COMMUNITY ECOLOGY 11(2): 179-186, 2010
1585-8553/$20.00 © Akadémiai Kiadó, Budapest
DOI: 10.1556/ComEc.11.2010.2.6

Identifying the drivers of pond biodiversity: the agony of model selection

M. Gioria[1,3], G. Bacaro[2] and J. Feehan[1]

[1]*School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland*
[2] *BIOCONNET, Biodiversity and Conservation Network, Department of Environmental Science "G. Sarfatti", University of Siena, Via P.A. Mattioli 4, 53100 Siena, Italy*
[3] *Corresponding author. E-mail: margherita.gioria@ucd.ie*

Keywords: Forward selection, Multivariate analysis, Species richness, Water beetle, Wetland plant.

Abstract: Ponds contribute substantially to the maintenance of regional biodiversity. Despite a growing body of literature on biotic-abiotic relationships in ponds, only few generalizations have been made. The difficulty in identifying the main drivers of pond biodiversity has been typically attributed to the heterogeneity of the local and regional conditions characterizing ponds. However, little is known on how the use of different analytical approaches and community response variables affects the results of analysis of community patterns in ponds. Here, we used a range of methods to model the response of water beetle and plant community data (species richness and composition) to a set of 12 environmental and management variables in 45 farmland ponds. The strength of biotic-abiotic relationships and the contribution of each variable to the overall explained variance in the reduced models varied substantially, for both plants and beetles, depending on the method used to analyze the data. Models of species richness included a lower number of variables and explained a larger amount of variation compared to models of species composition, reflecting the higher complexity characterizing multispecies response matrices. Only two variables were never selected by any of the model, indicative of the heterogeneity characterizing pond ecosystems, while some models failed to select important variables. Based on our findings, we recommend the use of multiple modeling approaches when attempting to identify the principal determinants of biodiversity for each response variable, including at least a non-parametric approach, as well as the use of both species richness and composition as the response variables. The results of this modeling exercise are discussed in relation to their practical use in the formulation of conservation strategies.

Abbreviations: AIC–Akaike Information Criterion; BIC–Bayesian Information Criterion; CCA–Canonical Correspondence Analysis; FP–Forward Procedure of variable selection; GLM–Generalized Linear Model; DISTLM–Distance-based Linear regression model; BIO-ENV–Permutational multivariate model of biotic-abiotic relationships; PERMANOVA–PERmutational Multivariate ANalysis Of VAriance; VIF–Variance Inflation Factor.

Introduction

Ponds play a central role as reservoirs of biodiversity regionally, supporting a disproportionately large number of species relative to their surface area, particularly within the agricultural matrix, where they may act as biodiversity 'hotspots' (Céréghino et al. 2008), supporting uncommon, unique, and rare species (Williams et al. 2004, Gioria et al. 2010). The conservation of ponds ultimately relies on a solid understanding of the principal determinants of pond biodiversity, which is a function of local and regional variables as well as of processes of dispersal and speciation, land use history, and degree of habitat patchiness (Heino 2000, Wood et al. 2003, Céréghino et al. 2008).

To date, extensive research efforts have focused on evaluating biotic-abiotic relationships in ponds. These studies have highlighted the importance of factors such as eutrophication, connectivity, and hydro-period in determining patterns of pond biodiversity (Heino 2000, Fairchild et al. 2003, Céréghino et al. 2008). Contrasting patterns have, however, emerged while assessing the relationship between pond community patterns and factors such as pond size, pond age, or water chemical gradients (e.g., Gee et al. 1997, Heino 2000, Oertli et al. 2002, Gioria et al. 2010).

The difficulties in making generalizations on the main drivers of patterns in pond biodiversity have often been attributed to 1) the heterogeneity in the biotic and biotic conditions that characterize each pond, 2) random colonization events, and 3) the tolerance of many taxa to large abiotic gradients (Heino 2000, Jeffries 2008). Although there is no doubt that certain environmental and habitat conditions play a critical role in determining pond biodiversity, differences in the approach used to analyze community patterns may have major effects on the outcomes of any ecological study.

Ecologists may employ a wide range of analytical approaches to assess biotic-abiotic relationships and new methods are occasionally proposed to address the difficulties associated with the analysis of species data or the drawbacks of previously-developed models (e.g., Clarke 1993, Guisan and Zimmerman 2000, Thuiller 2003, Blanchet et al. 2008). In the field of aquatic ecology, a remarkably different approach to the analysis of community data has been adopted by ma-

rine and freshwater ecologists over the years, despite the high similarity in the complexity and nature of such data. Freshwater ecologists have made extensive use of canonical correspondence analysis (CCA; ter Braak 1986) to evaluate biotic-abiotic relationships in ponds and lakes (e.g., Nicolet et al. 2004, Studinski and Grubbs 2007), despite the use of CCA to analyze community data has long been discouraged due to the implicit use of chi-square distances as the measure of dissimilarity between pairs of samples (see e.g., Clarke 1993, Legendre and Legendre 1998). The use of multivariate non-parametric or semi-metric approaches has therefore been strongly recommended (Clarke 1993, Anderson 2001). These methods, including analysis of similarities (ANOSIM, Clarke 1993), the BIO-ENV procedure proposed by Clarke and Ainsworth (1993), and permutational multivariate analysis of variance (PERMANOVA, Anderson 2001), have been widely applied to the analysis of marine community structure. In addition to these methods, Mantel tests (Mantel 1997) and Procrustean randomization tests (Gower 1971, Jackson 1995, see Peres-Neto and Jackson 2001 and references therein) have also been applied to the analysis of community patterns in aquatic ecosystems.

The aim of this paper is to compare the results of a number of methods that have been applied widely to the analysis of community patterns by using data from 45 farmland ponds in Ireland (see Gioria et al. 2010). To characterize and quantify the potential differences in the outcomes of these analytical approaches, we modelled the relationship between a range of environmental variables and community patterns (species richness and species composition) for water beetles and wetland plants. These taxa have been commonly used to evaluate biodiversity in ponds and tend to respond in a similar way to the same set of environmental variables (see Gioria et al. 2010 and references therein). Vascular plants are considered to be a good surrogate taxon for invertebrate biodiversity, in both terrestrial and aquatic systems, since they are sensitive to environmental changes and are characterized by a well-described ecology and taxonomy (e.g., Rodwell 1995, Sætersdal et al. 2003, Schaffers et al. 2008). In ponds, the use of water beetles as surrogates for invertebrate biodiversity and as indicators of anthropogenic disturbance has also been recommended (e.g., Foster et al. 1992, Menetrey et al. 2005, Bilton et al. 2006).

Specifically, we aimed at evaluating and comparing potential differences in: 1) the number of variables selected in the reduced models; 2) the contribution of each selected variable to the variance in the study response variable; 3) the overall variance explained by the reduced models; and 4) models of species richness *versus* model in species composition. This information is critical to a sound interpretation of any theoretical model and to the development of effective management practices and conservation programs.

Material and methods

To perform this modeling exercise, we used data from 45 permanent ponds located in two intensively farmed regions in Ireland (Wexford, 52°23'N, 6°23'W; Mullingar, 53°33'N, 7°25'W, see Gioria et al. 2010 for details on the sampling protocol). Beetle community data consist of species abundance data, while vegetation data are expressed as percentage cover. To model vegetation data, we used three categorical variables: grazing intensity (grazed, fenced, ungrazed); pond dominant substratum (mud, gravel); and pond age (2-10 years, >10 years), as well as nine continuous variables: pond surface area; maximum pond depth; maximum depth of sampling; conductivity; pH; alkalinity; ammonia (NH_3-N); and nutrients (NO_3-N, PO_4-P). To model beetle data, we used two additional explanatory variables: plant species richness and plant vegetation cover over the sampled area (Table 1).

Modeling species richness

The relationship between species richness and the sets of explanatory variables was investigated, separately for plants and beetles, using three procedures: 1) a generalized linear model (GLM; McCullagh and Nelder 1989); 2) a parametric multiple regression model based on a permutational forward selection procedure (FP; Blanchet et al. 2008); and 3) non-parametric distance-based multivariate analysis for linear models (DISTML; Anderson 2001, McArdle and Anderson 2001).

Species richness and environmental data were log-transformed prior to data analyses. Multi-collinearity between pairs of variables was examined using the Pearson's correlation coefficient (r) as well as variance inflation factor analysis (VIF; Montgomery and Peck 1982), which was performed on the full models. The maximum correlation between variables was always below 0.60, for both response variables (beetle and plant species richness), and individual VIF values were never above 10. We therefore retained all the explanatory variables in the full models.

For the GLM, we used the Poisson model family to model species richness (Guisan and Zimmermann 2000), applying the log-link function to relate the mean value of the response variables to their linear explanatory variables (Crawley 1993). The reduced model was constructed by following a number of steps. First, we computed the deviance for the null model to calculate the value of the intercept, and we constructed a full model using all the explanatory variables to quantify the total variance in the response variable explained by all the explanatory variables. We then used an iterative stepwise (backward and forward) procedure to identify the variables to be included in the reduced model (see Guisan and Zimmermann 2000). We used the Akaike Information Criterion (AIC; Hastie and Pregibon 1993) and the Bayesian Information Criterion (BIC; Schwarz 1978) as the model selection criteria (see Burnham and Anderson 2004 for a review of model selection criteria). The χ^2 statistic was used to test the significance of each variable retained in the reduced model ($\alpha = 0.05$, see Crawley 1993). The goodness-of-fit of the selected model was evaluated using the adjusted deviance D^2 (Guisan and Zimmermann 2000) and the χ^2 test was performed to evaluate whether there were statistically

Table 1. List of explanatory variables used to model a) water beetle and b) plant species richness and composition. The asterisk indicates the variables that were only used to model beetle community patterns.

Variable	Unit
Continuous	
Pond surface area	m^2
Pond maximum depth	m
Maximum sampling depth	m
pH	
Conductivity	microS/cm
Alkalinity	mgL^{-1}
NH_3-N	μgL^{-1}
NO_3-N	μgL^{-1}
PO_4-P	μgL^{-1}
Plant species richness*	
Percentage vegetation cover*	%
Categorical (dummy)	
Grazing intensity	three classes
Grazed	
Ungrazed	
Fenced	
Pond substratum	two classes
Mud	
Gravel	
Pond Age	two classes
2-10 years	
> 10 years	

significant differences between the full and the reduced model. We then performed a weighted analysis of deviance to evaluate the performance of the selected model in predicting the relationship between species richness and the predictor variables (McCullagh and Nelder 1989). To calculate the prediction error for the GLMs, we performed a leave-one-out cross-validation (Guisan and Zimmermann 2000). Finally, we calculated the coefficient of correlation (r) between the values predicted by the reduced model and the observed values of the response variables.

The forward selection procedure proposed by Blanchet et al. (2008) was here used since it was developed to prevent two well-known problems associated with the use of classic forward selection: 1) the overestimation of the explained variance and 2) an inflated Type I Error (Blanchet et al. 2008). This procedure is based on two stopping criteria: 1) the significance level α (here set at 0.05) and 2) an adjusted coefficient of determination (maximum R^2_{adj}), calculated by constructing a full model inclusive of all explanatory variables. This procedure performs a forward selection by permutation of residuals under the reduced model. When forward selection identifies a variable that brings one or the other criterion over the fixed threshold, that variable is rejected, and the procedure is stopped. One of the advantages of this method is that the selection of useless variables occurs less often and fewer variables are selected (Blanchet et al. 2008).

DISTLM is a non-parametric procedure that performs a distance-based analysis on a linear model for any dissimilarity matrix (McArdle and Anderson 2001). The purpose of DISTLM is to perform a permutational test for the multivariate null hypothesis of no relationship between two matrices on the basis of any distance measure of choice, using permutations of the observations. Here, we applied a forward selection of the predictor variables, running 9999 tests by permutation (α = 0.05). The Euclidean distance was used as the measure of dissimilarity between pairs of samples for log-transformed species richness data.

Modeling species composition

We used four analytical procedures to model patterns in species composition (sample × species matrices): 1) classical canonical correspondence analysis (CCA), using a stepwise variable selection; 2) a forward selection procedure described by Blanchet et al. (2008); 3) BIO-ENV (Clarke and Ainsworth 1993); and 4) DISTLM (forward selection of the predictor variables, 9999 permutations). Environmental variables were log-transformed prior to multivariate data analyses, while plant and beetle species composition data were log(x+1)-transformed.

Two measures of dissimilarity were applied to species composition data in non-parametric models: 1) the Bray-Curtis dissimilarity measure (d_{BC}; Bray and Curtis 1957) and 2) a modified Gower distance (d_{MG}) proposed by Legendre and Legendre (1998). The d_{BC} was selected due to its useful properties in the analysis of community data and its wide use in ecological studies (Clarke 1993, Legendre and Legendre 1998), while the d_{MG} was used since it has the advantage of being explicit about the contribution of differences in species identity and relative abundances (here, such differences were assigned weight: 1).

BIO-ENV is permutational procedure that aims at identifying the combination of environmental variables that maximizes the correlation between a biotic and an environmental data matrix (Clarke and Ainsworth 1993). We used the Spearman rank coefficient of correlation (ρ_s) as the measure of correlation between biotic and environmental variables. We used Euclidean distance as the measure of dissimilarity between pairs of samples for environmental data, and we performed 9999 random permutations of all combinations of variables.

Prior to performing these analyses, we used permutational multivariate analysis of variance (PERMANOVA; Anderson 2001) to test whether the factor 'region' (two levels) had any significant effect on beetle and plant assemblages. Since the effects of regional differences were not significant, this factor was not included in the analyses. GLM, FP, and CCA models were performed using the R software 2.10.1 (R Development Core Team 2010, see electronic supplement), while DISTML, BIO-ENV, and PERMANOVA procedures were performed using the statistical package PRIMER v.6 and PERMANOVA+ (Clarke and Warwick 2001, Anderson et al. 2008).

Results

The relationship between environmental variables and biotic communities varied from weak to strong, depending on 1) the response variable (species richness *versus* species

Table 2. Summary of the results of GLM (Poisson family, log-link function, for 1) beetle and 2) plant communities (model selection criterion: a) AIC, b) BIC). The variables grazing intensity and substratum are categorical.

Variable	Deviance reduction	Coefficient value	VIF	$p(\chi^2)$
1a) Beetle species richness, AIC, D^2_{adj} = 0.7274				
alkalinity	64.31	-	1.94	***
substratum	28.48		1.27	***
max depth	19.01	+	1.47	***
grazing	8.27		1.87	*
pH	4.71	+	1.10	*
Cross-validation estimate of prediction error: 19.5				
r = 0.8872, difference between full and minimal GLM, $p(\chi^2)$ = 0.969				
1b) Beetle species richness, BIC, D^2_{adj} = 0.697				
alkalinity	64.31	-	1.215	***
max depth	22.75	+	1.149	***
substratum	27.97		1.170	***
Cross-validation estimate of prediction error: 20.81				
r = 0.8872, difference between full and minimal GLM, $p(\chi^2)$ = 0.518				
2a) Plant species richness, AIC, D^2_{adj} = 0.5650				
pond age	29.21		1.63	***
pond area	20.62	+	1.45	***
max sampling depth	14.93	+	1.97	***
substratum	14.07		1.37	***
alkalinity	13.99	-	3.77	***
Grazing	9.64		2.22	**
Cross-validation estimate of prediction error = 37.9				
r = 0.8060, Difference between full and minimal GLM, $p(\chi^2)$ = 0.973				
2b) Plant species richness, BIC, D^2_{adj} = 0.5384				
pond area	35.56	+	1.146	***
max sampling depth	41.58	+	1.075	***
substratum	14.41		1.072	***
Cross-validation estimate of prediction error = 21.39				
r = 0.8060, Difference between full and minimal GLM, $p(\chi^2)$ = 0.218				

* = p <0.05; ** = p<0.01; *** = p<0.001

Table 3. Summary of results of DISTLM analyses used to model the relationship between the explanatory variables and species richness for a) beetle and b) plant communities. Data were log-transformed and the Euclidean distance was applied as a measure of dissimilarity between samples (9999 permutations).

DISTLM	variance	Cumulative variance	F	P
a) Beetle species richness, R^2 = 0.7304				
alkalinity	0.3828	0.3828	26.669	0.001
Area	0.1681	0.5509	15.718	0.001
Mud	0.1022	0.6530	12.073	0.001
max depth	0.0430	0.6960	5.656	0.026
pH	0.0343	0.7304	4.967	0.030
b) Plant species richness, R^2 = 0.5882				
max sampling depth	0.4165	0.4165	30.689	0.001
Gravel	0.1143	0.5308	10.231	0.005
Area	0.0574	0.5882	5.714	0.032

composition), 2) the model (three models of species richness, four models of species composition), 3) the measure of dissimilarity used in the non-parametric multivariate approaches, and on 4) the model selection criterion in the GLM and the DISTLM.

As for patterns of species richness, the GLM and DISTLM models explained a similar amount of variation (70-73% for beetles and 53-56% for plants), despite selecting a different number of variables and the different contribution percentage of each selected variable to the overall explained variance (Table 2 and 3). The FP model selected a lower number of variables compared to the GLM and the DISTLM, and explained a lower amount of variation (~63% and ~48% of variation in beetle and plant species richness, respectively, Table 4). The model selection criterion in the GLM (AIC *versus* BIC) had only a marginal effect on the overall explained variance, while a substantial difference was evident in the number of variables included in the reduced models. In the DISTLM models, the use of the BIC as compared to the AIC resulted in the selection of a lower number of variables, each of which explained the same amount of variance (Table 3). The same is true when species composition data were used as the response variable (Table 5).

Models of species composition explained a lower percentage of variation compared to those of species richness (Table 4 and 5). When modeling species composition, the FP, the DISTLM_AIC), and the BIO-ENV models selected a similar number of variables, although the FP reduced models explained a lower percentage of variance, particularly for plant species composition. The CCA models retained only two significant variables in both communities (Table 5); the use of the AIC versus the BIC did not affect the models.

In the DISTLM_BIC model, only 1-2 variables were selected, compared to up to seven variables included in the DISTLM_AIC. The choice of the dissimilarity measures (d_{BC} *versus* d_{MG}) in the non-parametric models (BIO-ENV and DISTLM) affected the overall explained variance in species composition, for both beetle and plant data, but did not affect the identity of the selected variables or their order of contribution to the explained variance (Table 5).

All the models included depth-related variables and type of substratum among the most significant determinants of species composition patterns, for both plant and beetle communities. The variable 'maximum sampling depth' was selected as the most important driver of patterns in beetle and plant species composition by the DISTLM_AIC and the BIO-ENV models, while the FP model of beetle species composition and CCA (both for plants and beetles) emphasized the contribution of 'maximum pond depth' to the total explained variance.

Discussion

Biotic-abiotic relationships in the study ponds ranged from weak to strong, for both plant and beetle communities. The use of different analytical approaches and response variables (species richness *versus* species composition) affected the overall variance explained by the reduced models, the identity of the variables selected in the reduced models, the total number of retained variables, and the contribution of the selected environmental variables to the explained variance.

Models of species richness explained a larger percentage of variation compared to models of species composition, showing a strong relationship between beetle species richness and environmental variables, while a moderate relationship was found for plant species richness. The lower variance

Table 4. Summary of the results of multiple regression models based on a modified forward selection procedure (Blanchet et al. 2008) for a) beetle and b) plant community data.

Variable	R^2_{adj}	cumulative R^2_{adj}	*F*	*P*
a) Beetle community				
Species richness				
alkalinity	0.3938	0.3938	27.934	0.0001
plant species richness	0.1458	0.5396	13.296	0.0006
gravel	0.0588	0.5690	6.009	0.0193
max depth	0.0687	0.6338	8.254	0.0067
Species composition				
max depth	0.1226	0.1226	7.148	0.0001
Mud	0.1176	0.2402	7.658	0.0001
plant species richness	0.0373	0.2776	3.171	0.0001
PO_4-P	0.0225	0.3000	2.316	0.0011
alkalinity	0.0156	0.3157	1.914	0.0071
pH	0.0160	0.3317	1.932	0.0066
b) Plant community				
Species richness				
max sampl. depth	0.3667	0.3667	24.901	0.0001
pond area	0.1115	0.4782	8.9763	0.0045
Species composition				
max sampling depth	0.0913	0.0913	5.423	0.000
Mud	0.0361	0.1274	2.777	0.000
pond area	0.0244	0.1518	2.207	0.002
conductivity	0.0193	0.1710	1.953	0.010
age >10	0.0130	0.1840	1.637	0.038

Table 5. Summary of results of DISTLM, BEST, and CCA to model the relationship between the explanatory variables and species composition for a) beetle and b) plant communities. The asterisk indicates the variables that were selected using the BIC criterion of model selection (maxSdepth = maximum sampling depth). Dissimilarity measure: BC = Bray-Curtis dissimilarity; MG = modified Gower distance.

DISTLM	*F*	*P*	variance	cumul. variance	DISTLM	*F*	*P*	variance	cumul. variance
a) Beetle species composition					a) Beetle species composition				
AIC_BC					*AIC_MG*				
maxSdepth*	6.202	0.001	0.1236	0.1236	maxSdepth*	5.150	0.001	0.1048	0.1048
mud*	5.763	0.001	0.1036	0.2271	mud*	4.899	0.001	0.0916	0.1963
PO_4-P	2.485	0.008	0.0432	0.2703	PO_4-P	2.059	0.002	0.0376	0.2339
conductivity	2.694	0.002	0.0450	0.3153	max depth	2.130	0.001	0.0378	0.2717
Alkalinity	1.972	0.020	0.0322	0.3475	conductivity	1.924	0.004	0.0334	0.3051
max depth	1.944	0.031	0.0310	0.3784					
pH	1.895	0.036	0.0295	0.4080					
b) Plant species composition					b) Plant species composition				
AIC_BC					*AIC_MG*				
maxSdepth*	7.903	0.001	0.1523	0.1523	maxSdepth*	4.442	0.001	0.0917	0.0917
Mud	3.618	0.002	0.0658	0.2180	mud	2.510	0.001	0.0501	0.1418
Area	2.560	0.004	0.0449	0.2630	area	1.877	0.012	0.0367	0.1785
conductivity	2.507	0.003	0.0425	0.3054	conductivity	1.919	0.006	0.0367	0.2152

BEST analysis		CCA	χ^2	*F*	*P*
a) Beetle species composition		a) Beetle species composition			
ρ_{BC} = 0.381	ρ_{MG} = 0.481	χ^2 = 0.3842			
maxSdepth	max sam. depth	max depth	0.1910	3.3882	<0.005
conductivity	conductivity	mud	0.1932	3.4277	<0.005
PO_4-P	PO_4-P				
plant species richness	plant species richness				
Mud	mud				
b) Plant species composition		b) Plant species composition			
ρ_{BC} = 0.288	ρ_{MG} = 0.276	χ^2 = 0.4690			
maxSdepth	max sam. depth	maxSdepth	0.2765	2.8992	<0.005
pH	pH	mud	0.1925	2.0179	<0.005
Mud	conductivity				

explained by the models of species composition reflects the higher complexity of multispecies abundance, consistent with previous investigations in terrestrial systems (e.g., Guisan et al. 1999, Su et al. 2004). The high variability of species composition data was also evident from the fact that only two explanatory variables were never selected by any of the models of species of composition. Some variables were significant determinants of patterns in species composition only (e.g., pond age), providing additional evidence that species richness may not summarize exhaustively community patterns, as previously suggested (Clarke 1993, Su et al. 2004, Fleishman et al. 2006).

When modeling species richness, the forward procedure proposed by Blanchet et al. (2008) resulted in the selection of a lower number of significant variables compared to GLM and explained a lower percentage of variation, indicating that

this method does indeed avoid calculating an over-inflated variance. These differences/advantages were not, however, evident when using the FP to model patterns in species composition. The use of parametric *versus* non-parametric approaches to model species richness (GLM and FP *versus* DISTLM) also resulted in differences in the number, identity, and contribution of the significant variables to the overall explained variance. The GLM and the DISTLM explained virtually the same amount of variation, probably due to the fact that species richness was normalized prior to the data analyses, allowing the use of a parametric measure of dissimilarity (Euclidean distance) in the DISTLM.

All the models selected pond-depth related variables and substratum among the significant determinants of plant and beetle community patterns, consistent with previous studies on plant and invertebrate assemblages in ponds and lakes (e.g., Rodwell 1995, Fairchild et al. 2003, Pakulnicka 2008). Some of the study models emphasized the role of maximum pond depth as a determinant of biodiversity patterns, while others identified in maximum depth of sampling the major determinants of such patterns. Despite being depth-related, these measures encompass different information, as indicated by their low correlation in permanent ponds. While maximum sampling depth provides information on the aquatic-terrestrial transition zones of a pond, maximum pond depth affects the presence of emergent and submerged plant species. Thus, the selection of only one pond depth-related measure may have important implications for the development of conservation strategies. The contribution of chemical variables to patterns in beetle and plant species composition was not always evident in the reduced models, consistent with previous investigations in ponds (Gee et al. 1997, Nicolet et al. 2004, Heino 2000, Jeffries 2008). This is likely due to the absence, in the ponds used for this study, of large chemical gradients and extreme conditions, as well as to a broad tolerance of the recorded species to physico-chemical variables (Gioria et al. 2010).

The differences in the measure of dissimilarity used in the non-parametric models and in the model selection criterion (AIC *versus* BIC) affected both the overall explained variance and the number of variables included in the reduced models. It is beyond the scope of this paper to discuss the differences in the properties of these or other dissimilarity measures (see Legendre and Legendre 1998) or criteria of model selection (see Burnham and Anderson 2004). Although we do not here suggest the use of multiple dissimilarity measures/criteria, we showed that ecologists must be aware that their choice in relation to the use of alternative measures/criteria may affect substantially the outcomes of any model and must therefore be based on the type of data we intend to analyze. The same can be said of the choice of the transformation applied to both biotic and abiotic variables prior to the analysis, although not presented in this study (M. Gioria, unpublished results).

CCA was the multivariate method that selected the lowest number of significant variables (pond-depth related measures and substratum) and did not include 'pond surface area' among the significant drivers of vegetation patterns. This indicates that extreme caution is required when interpreting the ecological significance of the results of only one theoretical model, and that multiple analytical approaches should be used in the analysis of biodiversity patterns. Each method has, in fact, advantages and drawbacks over others. In a comparison of CCA and GLM models of plant species distribution, Guisan et al. (1999) concluded that while CCA gives a broader view of ecological gradients in an area, particularly in the presence of rare species, GLM provides better species-specific models, although both approaches showed a similar ranking of model quality. In a study on littoral ascidians, Naranjo et al. (1996) compared the results of CCA and BIO-ENV and suggested a combined use of these methods, despite being conceptually different. Since the BIO-ENV procedure includes a stopping rule when ρ_s decreases with the inclusion of unimportant variables, these authors recommended the use of BIO-ENV to select the variables that would be subsequently used as the explanatory variables in a CCA.

The use of multi-trophic groups may aid in the interpretation of the results of theoretical models aimed at assessing the effects of environmental variables on biotic communities. Here, the combined use of plant and beetle data provided useful information on both the direct and the potential indirect of the environmental variables on the ecological quality of the study ponds. The role of the vegetation in providing food, shelter from predators, and a physical structure to invertebrate communities has been reported in both aquatic and terrestrial systems (e.g., Foster et al. 1990, 1992, Schaffers et al. 2008), and the strength of vegetation patterns in predicting water beetle species composition has been recently quantified (Gioria et al. 2010).

Information on species identity is also central to a proper interpretation of the results of any theoretical model. Here, the variable 'pond surface area' was not always selected among the determinants of plant or beetle community patterns. However, even if larger ponds support significantly more diverse communities, the role of small or temporary ponds in maintaining habitat connectivity and in acting as a refuge to some species may be central to the maintenance of biodiversity at the regional level, as previously shown in other studies (e.g., Nicolet et al. 2004, Biggs et al. 2005, Gioria et al. 2010). The role of species identity could be included in theoretical models by using weights accounting for a species' rarity and conservation value. This information could be already be available for certain regions and taxonomic groups (e.g., Foster et al. 1990).

As the field of pond ecology rapidly increases, there is a growing need for adopting common analytical approaches that allow for a synthesis of the results of a multiplicity of studies. This could be best achieved by progressively combining information from purely observational investigations with results obtained from experimental studies. In the long term, the use of balanced experimental designs and hypothe-

sis testing procedures would allow a more rigorous quantification of the effects of categorical variables and of the interactions among variables. It would also allow conducting meta-analysis studies and drawing more general conclusions about the effects of specific variables on communities and ecosystems.

Conclusions

Our results confirm that a critical approach is required when discussing biotic-abiotic relationships based on the results of theoretical models. As pointed out by Guthery et al. (2005) in a broader context, any theoretical model of biodiversity, in any system, should possess an interpretable and ecologically-sound meaning, and that no model selection criterion can replace an ecologist's experience. We cannot provide a remedy for the agony of selecting the most appropriate procedure to model biotic-abiotic relationships. Based on our results, however, we can make some general recommendations. First, we recommend the use of both species richness and species abundance data as the response variables of models of biodiversity. Second, we recommend the use of at least two modeling approaches for each response variable (univariate and multivariate), to increase the likelihood of detecting the major contributors of biodiversity patterns and to account for the highly variable nature of species data. Non-parametric modeling approaches should be included in studies of biotic-abiotic relationships, since these methods are flexible, robust, and have long been shown to be more appropriate for the analysis of community data, not being based on any assumption of multivariate normality (see Clarke 1993). Moreover, the use of non-parametric methods does not require data transformation, such as the one applied in this study for comparative reasons, avoiding having to reduce the contribution of species abundance data in the models to information similar to that provided by presence/absence data. Finally, we encourage the use of sampling designs and analytical approaches that allow for a synthesis of the results of a multiplicity of studies. This would be best achieved by progressively shifting from purely observational studies to experimental investigations and to the use of hypothesis testing analytical procedures.

Acknowledgements: The authors thank Dr. P. Ódor and two anonymous reviewers for their comments on the manuscript. Many thanks go to all the people that granted permission to sample the study ponds, particularly to Dr. E. Bannon and the Mullingar farming community. This study was supported by the Environmental Protection Agency Science, Technology, Research and Innovation for the Environment programme 2007-2013 (STRIVE Fellowship: 2007-FS-B-14-S5).

References

Anderson, M.J. 2001. Permutation tests for univariate or multivariate analysis of variance and regression. *Can. J. Fish. Aquat. Sci.* 58: 626-639.

Anderson, M.J., R.N. Gorley and K.R. Clarke. 2008. *PERMANOVA1 for PRIMER. Guide to software and statistical methods*. PRIMER-E Ltd., Plymouth, UK.

Biggs, J., P. Williams, M. Whitfield, P. Nicolet and A. Weatherby. 2005. 15 Years of pond assessment in Britain: results and lessons learned from the work of Pond Conservation. *Aquat. Conserv. Mar. Freshwater Ecosyst*. 15: 693-714.

Bilton, D.T., L. McAbendroth, A. Bedford and P.M. Ramsay. 2006. How wide to cast the net? Cross-taxon congruence of species richness, community similarity and indicator taxa in ponds. *Freshwater Biol.* 51: 578-590.

Blanchet, F.G., P. Legendre and D. Borcard. 2008. Forward selection of explanatory variables. *Ecology* 89: 2623-2632.

Bray, J. and J. Curtis. 1957. An ordination of the upland forest communities of Southern Wisconsin. *Ecol. Monog.* 27: 325-349.

Burnham, K.P. and D.R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Method. Res.* 33: 261-304.

Céréghino, R., J. Biggs, B. Oertli and S. Declerck. 2008. The ecology of European ponds: defining the characteristics of a neglected freshwater habitat. *Hydrobiologia* 597: 1-6.

Clarke, R.K. 1993. Non-parametric multivariate analyses of changes in community structure. *Austral. Ecol.* 18: 117-143.

Clarke, K.R. and M. Ainsworth. 1993. A method of linking multivariate community structure to environmental variables. *Mar. Ecol. Prog. Ser.* 92: 205-219.

Clarke, R.K. and R. Warwick. 2001. *Change in Marine Communities: an Approach to Statistical Analysis and Interpretation*. PRIMER-E, Plymouth.

Crawley, M.J. 1993. *Glim for Ecologists*. Blackwell, Oxford.

Fairchild, G.W., J. Cruz, A.M. Faulds, A.E.Z Short and J.F. Matta. 2003. Microhabitat and landscape influences on aquatic beetle assemblages in a cluster of temporary and permanent ponds. *J. N. Am. Benthol. Soc.* 22: 224-240.

Fleishman, E., R. Noss and B.R. Noon. 2006. Utility and limitations of species richness metrics for conservation planning. *Ecol. Indic.* 6: 543-553.

Foster, G.N., A.P. Foster, M.D. Eyre and D.T. Bilton. 1990. Classification of water beetle assemblages in arable fenland and ranking of sites in relation to conservation value. *Freshwater Biol.* 22: 343-354.

Foster, G.N., B.H. Nelson, D.T. Bilton, D.A. Lott, R. Merritt, R.S. Weyl and M.D. Eyre. 1992. A classification and evaluation of Irish water beetle assemblages. *Aquat. Conserv. Mar. Freshwater Ecosyst.* 2: 185-208.

Gee, J.H.R., B.D Smith., K.M. Lee and S.W. Griffiths. 1997. The ecological basis of freshwater pond management for biodiversity. *Aquat. Conserv. Mar. Freshwater Ecosyst.* 7: 91-104.

Gioria, M., A.P. Schaffers, G. Bacaro and J. Feehan. 2010. Predicting the conservation value of farmland ponds: use of vascular plants as a surrogate group. *Biol. Conserv.* 143: 1125-1133.

Gower, J.C. 1971. Statistical methods of comparing different multivariate analyses of the same data. In: F.R. Hodson, D.G. Kendall and P. Tautu (eds.), *Mathematics in the Archaeological and Historical Sciences*. Edinburgh University Press, Edinburgh. pp. 138-149.

Guisan, A., S.B. Weiss and A.D. Weiss. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecol.* 143: 107-122.

Guisan, A. and N.E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecol. Model.* 135: 147-186.

Guthery, F.S., L.A. Brennan, M.J. Peterson and J.J. Lusk. 2005. Information theory in wildlife science: critique and viewpoint. *J. Wildl. Manage.* 69: 457-465.

Hastie, T.J. and D. Pregibon. 1993. Generalized linear models. In J.M. Chambers and T.J. Hastie (eds.), *Statistical Models in S*. Chapman and Hall London, UK. pp. 194-244.

Heino, J. 2000. Lentic macroinvertebrate assemblage structure along gradients in spatial heterogeneity, habitat size and water chemistry. *Hydrobiologia* 418: 229-242.

Jackson, D.A. 1995. PROTEST: a Procrustean randomization test of community environment concordance. *Ecoscience* 2: 297-303.

Jeffries, M.J. 2008. The spatial and temporal heterogeneity of macrophyte communities in thirty small, temporary ponds over a period of ten years. *Ecography* 31: 765-775.

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

Mantel, N.A. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220.

McArdle, B.H. and M.J. Anderson. 2001. Fitting multivariate models to semi-metric distances: a comment on distance-based redundancy analysis. *Ecology* 82: 290-297.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall, London.

Menetrey, N., L. Sager, B. Oertli and J.-B. Lachavanne. 2005. Looking for metrics to assess the trophic state of ponds. Macroinvertebrates and amphibians. *Aquat. Conserv. Mar. Freshwater Ecosyst*. 15: 653-664.

Montgomery, D.C. and L.A. Peck. 1982. *Introduction to Linear Regression Analysis*. John Wiley and Sons, New York.

Naranjo, S.A., J.L. Carballo and J.C. Garcia-Gomez. 1996. Effects of environmental stress on ascidian populations in Algeciras Bay (southern Spain). Possible marine bioindicators? *Mar. Ecol. Prog. Ser.* 144: 119-131.

Nicolet, P., J. Biggs, G. Fox, M.J. Hodson, C. Reynolds, M. Whitfield and P. Williams. 2004. The wetland plant and macroinvertebrate assemblages of temporary ponds in England and Wales. *Biol. Conserv.* 120: 261-278.

Oertli, B., D.A. Joye, E. Castella, R. Juge, D. Cambin and J.-B. Lachavanne. 2002. Does size matter? The relationship between pond area and biodiversity. *Biol. Conserv.* 104: 59-70.

Pakulnicka, J. 2008. The formation of water beetle fauna in anthropogenic water bodies. *Oceanol. Hydrobiol. St.* 37: 31-42.

Peres-Neto, P.R. and D.A. Jackson. 2001. How well do multivariate data sets match? The robustness and flexibility of a Procrustean superimposition approach over the Mantel test. *Oecologia* 129:169-178.

R Development Core Team. 2010. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna.

Rodwell, J.S. 1995. *British Plant Communities*, vol. 4. Cambridge University Press, Cambridge.

Sætersdal, M., I. Gjerde, H.H. Blom, P.G. Ihlen, E.W., Myreseth, R. Pommeresche, J. Skartveit, T. Solhøyb and O. Aasc. 2003. Vascular plant as a surrogate species group in complementary site selection for bryophytes, macro-lichens, spiders, carabids, staphylinids, snails, and wood living polypore fungi in a northern forest. *Biol. Conserv.* 115: 21-31.

Schaffers, A.P., I.P. Raemakers, K.V. Sýkora and C.J.F. ter Braak. 2008. Arthropod assemblages are best predicted by plant species composition. *Ecology* 89: 782-794.

Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6: 461-464.

Studinski, J.M. and S.A. Grubbs. 2007. Environmental factors affecting the distribution of aquatic invertebrates in temporary ponds in Mammoth Cave National Park, Kentucky, USA. *Hydrobiologia* 575: 211-220.

Su, J.C., D.M., Debinski, M.E. Jakubauskas and K. Kindscher. 2004. Beyond species richness: community similarity as a measure of cross-taxon congruence for coarse-filter conservation. *Conserv. Biol.* 18: 167-173.

ter Braak, C.J.F. 1986. Canonical correspondence analysis: A new eigenvariable technique for multivariate direct gradient analysis. *Ecology* 67: 1167-1179.

Thuiller, W. 2003. BIOMOD: Optimising predictions of species distributions and projecting potential future shifts under global change. *Glob. Change Biol.* 9: 1353-1362.

Williams, P., M. Whitfield, J. Biggs, S. Bray, G. Fox, P. Nicolet and D. Sear. 2004. Comparative biodiversity of rivers, streams, ditches and ponds in an agricultural landscape in Southern England. *Biol. Conserv.* 115: 329-341.

Wood, P.J., M.T. Greenwood and M.D. Agnew. 2003. Pond biodiversity and habitat loss in the UK. *Area* 35: 206-216.

Received March 30, 2010
Revised June 25, 2010
Accepted September 8, 2010

Electronic supplement: R packages and functions

The file may be downloaded from the web site of the publisher at www.akademiai.com.