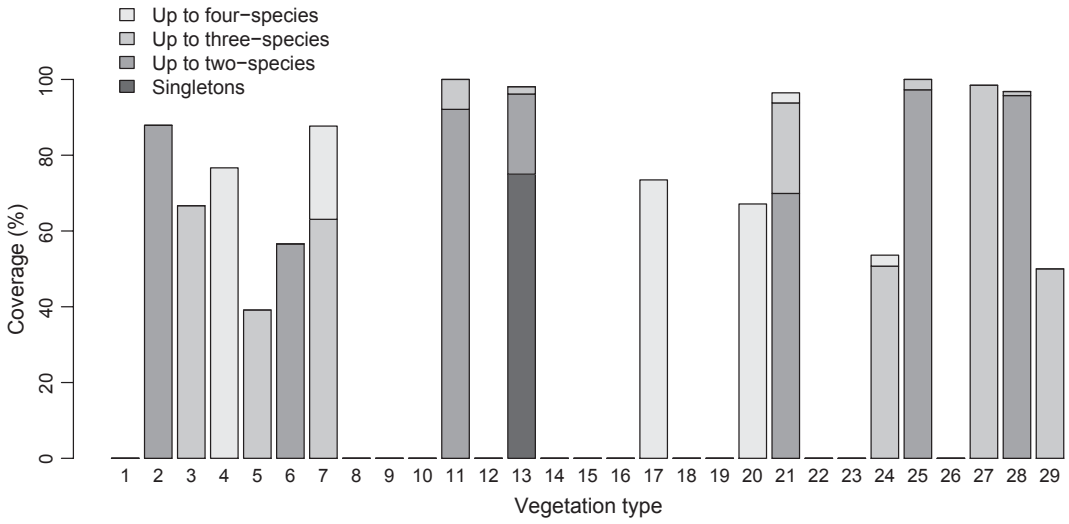

Fig. 4. Coverage of the 29 NZ woody vegetation types (i.e. percentage of each type covered by valid indicators) obtained when considering indicators of increasing order, from singletons up to combinations of four species. Valid indicators are those whose 95% confidence interval lower bound for A is higher than 0·6 and the corresponding lower bound for B is higher than 0·25. The indicator value results obtained using up to four-species combinations are given in Appendices S1 and S2 in Supporting Information.

frequent or dominant but are not restricted to the target site group can be taken into account for indication purposes. For example, one of the valid indicators that we found for the New Zealand vegetation type #2 was the combination of *D. uniflorum* and *F. novae-zelandiae*. Whereas *D. uniflorum* is mostly restricted to this montane shrubland type (A = 0·49, B = 1·00), *F. novae-zelandiae* is a common grass with only 7% of its occurrences being in shrublands of this type (A = 0·07, B = 0·82). Nevertheless, adding the information provided by the occurrence and abundance of *F. novae-zelandiae* to the indicator species *D. uniflorum* raises the positive predictive value (from A = 0·49 to A = 0·80) without markedly decreasing sensitivity (from B = 1·00 to B = 0·82).

Despite its advantages, our method is limited to the number of species it can include in species combinations. If many species are combined, the sensitivity of the indicator will be very low and the uncertainty of both sensitivity and positive predictive value estimates will increase. Furthermore, the number of combinations to evaluate grows rapidly with the number of species combined imposing computational limits. From a practical perspective, however, indicators are often meant to avoid sampling the entire community in the field, so we do not believe this is a serious disadvantage if this is the goal.

Although they are related, the concept of indicator species with species combinations differs from that of species associations (e.g. Legendre 2005; Gotelli & Ulrich 2010). Associated species are groups of highly correlated species in one form or another, whereas a group of species that jointly provides an indicator for the target site group should not be maximally correlated, but rather complementary in their indicative power of a given group of sites. As a consequence, a species combination, as defined in the present paper, may not necessarily form a species association.

Considering species combinations does not mean that powerful indicators will magically arise for any site group (e.g. Fig. 3b). As the definition of site groups is completely left to the user, it is not easy to anticipate the cases where considering species combinations will be useful. However, our simulation results and experience with real data sets suggest that valid indicators are more likely to be obtained if the frequency of occurrence of at least some species is larger than the proportion of sites belonging to the site group. Moreover, if the beta diversity of the data set is too low (i.e. low species turnover), the species in the taxonomic group under study may perceive the environmental conditions of the site group as similar to the conditions prevailing at other sites. In this latter case, several taxonomic groups may have to be tried and compared to find valid indicators (McGeoch & Chown 1998). Finally, one may fail to obtain valid indicators because of a small sample size. Even in large data sets, valid indicators are unlikely to occur if the number of sites belonging to the site group is too small (e.g. $<$10 sites). This lack of information will normally be apparent when looking at the breadth of the confidence intervals for A and B.

HOW DOES THIS EXTENSION OF INDICATOR VALUE ANALYSIS RELATE TO OTHER APPROACHES?

Although we focused on indicator value indices here, many researchers prefer to use correlation indices, such as the *phi* coefficient, to measure the association between species and site groups, especially in a vegetation context (Chytrý *et al.* 2002; Willner, Tichý & Chytrý 2009). Our approach can be extended to correlation indices. Indeed, one can straightforwardly use matrix **C** to calculate correlation indices for species combinations. Graphs can also be easily obtained to examine the coverage of the target site group as a function of the desired level of correlation between indicators and the site group (Fig. 3). As correlation indices cannot be decomposed into components that express the predictive value of indicators and hence do not

allow one to obtain the kind of rules we were interested in, we did not explicitly describe these parallel methods here.

The generalization from single species to species combinations is also compatible with other modifications that have been proposed to overcome the limitations of the original indicator value method (e.g. Podani & Csányi 2010; Urban *et al.* 2012). For example, Urban *et al.* (2012) have recently pointed out that indicator value analyses can produce biased estimates if differences between site groups in species detection probabilities are not taken into account. Their proposal, consisting in using N-mixture models and a Bayesian inference framework, could be incorporated into our approach by replacing the input data table **X** with a 'corrected' table of abundance/occurrence produced by their method. Other extensions/modifications of the indicator value method are, by definition, less amenable to the consideration of species combinations (e.g. Baker & King 2010; De Cáceres, Legendre & Moretti 2010b). For example, De Cáceres, Legendre & Moretti (2010b) extended the indicator value method to determine whether the a given species was associated with multiple site groups instead of a single one. As the focus of that extension was to characterize the niche breadth of the species, considering species combinations does not really make sense in that context.

POTENTIAL APPLICATIONS

The use of indicator species to monitor or assess environmental conditions, or to determine habitat or community types, is a firmly established tradition for both theoretical and applied purposes. A suite of indicator variables rather than single indicators has been recommended to increase the reliability of bioindication systems (Carignan & Villard 2002). Therefore, one of the applications we envisage for the new method is the development of multispecies ecological or environmental indicators (McGeoch 1998; Niemi & McDonald 2004; Butler *et al.* 2012).

Another important application, illustrated in this paper, is motivated by the need for vegetation classification schemes to provide rules that allow new vegetation observations to be assigned to previously defined vegetation types. As vegetation types are often defined using the complete composition of vascular plants, in order to be consistent with the original classification, assignment rules should also be based on full vegetation plots (De Cáceres & Wiser 2012). When complete composition is available, there are several alternatives for assigning vegetation plot records to predefined vegetation types (e.g. Kočí, Chytrý & Tichý 2003; van Tongeren, Gremmen & Hennekens 2008; De Cáceres *et al.* 2009), which are preferable to the approach presented here. However, in many cases, vegetation surveys need to be conducted rapidly and with limited resources, such as for vegetation mapping. In such situations, it may be desirable to survey the largest possible number of localities, but simplify the fieldwork protocol by focusing on a small subset of species that have high predictive value. Our method provides a useful tool for this task. If, at a given site, one finds a species combination with high predictive value, the site can be assigned with confidence to the indicated type. If none of the valid indicators is found, then a full vegetation plot may need to be conducted.

Users of the method should bear in mind that when site groups have been defined using species composition data, they are by definition nonindependent from species. In these cases, the indicator value statistic will be larger than the value expected under the null hypothesis of independence, leading to a high rate of rejection in inferential tests (De Cáceres & Legendre 2009; De Cáceres, Legendre & Moretti 2010b). When confidence intervals are being used to assessing the uncertainty of the estimation, however, they are still valid.

Acknowledgements

MDC was supported by a postdoctoral grant (2009 BP-B 00342) from the Catalan Agency for Management of University and Research Grants. MDC and LB were supported by a research projects BIONOVEL (CGL2011-29539/BOS) and MONTES (CSD2008-00040) funded by the Spanish Ministry of Education and Science. SKW was supported by New Zealand Department of Conservation and the New Zealand Ministry of Science and Innovation (C09X0916, http://www.frst.govt.nz/). The work was also supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grant no. 7738 to PL. We acknowledge the use of data drawn from the National Vegetation Survey Databank and all of those people who wore out their boots collecting the original data. Authors would like to thank Milan Chytrý, Janós Podani and two anonymous reviewers for their enormously useful comments on previous versions of this manuscript.

References

Baker, M.E. & King, R.S. (2010) A new method for detecting and interpreting biodiversity and ecological community thresholds. *Methods in Ecology and Evolution*, **1**, 25–37.

Bakker, J.D. (2008) Increasing the utility of indicator species analysis. *Journal of Applied Ecology*, **45**, 1829–1835.

Basset, Y., Mavoungou, J.F., Mikissa, J.B., Missa, O., Miller, S.E., Kitching, R. L. & Alonso, A. (2004) Discriminatory power of different arthropod data sets for the biological monitoring of anthropogenic disturbance in tropical forests. *Biodiversity and Conservation*, **13**, 709–732.

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

Butler, S.J., Freckleton, R.P., Renwick, A.R. & Norris, K. (2012) An objective, niche-based approach to indicator species selection. *Methods in Ecology and Evolution*, **3**, 317–326.

Carignan, V. & Villard, M.-A. (2002) Selecting indicator species to monitor ecological integrity: a review. *Environmental monitoring and assessment*, **78**, 45–61.

Chazdon, R.L., Chao, A. & Colwell, R.K. (2011) A novel statistical method for classifying habitat generalists and specialists. *Ecology*, **92**, 1332–1343.

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

De Cáceres, M., Font, X. & Oliva, F. (2010a) The management of vegetation classifications with fuzzy clustering. *Journal of Vegetation Science*, **21**, 1138–1151.

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

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

De Cáceres, M. & Wiser, S.K. (2012) Towards consistency in vegetation classification. *Journal of Vegetation Science*, **23**, 387–393.

De Cáceres, M., Font, X., Vicente, P. & Oliva, F. (2009) Numerical reproduction of traditional classifications and automatic vegetation identification. *Journal of Vegetation Science*, **20**, 620–628.

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

Gotelli, N.J. & Ulrich, W. (2010) The empirical Bayes approach as a tool to identify non-random species associations. *Oecologia*, **162**, 463–477.

Hill, M.O. (1979) *TWINSPAN – A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-Way Table by Classification of the Individuals and Attributes*. Cornell University, Ithaca, NY.

Hilty, J. & Merenlender, A. (2000) Faunal indicator taxa selection for monitoring ecosystem health. *Biological Conservation*, **92**, 185–197.
Kočí, M., Chytrý, M. & Tichý, L. (2003) Formalized reproduction of an expert-based phytosociological classification: a case study of subalpine tall-forb vegetation. *Journal of Vegetation Science*, **14**, 601–610.
Legendre, P. (2005) Species associations: the Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics*, **10**, 226–245.
Manly, B.F.J. (1997) *Randomization, bootstrap and Monte Carlo methods in biology*. 2nd edn. Chapman and Hall, London.
McGeoch, M.A. (1998) The selection, testing and application of terrestrial insects as bioindicators. *Biological Reviews*, **73**, 181–201.
McGeoch, M.A. & Chown, S.L. (1998) Scaling up the value of bioindicators. *Trends in Ecology and Evolution*, **13**, 46–47.
McGeoch, M.A., Van Rensburg, B.J. & Botes, A. (2002) The verification and application of bioindicators: a case study of dung beetles in a savanna ecosystem. *Journal of Applied Ecology*, **39**, 661–672.
Murtaugh, P.A. (1996) The statistical evaluation of ecological indicators. *Ecological Applications*, **6**, 132–139.
Niemi, G.J. & McDonald, M.E. (2004) Application of ecological indicators. *Annual Review of Ecology, Evolution, and Systematics*, **35**, 89–111.
Pignatti, S. (1980) Reflections on the phytosociological approach and the epistemological basis of vegetation science. *Vegetatio*, **42**, 181–185.
Podani, J. & Csányi, B. (2010) Detecting indicator species: some extensions of the IndVal measure. *Ecological Indicators*, **10**, 1119–1124.
Qaqish, B.F. (2003) A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, **92**, 455–463.
R Development Core Team (2012) *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
van Tongeren, O., Gremmen, N. & Hennekens, S. (2008) Assignment of relevés to pre-defined classes by supervised clustering of plant communities using a new composite index. *Journal of Vegetation Science*, **19**, 525–536.
Urban, N.A., Swihart, R.K., Malloy, M.C. & Dunning, J.B. (2012) Improving selection of indicator species when detection is imperfect. *Ecological Indicators*, **15**, 188–197.
Wardle, P. (1991) *Vegetation of New Zealand*. Cambridge University Press, Cambridge, UK.
Willner, W., Tichý, L. & Chytrý, M. (2009) Effects of different fidelity measures and contexts on the determination of diagnostic species. *Journal of Vegetation Science*, **20**, 130–137.
Wiser, S.K., Bellingham, P.J. & Burrows, L.E. (2001) Managing biodiversity information : development of New Zealand's National Vegetation Survey databank. *New Zealand Journal of Ecology*, **25**, 1–17.
Wiser, S.K. & De Cáceres, M. (in press) Implementing dynamic vegetation classifications: an example with New Zealand's woody vegetation. *Journal of Vegetation Science*. doi: 10.1111/j.1654-1103.2012.01450.x.
Wiser, S.K., Hurst, J.M., Wright, E.F. & Allen, R.B. (2011) New Zealand's forest and shrubland communities: a quantitative classification based on a nationally representative plot network. *Applied Vegetation Science*, **14**, 506–523.

Received 26 June 2012; accepted 2 August 2012
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. Site group characteristics and indicator value analysis results for the 29 New Zealand's woody vegetation types.

Table S2. List of valid indicators for each of the 29 New Zealand's woody vegetation types.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.