plant biology

Plant Biology ISSN 1435-8603

RESEARCH PAPER

Selecting predictors for discriminant analysis of species performance: an example from an amphibious softwater plant

F. Vanderhaeghe[1,2], A. J. P. Smolders[3,4], J. G. M. Roelofs[3,4] & M. Hoffmann[1,2]

1 Department of Biology, Terrestrial Ecology Unit, Ghent University, Ghent, Belgium
2 Research Institute for Nature and Forest, Brussels, Belgium
3 Department of Aquatic Ecology & Environmental Biology, Institute for Water and Wetland Research, Radboud University Nijmegen, Nijmegen, The Netherlands
4 B-WARE Research Centre, Radboud University Nijmegen, Nijmegen, The Netherlands

Keywords
Data mining; goodness-of-fit test; model evaluation; overfitting; Pearson chi-square test; principal components analysis; step-wise analysis.

Correspondence
F. Vanderhaeghe, Research Institute for Nature and Forest, Kliniekstraat 25,1070 Brussels, Belgium.
E-mail: floris.vanderhaeghe@inbo.be

Editor
T. Elzenga

Received: 14 February 2011; Accepted: 23 June 2011

doi:10.1111/j.1438-8677.2011.00497.x

ABSTRACT

Selecting an appropriate variable subset in linear multivariate methods is an important methodological issue for ecologists. Interest often exists in obtaining general predictive capacity or in finding causal inferences from predictor variables. Because of a lack of solid knowledge on a studied phenomenon, scientists explore predictor variables in order to find the most meaningful (*i.e.* discriminating) ones. As an example, we modelled the response of the amphibious softwater plant *Eleocharis multicaulis* using canonical discriminant function analysis. We asked how variables can be selected through comparison of several methods: univariate Pearson chi-square screening, principal components analysis (PCA) and step-wise analysis, as well as combinations of some methods. We expected PCA to perform best. The selected methods were evaluated through fit and stability of the resulting discriminant functions and through correlations between these functions and the predictor variables. The chi-square subset, at $P < 0.05$, followed by a step-wise sub-selection, gave the best results. In contrast to expectations, PCA performed poorly, as so did step-wise analysis. The different chi-square subset methods all yielded ecologically meaningful variables, while probable noise variables were also selected by PCA and step-wise analysis. We advise against the simple use of PCA or step-wise discriminant analysis to obtain an ecologically meaningful variable subset; the former because it does not take into account the response variable, the latter because noise variables are likely to be selected. We suggest that univariate screening techniques are a worthwhile alternative for variable selection in ecology.

INTRODUCTION

Modelling the response of plant species to environmental factors is important in the methodology of many ecological studies. Several recommendations now exist regarding the choice of predictors. In their reviews on species distribution models, Elith & Leathwick (2009) and Austin (2007) point out that best results are obtained when proximal (causal) predictors are selected first, based on existing knowledge. Austin (2007) explicitly refers to a 'move away from using all possible predictors and use existing knowledge to best advantage'. This view is now largely accepted. Ecologists have been making more use of model selection criteria in order to model the response of a species to its multivariate environment (Johnson & Omland 2004; Rushton *et al.* 2004). In this approach, only those variables are sampled that are predefined in the models to be evaluated. If the models contain variables that represent the most important causal factors for a species' response, these models have a larger predictive or explanatory value (MacNally 2000; Ginzburg & Jensen 2004).

However, the model selection method has its limitations (Ginzburg & Jensen 2004; Rushton *et al.* 2004). It is often not possible to know which predictor variables will be most decisive for the response variable; for example, lack of sufficient knowledge of the species' ecology or the ecosystem (*e.g.* Van Sickle *et al.* 2006). Next, it can be challenging to find new, important variables beside the 'generally accepted' ones, like pH, nitrogen and phosphorus concentrations in the case of plant species, given a dataset with detailed environmental information (many variables). This is especially the case for (semi-)aquatic plant species, where literature on the species' environmental niche is often scattered and incomplete. For this reason, Vanderhaeghe *et al.* (2005) sampled many variables during a field survey in order to elucidate the most important predictor variables for *Eleocharis multicaulis* (Smith) Desv., an amphibious plant of west European softwater lakes. In such cases, we must turn to preliminary variable selection from a larger dataset of many potentially relevant variables before model fitting (James & McCullogh 1990; Neter *et al.* 1996; Quinn & Keough 2002). If in this way we obtain a parsimonious model with a good fit to the data, chances are high that causal factors were selected (MacNally 2000; Austin 2002). Therefore, ecologists have designed ways to reduce multivariate information (Austin 1985; James

& McCullogh 1990; Manly 1994), among which canonical ordinations (ter Braak 1995) are very popular. Through inspection of the coefficients of a canonical function, many authors have interpreted the meaning of the original variables in relation to the observed phenomenon. However, parsimony of these models has often been ignored.

In our analysis of the realised niche of *E. multicaulis* (Vanderhaeghe *et al.* 2005), we applied canonical discriminant function analysis (discriminant analysis, DA) in order to find the main predictors that distinguish between three performance categories of the species (absent, low and high cover). A major advantage of DA is that no distributional assumptions are made for the predictor variables. In this paper we explain the backgrounds of the applied variable selection. We specifically ask how variables can be selected for DA, and we therefore compared several variable selection methods. While only the best performing variable selection technique was applied and ecologically interpreted in Vanderhaeghe *et al.* (2005), the preceding comparison of variable selection methods is the subject of the current paper. Hence, no new ecological information is provided here. The results of the comparison of variable selection methods can be useful in future explorations of large multivariate datasets in which selecting the right predictors is not self-evident. An important issue in variable selection is to disentangle the web of multicollinearity among the variables in order to select those that have most potential to be causal (MacNally 2000; Graham 2003; Zuur *et al.* 2010). Williams & Titus (1988) recommend a 1:3 ratio of variables to observations as a maximum (after selection), in order to obtain narrow confidence intervals of the canonical coefficients and thus achieve a reliable interpretation.

Several methods of variable selection have been put forward. One approach is to conduct a principal components analysis (PCA) and select original variables by means of the factor loadings (Jolliffe 1972a,b; Krzanowski 1987; King & Jackson 1999). This results in an effective reduction of multicollinearity among the final predictor variables. In some studies the principal components, which are linear combinations of the original variables, are used as actual predictor variables for the response model (Manel *et al.* 2001; Graham 2003). Step-wise canonical ordinations (in contrast to direct analysis) are an alternative approach. They combine a forward selection procedure and a backward elimination procedure at each intermediate step of model fitting; mostly using P-values as criteria for entering and removing variables. This approach has been criticised because the selected subset is considered highly variable, thus dependent on the specific sample (*e.g.* Flack & Chang 1987; James & McCullogh 1990; Guisan & Zimmerman 2000; MacNally 2000; Guisan *et al.* 2002; Quinn & Keough 2002). Univariate screening of predictor variables, *e.g.* through their partial correlation with the response variable, constitutes another algorithm to obtain a subset of variables, although it has been criticised, exactly because of its univariate nature as well as for its compromised type-I error rates (MacNally 2000). Finally, hierarchical partitioning (Chevan & Sutherland 1991; MacNally 2000) quantifies the independent effect of each predictor on the response, so that they can be ranked.

In our application of DA, we compared three procedures to reduce the set of variables to enter: step-wise DA, PCA and univariate screening with Pearson chi-square calculation. In step-wise DA, the variable subset is formed during the actual DA procedure, while in the other two methods this is accomplished beforehand. The Pearson chi-square procedure consists of univariate screening of all predictors in relation to the response variable (Garson & Moser 1995). It makes no distributional assumptions regarding the predictor variables. Eventually, we performed several combinations of these three methods. The aim of the present study was to evaluate these selection procedures. From other authors' findings (see above), we expected the worst result with the step-wise method, while PCA would work best because of its ability to reduce multicollinearity.

METHODS

The dataset

The same dataset was used as in Vanderhaeghe *et al.* (2005), and stems from a field survey in summer 2001 and winter 2002. Data were collected from plots of 2 m^2 on the shores of 26 shallow softwater lakes in sandy areas of Belgium and the Netherlands. One to three plots were sampled per lake. The statistical sample consists of 46 units (plots), 232 predictor variables and one response variable, the cover of *E. multicaulis*, which is an uncommon species in the investigated region. We selected this species because it is typical of the *Eleocharition multicaulis* (Vanden Berghen 1969) alliance, and because this plant community had not previously been subject to more elaborate research. Both simple and derived variables comprise the predictor dataset (derived variables are typically ratios or summer–winter differences of simple variables). Although the sampling design implies a limited interdependence among plots, we assume that this effect can be ignored as many lakes were sampled and plots within one lake were chosen to be dissimilar and distant. For a summary of the predictor variables, see Appendix S1. The response variable was split into three classes in order to reflect the major variation of the species' response and in order to obtain enough sample units per response class. The following classes were chosen: absent (cover = 0%; 16 cases), low cover (cover 10% or less; 22 cases) and high cover (cover > 10%; eight cases). To improve the performance of PCA, six possible transformations were applied to different predictors in order to normalise them (monotone functions). A total of 101 (44%) of the 232 predictors attained a normal distribution (Kolmogorov-Smirnov test, $P > 0.05$).

Analyses were done with the statistical package SPSS 11.0 for Windows (SPSS Inc 2001).

Variable Selection

Principal components analysis

For each principal component, the variable with the highest loading was retained (method 'B4' of King & Jackson (1999)). The first p components with eigenvalues larger than those generated by the broken-stick model (Frontier 1976) were taken into account. However, when this led to $p > 15$, only the first 15 principal components were taken, in agreement with Krzanowski (1987). This was done to keep the ratio of variables to observations below 1:3, in order to achieve reliable canonical coefficients in DA.

Step-wise discriminant analysis
Discriminant analysis will be explained below. In the step-wise procedure, the probability of the F-value in ANOVA was used for inclusion and exclusion of variables, using $P < 0.05$ and $P > 0.10$ as criteria for entry and removal, respectively.

Pearson chi-square screening
For each variable, the sample was divided into three quantiles of equal frequency. In this way the frequency distribution of the response variable could be compared between the quantiles, by means of a 3×3 contingency table, one table per predictor. We acknowledge that there is no control for type-I error when performing so many successive tests. In this case, the use of chosen P-levels is merely a criterion to delimit a subset of variables and to measure their association with the response variable; it is not intended for use in statistical inference. This aspect typifies variable selection methods in general (Quinn & Keough 2002). We therefore prefer the term 'screening' instead of 'testing'. Two P-levels were used for variable selection, 0.01 and 0.05. We did not accept significant results when more than 50% of the contingency table cells contained expected frequencies of < 5. The $P < 0.01$ criterion yielded ten variables (4.3%), the $P < 0.05$ yielded 31 (13.4%). Only the $P < 0.01$ subset was used for direct DA, as the other subset contained more than 15 variables.

Combinations
Three combinations of the previously described methods were made, starting from the variable subsets obtained from the Pearson chi-square screening. These were step-wise DA, applied to both subsets (with $P < 0.01$ and $P < 0.05$, respectively), and PCA applied to the largest subset (with $P < 0.05$). We thus compared six selection methods, through the results of DA. The three variable subsets in which step-wise analysis was not involved, were used for direct DA. Because of the high number of variables, we did not perform hierarchical partitioning (Chevan & Sutherland 1991; MacNally 2000); this technique may however be effective in datasets with a shorter list of potential predictors.

Evaluation with Discriminant Analysis

In direct DA, all variables entered appear with a coefficient in the discriminant functions (DFs), which are the canonical functions in this type of ordination. In the case of three groups, two orthogonal DFs are constructed, leading to a total of 12 calculated DFs for all scenarios together. The discriminant coefficients are chosen to maximise the F-ratio of a one-way ANOVA, in which the three cover classes play the role of the grouping factor and the DF is the dependent variable. Homogeneity of variances among groups was subjectively evaluated on the basis of the ordination diagram, as proposed in Quinn & Keough (2002). To interpret the relative contribution of each original variable, standardised discriminant coefficients were used because these correspond to scaled variables with unit variance. The match between the six selected subsets was evaluated with Jaccard's similarity coefficient (Krebs 1999).

We considered three criteria as important for ecologists when using DA. First, the resulting DFs must clearly separate the response groups (model fit). This was verified by their classification success (percentage of correctly classified sample units on the basis of DF scores) and by means of the F-ratios. Second, the DFs must be stable to small changes in the sample, in order to have a general value. This was verified in two ways. We applied jack-knife classification, in which each sample unit in turn is assigned to one of the categories, based on the DFs calculated from all remaining sample units (Manel *et al.* 2001). When a large drop was observed in the classification success between standard classification and jack-knife classification, the DA is considered unreliable. Beside the jack-knife classification criterion, results were considered suspicious if the variable subset contained more than 15 variables, the maximum allowed for stable coefficients in our case (Williams & Titus 1988). Third, we considered the result acceptable only if at least one original variable was present in the DF of those that most strongly correlate with the DF. For this purpose, Spearman rank correlations between the 232 predictor variables and the 12 DFs were calculated, and for each DF the predictor variables with significant correlation ($P < 0.05$) were assigned a rank number according to correlation strength. We did not perform any type I error correction when calculating these correlations, because it was our aim to select and rank the predictors, not to make a multiple statistical inference regarding the correlations. Remember that each DF contains p predictor variables; the p predictors with lowest correlation rank (most strongly correlating) of all 232 predictors were considered and it was determined which of them are present in the DF. Furthermore, considering the p predictor variables of a DF, we wanted a close relationship between standardised coefficients and Spearman rank correlations; this was verified by the Pearson correlation coefficient between these two measures.

RESULTS

Overall similarity between the variable subsets is low, largely due to the differing number of variables selected (Table 1; see Appendix S2 for Jaccard similarities). The step-wise DA subset has too many variables to yield reliable coefficients, but we will examine the result for other properties. Most subsets originating from the chi-square selections are relatively similar.

The ability of the subsets to distinguish the cover classes can be derived from Table 1. A clear separation of classes (high F-ratio) with a perfect classification is reached with the step-wise analysis (Fig. 1A). The only other analyses that yielded good separation are the step-wise DA of the $P < 0.05$ chi-square subset (Fig. 1B) and the simple PCA selection (diagram similar to Fig. 1B).

The stability of the DFs is derived from the drop in classification success when jack-knife classification is done (Table 1). At first sight, the step-wise analysis seems the best solution (no drop); however, this result is achieved through incorporation of 24 variables in the DFs, making the coefficients unreliable. So, the most stable solutions in our case are those with a small drop in classification success: the step-wise DAs of the chi-square subset, $P < 0.01$, retaining two variables (drop: 7%), and $P < 0.05$, retaining five (drop: 8%).

Table 1. F-ratios for the discriminant functions and classification success of each discriminant analysis. A clear separation is established for the lower three scenarios.

type of analysis	number of retained variables	discriminant function	$F_{2,43}$	P	standard classification success (%)	Jack-knife classification success (%)	percentage drop[a]
chi (0.01) direct	10	DF1	18.7	<0.001	61	41	32
		DF2	0.9	0.426			
chi (0.01) steps	2	DF1	12.0	<0.001	61	57	7
		DF2	0.2	0.824			
chi (0.05) PCA direct	4	DF1	3.1	0.057	46	39	14
		DF2	0.7	0.512			
chi (0.05) steps	5	DF1	22.9	<0.001	78	72	8
		DF2	10.2	<0.001			
PCA direct	15	DF1	29.8	<0.001	65	39	40
		DF2	8.1	0.001			
steps	24	DF1	975.0	<0.001	100	100	0
		DF2	427.2	<0.001			

DF1 = first discriminant function; DF2 = second discriminant function; chi (0.01) direct = direct discriminant analysis with the chi-square subset P < 0.01; chi (0.01) steps = step-wise discriminant analysis with the chi-square subset P < 0.01; chi (0.05) PCA direct = direct discriminant analysis with the PCA-selected variables from the chi-square subset P < 0.05; chi (0.05) steps = step-wise discriminant analysis with the chi-square subset P < 0.05; PCA direct = direct discriminant analysis with PCA-selected subset; steps = step-wise discriminant analysis.
[a]Percentage is calculated relative to standard classification success.

The DFs from the chi-square derived analyses retain variable proportions of the predictors that correlate well with the DF (0–100%, see Appendix S3), with the lowest proportions occurring in the PCA-selected subset and the step-wise DA. All six analyses retain at least one of the best correlating predictors. However, the standardised DF coefficient is not always reflected by the Spearman correlation with the DF (Fig. 2). A significant relationship between the two parameters is present for the step-wise DA, the analysis of the PCA-selected subset and the step-wise DA of the P < 0.05 chi-square subset. The last analysis is marked by very high Pearson correlation coefficients between the two measures (Table 2). However, for the PCA-selected subset and the step-wise DA, the actual Spearman rank correlation coefficients of the variables retained in the DFs are generally low compared to the other variable subsets (Fig. 2; see also Appendix S3). The smaller variable subsets of the four Pearson chi-square screening approaches are shown in Table 3.

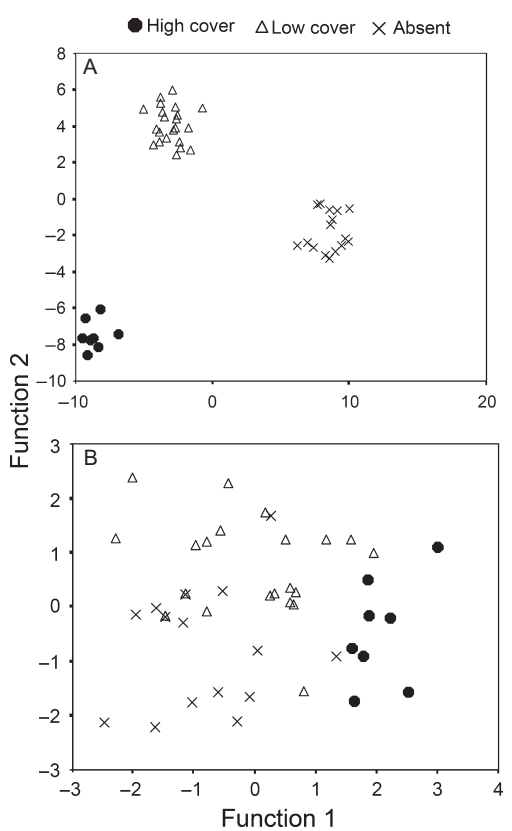

Fig. 1. Examples of ordination diagrams for two discriminant analyses. A. Step-wise discriminant analysis, starting from all 232 predictors; the cover classes are strongly separated. B. Step-wise discriminant analysis using the P < 0.05 chi-square subset; the cover classes are rather well separated.

DISCUSSION

In our study, the step-wise DA of the P < 0.05 chi-square subset combines all desirable properties: it is able to effectively discriminate between the cover classes, the analysis meets the variables to observations ratio condition that should lead to stable DF coefficient estimation, and the interpretation of the coefficients is supported by the pattern of Spearman rank correlations. At least one of these criteria is not accomplished in any of the other methods. In particular, our study suggests failure of the popular step-wise procedure and, in contrast with our expectations, also the PCA selection procedure. Moreover, the causal relations of the chi-square selected predictor variables with the performance of the plant species can be ecologically explained (see Vanderhaeghe *et al.* 2005), whereas several variables retained in the step-wise and PCA procedure would be difficult to explain in a plant ecological context [*e.g.* lake surface or water colour (absorption at 450 nm)]. Thus, the models we compared statistically also differ in ecological meaningfulness. Austin (2007) encourages the use of this criterion when evaluating models.

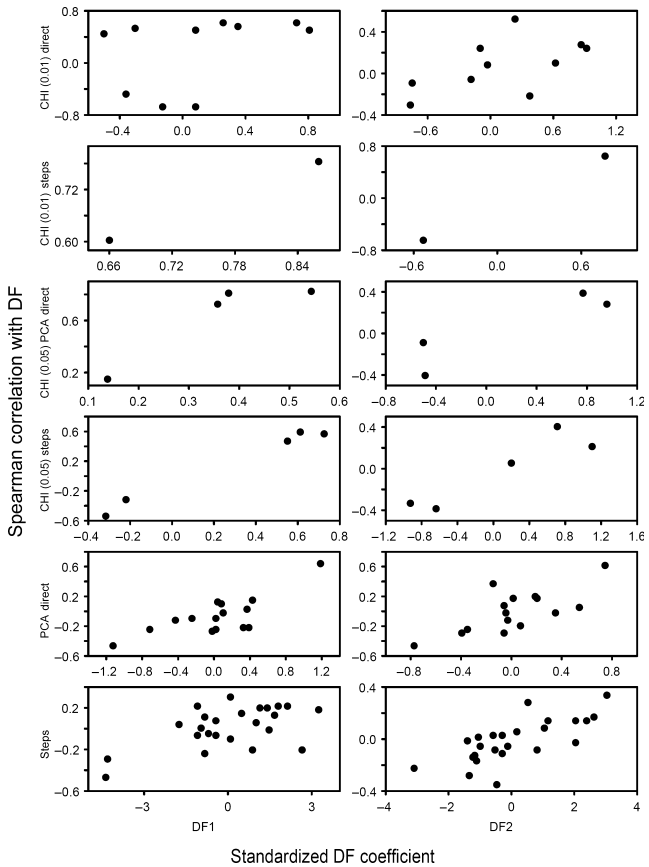

Fig. 2. Graphic relationship between standardised discriminant function coefficients and the corresponding Spearman rank correlation between the retained predictors and the discriminant function. Standardised discriminant function coefficients are not strictly associated with the Spearman rank correlation between the retained predictors and the discriminant function. Only for the lowest three analyses (six graphs) was there a significant correlation between the two measures ($P < 0.05$). Abbreviations as in Table 1.

Table 2. Pearson correlation coefficients for each discriminant function, between the standardised function coefficients and the Spearman rank correlation coefficient of the corresponding predictors with the discriminant function.[a] Only the bold results are significant ($P < 0.05$).

	pearson correlation	
type of analysis	DF1	DF2
chi (0.01) direct	0.40	0.57
chi (0.01) steps	1.00	1.00
chi (0.05) PCA direct	0.90	0.91
chi (0.05) steps	**0.99**	**0.92**
PCA direct	**0.76**	**0.72**
steps	**0.57**	**0.73**

[a]Abbreviations as in Table 1.

James & McCullogh (1990) and Quinn & Keough (2002) summarise the criticisms against the use of step-wise procedures in linear methods. Step-wise procedures attempt to maximise the percentage of variation accounted for by the linear function. However, meaningless variables are likely to be selected to serve this purpose, as shown in a simulation study by Flack & Chang (1987). Our results support these findings. Although the step-wise selection method gave promising results, the rather poor relationship with the Spearman correlations, the low involvement of predictors that correlate well with the DFs and the high number of selected variables make this model unlikely to be generally applicable. Similarly, Van Sickle *et al.* (2006) compared step-wise DA with best subsets DA, the latter being an approach not considered by us. These authors came to similar conclusions regarding step-wise DA and they equally discourage its use.

Multicollinearity is a problem of all methods in which linear combinations of predictor variables are involved (Graham 2003). Especially when scientists are interested in the causal (explanatory) value of a model, rather than in its predictions, multicollinearity brings the risk of retaining non-causal predictor variables that are correlated with a causal, but discarded, variable. Multicollinearity is the main reason that several authors recommend interpretation by means of correlations with the canonical function (linear function) instead of the actual canonical coefficients (*e.g.* Manly 1994; Quinn & Keough 2002). However, in observational studies an orthogonal design is rarely present, in which one predictor varies while the others remain constant (Johnson & Omland 2004). Hence, the additive effect of each predictor variable can only be interpreted from the canonical coefficients, as suggested by Rencher (1988, 1992), Williams & Titus (1988), Tardif & Hardy (1995) and MacNally (2000). We have therefore used the presence of the best correlating predictors in the canonical function and a good relationship between correlations and coefficients as model eligibility criteria.

Retaining the variables with the highest loading in PCA reduces the multicollinearity problem, because this selection method effectively excludes redundant variables. King & Jackson (1999) selected variables from a climate dataset using PCA, in order to conduct a canonical correlation analysis between the climate data and data on lake thermal stratification. However, selection by means of PCA does not take into account the direct relevance of the predictor variables to the response variable(s), implying a potential for withholding less meaningful, but less redundant, predictor variables (Graham 2003). A similar reasoning was made by ter Braak (1995, p. 136), in the context of correspondence analysis. The low jack-knife classification success in our case indicates that PCA did not mark the necessary variables for prediction of the performance of *E. multicaulis*. The algorithm of PCA subset selection contrasts with that of the chi-square screening approaches, in which the direct relation of each predictor with the response variable is the first criterion to retain or reject a predictor variable.

Univariate chi-square screening before embarking on any multivariate analysis turned out to be a very satisfactory method. Conducting a step-wise procedure with these variables (subset $P < 0.05$) does not run the risk of obtaining an ecologically less meaningful variable subset. On the contrary, it indicates their predictive ability through further selection and assigning coefficients. It is probable that the rather low falls in classification success, when jack-knife classification is performed (Table 1), are due to the selection of the ecologi-

Table 3. Retained variables of the Pearson chi-square screening approaches, marked with their maximum absolute Spearman rank correlation (with either DF1 or DF2) when they belonged to the most strongly correlated predictors, and otherwise with 'X'. The PCA-selected subset and the step-wise DA subset are not shown in full (see Appendix S4); they mainly do not coincide with the chi-square subsets shown here, and contain few highly correlated predictors, of which most belong to chi-square subsets.[a]

covariate[b]	chi (0.01) direct	chi (0.01) steps	chi (0.05) PCA direct	chi (0.05) steps	PCA direct	steps
number of retained variables	10	2	4	5	15	24
seasonality of mineral soil layer Si concentration nacl extraction	0.67					
mineral soil layer winter Si concentration nacl extraction	0.65					
seasonality of mineral soil layer total N content	0.52	0.66		X		
mineral soil layer winter total N content	0.55					
seasonality of mineral soil layer K concentration	0.58					
seasonality of surface water ammonium/nitrate concentration	0.64	0.79		0.60		0.45
mineral soil layer winter Si concentration	0.64					
cover *Juncus bulbosus*	X					
surface water summer ion ratio (IR)	X					
surface water winter Cl proportion	X					
surface water winter K concentration				0.52	0.45	
seasonality of surface water divalent/monovalent cation ratio				0.57	0.65	0.34
cover *Mentha aquatica*				X	0.44	
cover *Agrostis canina*			X			
mineral soil layer summer Mg concentration			X			
surface water summer Mg concentration			0.81			
surface water winter pH			0.82			

[a]Abbreviations as in Table 1.

[b]Seasonality variables are the difference between summer and winter values. Other variables are always confined to winter or summer conditions, with cover values for the summer period. See Vanderhaeghe *et al.* (2005) for technical aspects and ecological interpretation of results.

cally most meaningful variables beforehand. In general, however, by excluding variables that are not significant in univariate analysis, there is a risk of losing actually causal variables when their effects cancel each other out in the specific dataset (MacNally 2000). Although this is theoretically possible, we expect that the ecologically most significant predictors will, in most cases, be significant on the univariate level when the sample is not too small.

In conclusion, we agree with other authors that purely step-wise methods are not recommended for achieving a good explanatory ecological model when starting from many predictor variables. From our results, ecologists should be prudent when using PCA subset selection. PCA can be used if the only purpose is to obtain a limited dataset that still contains much of the variation of the original data. PCA, however, will not necessarily withhold the important variables for the interpretation of an extra phenomenon. In general, we suggest evaluating any selected variable subset by means of the efficacy and credibility of the obtained results from the analysis of real interest (discriminant analysis in our case), using objective statistical criteria as well as the ecological interpretability and credibility of the models, thereby supporting the view of Austin (2007). The univariate evaluation of variables in relation to the response variable is a method with potential, *e.g.* chi-square screening in the case of a categorical response variable. Subsequent selection with one of the previous methods (PCA or step-wise) can prove useful. Other methods that we did not consider, *e.g.* best-subsets comparison and hierarchical partitioning, may also be useful. There is clearly a need for simulation studies on these subjects, so that more generally applicable conclusions can be drawn than presently possible from the ecological literature.

ACKNOWLEDGEMENTS

We are grateful to the many scientists, owners and nature managers who gave us access to the softwater lake study sites. Special thanks to Luc Lens, Beatrijs Bossuyt, Leon van den Berg, Nigel Yoccoz and Jan van Groenendael for useful comments that improved the manuscript. The first author did part of this work as Research Assistant of the Fund for Scientific Research – Flanders (Belgium) (F.W.O.-Vlaanderen).

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix S1. The number of different types of predictor measured.

Appendix S2. Overall characteristics of the six variable subsets.

Appendix S3. Match between the variable subsets and the variables that best correlate with the discriminant functions.

Appendix S4. Retained variables in all approaches.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

REFERENCES

Austin M.P. (1985) Continuum concept, ordination methods, and niche theory. *Annual Review of Ecology and Systematics*, **16**, 39–61.

Austin M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Austin M.P. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

ter Braak C.J.F. (1995) Ordination. In: Jongman R.H.G., ter Braak C.J.F., van Tongeren O.F.R. (Eds), *Data analysis in community and landscape ecology*. Cambridge University Press, Cambridge, UK, pp 91–173.

Chevan A., Sutherland M. (1991) Hierarchical partitioning. *The American Statistician*, **45**, 90–96.

Elith J., Leathwick J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology and Systematics*, **40**, 677–697.

Flack V.F., Chang P.C. (1987) Frequency of selecting noise variables in subset regression analysis: a simulation study. *The American Statistician*, **41**, 84–86.

Frontier S. (1976) Etude de la décroissance des valeurs propres dans une analyse en composantes principales: comparaison avec le modèle du bâton brisé. *Journal of Experimental Marine Biology and Ecology*, **25**, 67–75.

Garson G.I., Moser E.B. (1995) Aggregation and the Pearson chi-square statistic for homogeneous proportions and distributions in ecology. *Ecology*, **76**, 2258–2269.

Ginzburg L.R., Jensen C.X.J. (2004) Rules of thumb for judging ecological theories. *Trends in Ecology and Evolution*, **19**, 121–126.

Graham M.H. (2003) Confronting multicollinearity in ecological multiple regression. *Ecology*, **84**, 2809–2815.

Guisan A., Zimmerman N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Guisan A., Edwards T.C. Jr, Hastie T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100.

James F.C., McCullogh C.E. (1990) Multivariate analysis in ecology and systematics: Panacea or Pandora's box? *Annual Review of Ecology and Systematics*, **21**, 129–166.

Johnson J.B., Omland K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

Jolliffe I.T. (1972a) Discarding variables in a principal components analysis I: artificial data. *Applied Statistics*, **21**, 160–173.

Jolliffe I.T. (1972b) Discarding variables in a principal components analysis II: real data. *Applied Statistics*, **22**, 21–31.

King J.R., Jackson D.A. (1999) Variable selection in large environmental data sets using principal components analysis. *Environmetrics*, **10**, 67–77.

Krebs C.J. (1999) *Ecological methodology*, 2nd edition. Addison Wesley Longman, New York, USA.

Krzanowski W.J. (1987) Selection of variables to preserve multivariate data structure using principal components. *Applied Statistics*, **36**, 22–33.

MacNally R. (2000) Regression and model-building in conservation biology, biogeography and ecology: the distinction between – and reconciliation of – 'predictive' and 'explanatory' models. *Biodiversity and Conservation*, **9**, 655–671.

Manel S., Williams H.C., Ormerod S.J. (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.

Manly B.F.J. (1994) *Multivariate statistical methods: a primer*, 2nd edition. Chapman & Hall, London, UK.

Neter J., Kutner M.H., Nachtsheim C., Wasserman W. (1996) *Applied linear statistical models*, 4th edition. McGraw-Hill, New York, NY, USA.

Quinn G.P., Keough M.J. (2002) *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge, UK.

Rencher A.C. (1988) On the use of correlations to interpret canonical functions. *Biometrika*, **75**, 363–365.

Rencher A.C. (1992) Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician*, **46**, 217–225.

Rushton S.P., Ormerod S.J., Kerby G. (2004) New paradigms for modelling species distributions. *Journal of Applied Ecology*, **41**, 193–200.

SPSS Inc (2001) *SPSS for Windows.* Release 11.0.1. SPSS Inc., IL, USA.

Tardif B., Hardy J. (1995) Assessing the relative contribution of variables in canonical discriminant analysis. *Taxon*, **44**, 69–76.

Van Sickle J., Huff D.D., Hawkins C.P. (2006) Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates. *Freshwater Biology*, **51**, 359–372.

Vanderhaeghe F., Smolders A.J.P., Ruysschaert S., Roelofs J.G.M., Hoffmann M. (2005) Understanding the realised niche of an amphibious softwater plant, *Eleocharis multicaulis*. *Archiv für Hydrobiologie*, **163**, 329–348.

Williams B., Titus K. (1988) Assessment of sampling stability in ecological applications of discriminant analysis. *Ecology*, **69**, 1275–1285.

Zuur A.F., Ieno E.N., Elphick C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14.