Aquat Ecol (2007) 41:427–441
DOI 10.1007/s10452-007-9081-7

ORIGINAL PAPER

Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks

Wim Gabriels · Peter L. M. Goethals · Andy P. Dedecker · Sovan Lek · Niels De Pauw

Received: 1 June 2006 / Accepted: 10 January 2007 / Published online: 6 February 2007

Abstract The effect of environmental conditions on river macrobenthic communities was studied using a dataset consisting of 343 sediment samples from unnavigable watercourses in Flanders, Belgium. Artificial neural network models were used to analyse the relation among river characteristics and macrobenthic communities. The dataset included presence or absence of macroinvertebrate taxa and 12 physicochemical and hydromorphological variables for each sampling site. The abiotic variables served as input for the artificial neural networks to predict the macrobenthic community. The effects of the input variables on model performance were assessed in order to identify the most diagnostic river characteristics for macrobenthic community composition. This was done by consecutively eliminating the least important variables and, when beneficial for model performance, adding previously removed ones again. This stepwise input variable selection procedure was tested not only on a model predicting the entire macrobenthic community, but also on three models, each predicting an individual taxon. Additionally, during each step of the stepwise leave-one-out procedure, a sensitivity analysis was performed to determine the response of the predicted macroinvertebrate taxa to the input variables applied. This research illustrated that a combination of input variable selection with sensitivity analyses can contribute to the development of reliable and ecologically relevant ANN models. The river characteristics predicting presence or absence of the benthic macroinvertebrates best were the Julian day, conductivity, and dissolved oxygen content. These conditions reflect the importance of discharges of untreated wastewater that occurred during the period of investigation in nearly all Flemish rivers.

Keywords Benthic macroinvertebrates · Biotic Sediment Index · Ecological modelling · River sediment · River pollution

Abbreviations

ANN	Artificial neural network
BBI	Belgian Biotic Index
BSI	Biotic Sediment Index
CCI	Number of correctly classified instances

W. Gabriels · P. L. M. Goethals (✉) · A. P. Dedecker · N. De Pauw
Department of Applied Ecology and Environmental Biology, Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, J. Plateaustraat 22, 9000 Ghent, Belgium
e-mail: Peter.Goethals@UGent.be

S. Lek
LADYBIO UMR 5172, CNRS-University Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex, France

RIVPACS	River In Vertebrate Prediction And Classification System
RMSE	Root mean square error
SLOO	Stepwise leave-one-out

Introduction

Development and use of models predicting macroinvertebrate community composition has gained a lot of interest during the past decade. Such models are of considerable value for decision support in river management (Goethals and De Pauw 2001). Another application is prediction of the macroinvertebrate community that would be present at a river site in the absence of environmental stress. The European Water Framework Directive (EU 2000) requires EU member states to assess the ecological status of water bodies by comparing the actual and reference status of biological communities. When no reference sites are available, reference status may be based on modelling (Logan and Furse 2002). Software packages such as RIVPACS (Wright 2000) and AUSRIVAS (Davies 2000) offer site-specific predictions of the macroinvertebrate fauna to be expected in the absence of major environmental stresses. Based on these predictions and the fauna present, an environmental quality index can be calculated (Wright 2000; Clarke et al. 2003). A variety of modelling techniques are applied in this context. RIVPACS and many related assessment systems are based on classical multivariate techniques. During recent years however, data mining techniques are increasingly being used, such as artificial neural networks (ANNs) (e.g. Hoang et al. 2001; Dedecker et al. 2004) and decision trees (e.g. Dzeroski et al. 1997; D'heygere et al., 2003). Various authors have shown that ANNs provide powerful predictive models, which in many cases outperform the more traditional modelling tools (e.g. Paruelo and Tomasel 1997; Guégan et al. 1998; Walley and Fontama 1998; Lek and Guégan 1999). ANNs are known for their capacity to process non-linear relationships (Hornik et al. 1989; Chen et al. 1990). This feature makes these models particularly useful for applications in ecological system analysis (e.g. Gevrey et al. 2004).

Neural networks can be valuable instruments to find the dominant sources of stress affecting river communities. However, selection of variables that best describe river status is important for effective model development. A large number of input variables can provide an accurate description of the studied issue, but results in more complex models that are difficult to characterise and require more computational processing time, and often more data for effective discrimination (Maier and Dandy 2000). Several procedures have been tested to select input variables for ANNs, such as a progressive elimination of the least important variables (Walley and Fontama 1998), sensitivity analysis (Schleiter et al. 1999; Hoang et al. 2001), a senso-net (Schleiter et al. 2001) and genetic algorithms (Goethals 2005; D'heygere et al. 2006). A review of methods for analysing variable contribution in ANNs is given by Gevrey et al. (2003).

In this article, a new approach for input variable selection is proposed and tested. The ANN input variables for predicting benthic macroinvertebrate communities were selected by a stepwise leave-one-out (SLOO) procedure. Variables were excluded or added based on their effect on model performance as assessed by Cohen's κ (1960). In this manner, the effect of the prevalence of the different macroinvertebrate taxa was compensated for during the river characteristics selection procedure. This is in contrast to methods merely making use of the root mean square error (RMSE) and the number of correctly classified instances (CCI). Simultaneously, during each step of the SLOO procedure, a sensitivity analysis (Lek et al. 1995, 1996a, b) was performed to determine the response of the predicted macroinvertebrate taxa to the applied input variables. The emphasis is put on organic pollution due to urban wastewater discharges and nutrient enrichment due to agricultural land use, because these were assumed to be the main sources of impact on the aquatic community in Flanders during the period of sampling (1996–1998) (De Cooman et al. 1999; Goethals 2005).

Materials and methods

Dataset

Between 1996 and 1998, 360 sediment samples were collected in unnavigable watercourses throughout Flanders, Belgium (Fig. 1). The samples were taken by means of a Van Veen grab sampler (2 l volume), zigzagging across the watercourse over a length of 50 m (Ministry of the Flemish Community 2000). Between 25 and 40 sub-sample grabs (up to a total volume of approximately 40 l) were collected and mixed together to form a homogeneous sample. From this mixture, a random subsample of approximately 13 l was kept separate for studying the macroinvertebrate community (De Pauw and Heylen 2001). For each sample, all present macroinvertebrate taxa were recorded. The identification level for these taxa was genus or family, except for the Diptera family Chironomidae, which was divided into the group *thummi-plumosus* and the group non *thummi-plumosus* (cf. De Pauw and Heylen 2001). The total dataset comprised 92 different taxa. For each sample a number of abiotic variables was recorded, including in situ measurements of sediment pore water, physicochemical properties of the sediment and granulometric characteristics. The environmental variables used in this study are summarised in Table 1. Seventeen samples were excluded from the dataset due to missing data.

These data were collected within the context of the development and optimisation of the TRIAD methodology for assessment of freshwater sediments (e.g. Chapman et al. 1991) of rivers in Flanders (Ministry of the Flemish Community 2000). The TRIAD assessment is based on biological, ecotoxicological and physicochemical data. The biological component consists of determining the Biotic Sediment Index (BSI) (De Pauw and Heylen 2001) and the percentage mentum deformities in *Chironomus* larvae (Heylen and De Pauw 2003). The BSI is a modification of the Belgian Biotic Index (BBI) (De Pauw and Vanhooren 1983) and is based on the taxonomic diversity of the benthic macroinvertebrate community and the presence or absence of specific

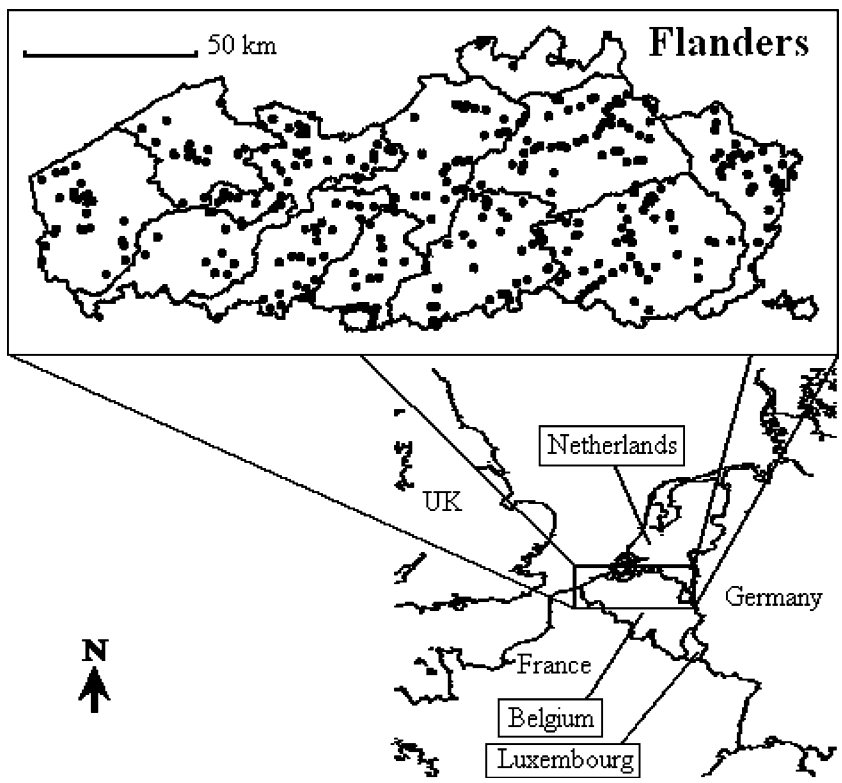

Fig. 1 Overview of the distribution of the 360 sediment sampling sites in unnavigable watercourses throughout Flanders, Belgium

Table 1 Environmental variables in the dataset used in the present study

Variable	Abbreviation	Units	Min	Mean	Max
Date of sampling	DAY	Julian day (1–365)	20	175	338
River width	WIDT	m	0.4	3.8	15.0
River depth	DEPT	m	0.01	0.61	3.00
Stream velocity class	VELO	0 (stagnant) to 4 (fast)	0	1.9	4
Clay fraction in sediment	CLAY	%	0	11	65
Silt fraction in sediment	SILT	%	0	20	80
Sand fraction in sediment	SAND	%	0	69	100
pH	PH	–	3.38	7.42	9.06
Dissolved oxygen	DO	mg/l	0.1	5.7	13.2
Electric conductivity	COND	mS/cm	0.11	0.91	16.66
Total phosphorus in sediment	TP	mg P/kg dry matter	17	1,759	42,200
Kjeldahl nitrogen in sediment	TKN	mg N/kg dry matter	100	2,022	1,1200

indicator taxa in the sediment sample. The BSI ranges from 10 for unimpacted sediments to 0 for severely polluted sediments.

Development and assessment of an ANN predicting all taxa simultaneously

Three-layered feed-forward neural networks with bias were constructed to predict the benthic macroinvertebrate community composition. The neural network which was initially developed consisted of an input layer with 12 neurons (one for each input variable mentioned in Table 1), a hidden layer with a number of neurons optimised by trial and error, and an output layer with 92 neurons, corresponding to all macroinvertebrate taxa present in the dataset. The trial and error process was conducted by consecutively training and validating ANNs with varying numbers of hidden neurons until no further improvement of model performance, as assessed by Cohen's κ (see further), was obtained. All neural networks were trained using the error backpropagation algorithm with momentum and adaptive learning rate (Hagan et al. 1996). All river characteristics, that were used as input variables, were rescaled to the interval [–1 1] prior to presenting them to the ANN. Output values equalled zero for absence and one for presence.

Model performance was assessed with cross-validation (Witten and Frank 2000). When using cross-validation, the original dataset is equally split into n subsets. Subsequently, n models are trained and validated, each subset in turn being used as validation set for a model that is trained using the other $n - 1$ subsets. These n validations together are used for evaluation of model architecture. This method is particularly useful when only a limited number of data are available for training and validating a model. In this case, 7-fold cross-validation was used, hence a training set of 294 patterns and a validation set of 49 patterns was available for each fold. The ANN output values, continuous values between zero and one, were rounded in order to enable a comparison with the discrete absence/presence values from the dataset. Values larger than or equal to 0.5 were rounded up to 1. For each validation site, the actual presence or absence and the one predicted by the model could be compared for all 92 taxa. This gave rise to 92 times 343, or 31566 cases to be compared each time. The assessment was based on the calculation of the percentage of CCI (Witten and Frank 2000) and Cohen's κ (1960). Both CCI and κ can be used for comparing model predictions, but the κ value takes a correction into account for the expected number of correct predictions due to randomness, which is strongly related to taxon prevalence (Manel et al. 2001). Therefore κ provides a more reliable representation of model performance (Cohen 1960). Kappa values are evaluated as follows in medical applications: 0.00–0.40: slight to fair; 0.40–0.60: moderate; 0.60–0.80: substantial; 0.80–1.00: almost perfect (Manel et al. 2001 after Landis and Koch, 1977). However, these κ values also represent the information that is in the dataset, and each dataset has a limit regarding extractable information. Consequently, also differences between classes can be expected between disciplines in general and datasets in particular. As a result, the κ cannot

be seen as an absolute value to make a model evaluation, but should rather be seen as a good way to compare models, and study the effect of removing variables. In an ecological context, Randin et al. (2006) assess κ values as follows: 0.00–0.40: poor; 0.40–0.75: good; 0.75–1.00: excellent. The following assessment scheme, based on the two cited schemes, will be used throughout this article:

0.00–0.20: poor;
0.20–0.40: fair;
0.40–0.60: moderate;
0.60–0.80: substantial;
0.80–1.00: excellent.

Development and assessment of ANNs predicting individual taxa

In order to study the effects of the river characteristics on individual taxa, three models were developed which were similar to the previous ones, each time using one individual taxon as output. Those taxa with prevalence closest to 25%, 50% and 75%, respectively, were chosen as focus taxa for this study. In this manner, representatives of different tolerance classes could be compared. These taxa were *Pisidium* (Bivalvia, Sphaeriidae) (27.1%), *Erpobdella* (Hirudinea, Erpobdellidae) (37.3%) and Chironomidae, group *thummi-plumosus* (Insecta, Diptera) (73.8%). The input variables were the same as used for the whole community. Neural network architecture was also identical to the previous one, except for the number of hidden neurons, which was again optimised by trial and error. Assessment was once again carried out by means of 7-fold cross-validation.

Input variable selection

To study the impact of the input variables on the ANN predictions, a SLOO procedure for variable selection was proposed and tested. Throughout this selection procedure, all characteristics of the ANN remained unaltered, except for the number of input neurons.

The SLOO procedure is outlined in Fig. 2. It is an iterative process starting with the 12 input variables until only one variable remains. Each phase of this process consists of two steps. In the first step a new series of ANN models is trained and validated. The number of models equals the number of remaining variables, each model being constructed by excluding a different variable. The best performing model, according to Cohen's κ, is selected, and hence the least important variable is excluded. In the second step, a new series of models is built, this time by adding all previously removed variables again. If one of these models performs better than any previous model with the same number of variables, the corresponding variable is again included. Otherwise no variable is added in the second step. This iterative process of stepwise removing the least important

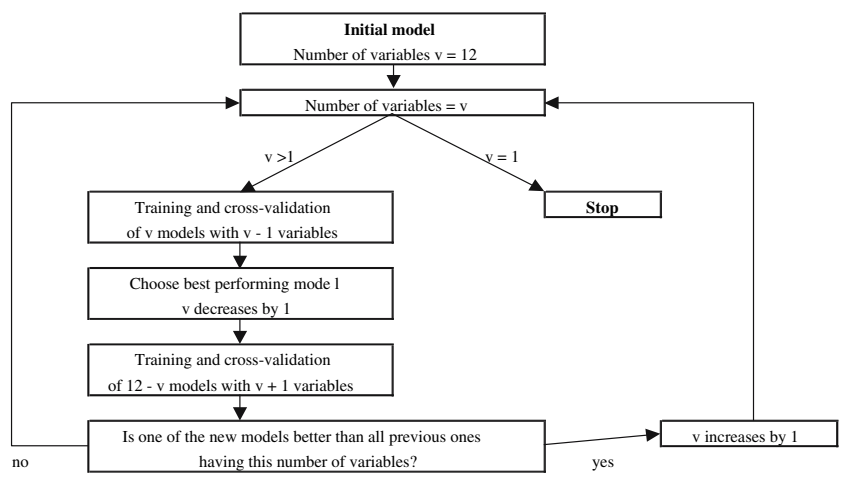

Fig. 2 Summary of the stepwise leave-one-out procedure used in the present study to select input variables for neural network models predicting absence/presence of macroinvertebrate taxa

variables for predicting macroinvertebrate taxa is continued until only one variable is left.

This input variable selection procedure was tested for the ANN predicting all taxa simultaneously as well as for the three models predicting individual taxa.

Sensitivity analysis

During each step of the SLOO procedure, a sensitivity analysis was performed based on Lek et al. (1995, 1996a, b). Olden et al. (2004) illustrated that other methods can be better to select input variables of models. However the method of Lek et al. (1995) was selected because it has as major advantage that it directly illustrates the relation between input variables and the predicted variable. This extra information is very useful, because this provides direct insight regarding the ecological relevance of this relation as well. In this way, the stability of these ecological relations could be monitored during the selection process. This allowed the interpretation of the impact of river characteristics on the probability of occurence of the three focus taxa (*Pisidium*, *Erpobdella* and Chironomidae, group *thummi-plumosus*). Twelve values of a variable were taken at equal intervals covering the whole range of the variable within the dataset, starting with the minimum and concluding with the maximum. These values were separately presented to the ANN, while all other variables were kept constant at their mean value within the dataset. In this way, the effect of one variable on ANN predictions throughout its range within the dataset could be visualised. (Lek et al 1996b). For all three focus taxa, this sensitivity analysis was performed for all remaining variables following each step of the SLOO procedure.

Results

Development and assessment of an ANN predicting all taxa simultaneously

For the initial ANN model, optimisation of the number of hidden neurons with trial and error resulted in a network architecture with 12 hidden neurons. The percentage of correctly classified presence was 44.6%, while the percentage of correctly classified absence was 98.9% (Table 2). Average CCI percentage was 95.0%. Since the number of taxa absent was usually far higher than the number of taxa present, the total CCI was far closer to the CCI for absent taxa. Kappa equalled 0.537, which corresponds to moderate model performance.

Development and assessment of ANNs predicting individual taxa

Optimisation with trial and error resulted in a neural network architecture with eight hidden neurons for the individual taxa ANNs. CCI values are close to 70% for all three taxa, but κ values can be characterised as poor for *Pisidium* and Chironomidae, group *thummi-plumosus* and fair for *Erpobdella* (Table 3).

Input variable selection for the ANN predicting all taxa simultaneously

Performance remained virtually unchanged when the number of input variables was reduced (Fig. 3; Table 4). Throughout the selection

Table 2 Confusion matrix of the results obtained with 7-fold cross-validation using all input variables to predict absence/presence of 92 macroinvertebrate taxa in the dataset

		Predicted	
		Present	Absent
Actually	Present	1015 (44.6%)	1261 (55.4%)
	Absent	323 (1.1%)	28957 (98.9%)

The total percentage of CCI was 95.0%. Cohen's κ equalled 0.537

Table 3 Results obtained with 7-fold cross-validation using all input variables to predict absence/presence of *Pisidium*, *Erpobdella* and Chironomidae, group *thummi-plumosus*

Taxon	CCI (%)	Kappa
Pisidium	71.1	0.165
Erpobdella	70.3	0.326
Chironomidae, group t.-p.	69.1	0.068

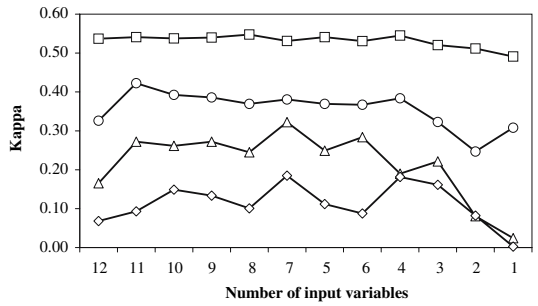

Fig. 3 Influence of the number of input variables on Cohen's κ throughout the input variable selection procedure for four ANN models predicting absence or presence of macroinvertebrate taxa. For each model, only the highest κ found for each number of input variables is plotted. Squares: ANN for all taxa simultaneously; circles: ANN for *Erpobdella*; triangles: ANN for *Pisidium*; rhombuses: ANN for Chironomidae, group *thummi-plumosus*

Table 4 Summary of the selection procedure of the input variables to predict absence/presence of 92 macroinvertebrate taxa in the dataset

Step	Number of input variables	Variable removed	Variable added	CCI (%)	Kappa
1	12	–	–	95.0	0.537
2	11	DEPT	–	95.0	0.541
3	10	VELO	–	94.9	0.538
4	9	TP	–	94.8	0.540
5	8	COND	–	95.0	0.548
6	7	TKN	–	94.8	0.531
7	6	SILT	–	95.0	0.541
8	5	SAND	–	94.9	0.531
9	4	DO	–	94.7	0.516
10	3	PH	–	94.1	0.490
11	4	–	COND	95.0	0.535
12	3	CLAY	–	94.5	0.507
13	4	–	SAND	95.0	0.545
14	3	WIDT	–	94.8	0.519
15	2	SAND	–	94.6	0.506
16	3	–	DO	94.7	0.520
17	2	DO	–	94.6	0.506
18	1	COND	–	94.4	0.491
19	2	–	CLAY	94.8	0.512
20	1	CLAY	–	94.4	0.491

See Table 1 for variable abbreviations

procedure, the decrease of κ in comparison to the initial model is never more than 0.05, and in some cases κ even increases. Model performance remained moderate until only one variable remained. The highest values for κ and CCI, 0.548 and 95.0%, respectively, were obtained when eight input variables were used. When only the variables day, width, and clay were considered as input variables, the lowest κ (0.490) and CCI value (94.1%) was obtained. When all variables were listed in order of importance, expressed as the smallest set of variables in which each variable still appeared, Julian day, clay fraction, conductivity, and dissolved oxygen were the four most important variables (Table 5).

Input variable selection for ANNs predicting individual taxa

The response of κ to number of ANN input variables generally rose to an asymptote for predictions of *Pisidium*, *Erpobdella* and Chironomidae, group *thummi-plumosus* (Fig. 3). The best model performance for *Pisidium* (κ = 0.322 and CCI = 76.4%) was obtained when the input variables pH, width, silt, total phosphorus and Kjeldahl nitrogen were removed. When only one variable was left (day), κ decreased to a minimum of 0.024 although the CCI remained above 70.0%. The κ and the CCI values for *Erpobdella* were between 0.247 and 0.423 and between 66.8% and 74.1%, respectively, when two input variables

Table 5 Ranking of the 12 input variables in order of importance for predicting absence/presence of 92 macroinvertebrate taxa simultaneously, and for the three individual taxa, according to the input variable selection procedures for these four ANNs

Rank of variable	Model			
	All taxa	*Pisidium*	*Erpobdella*	Chironomidae, group *thummi-plumosus*
1	DAY	DAY	COND	PH
2	CLAY	DO	DO	DO
3	COND	SILT	DAY	DAY
4	DO	CLAY	CLAY	TP
5	SAND	DEPT	PH	CLAY
6	WIDT	COND	VELO	TKN
7	PH	SAND	TKN	DEPT
8	SILT	TP	SAND	SILT
9	TKN	WIDT	DEPT	SAND
10	TP	TKN	TP	VELO
11	VELO	VELO	WIDT	WIDT
12	DEPT	PH	SILT	COND

The variable rank equals the lowest number of variables in which it was still included (in other words, the variable that was excluded first has rank 12). See Table 1 for variable abbreviations

(dissolved oxygen and conductivity) were used and only one variable (silt fraction) was removed. For Chironomidae, group *thummi-plumosus*, poor results were obtained based on κ. The highest κ value (0.185) was reached when seven input variables were used. The CCI was higher than the initial percentage of 69.1 after each step of the input variable selection procedure.

The most important variables for the three focus taxa, based on the variable selection procedure, were quite similar. The three most important variables were Julian day, dissolved oxygen concentration and silt fraction for *Pisidium*; conductivity, dissolved oxygen concentration and Julian day for *Erpobdella*; and pH, dissolved oxygen concentration and Julian day for Chironomidae, group *thummi-plumosus* (Table 5).

The effects of the input variables on the probability of presence of *Pisidium*, *Erpobdella* and Chironomidae, group *thummi-plumosus*, respectively, showed a variety of responses (Figs. 4–6). Only the curves for 12, 8 and 4 variables are presented for each taxon (see figures a–c). For *Pisidium*, conductivity and dissolved

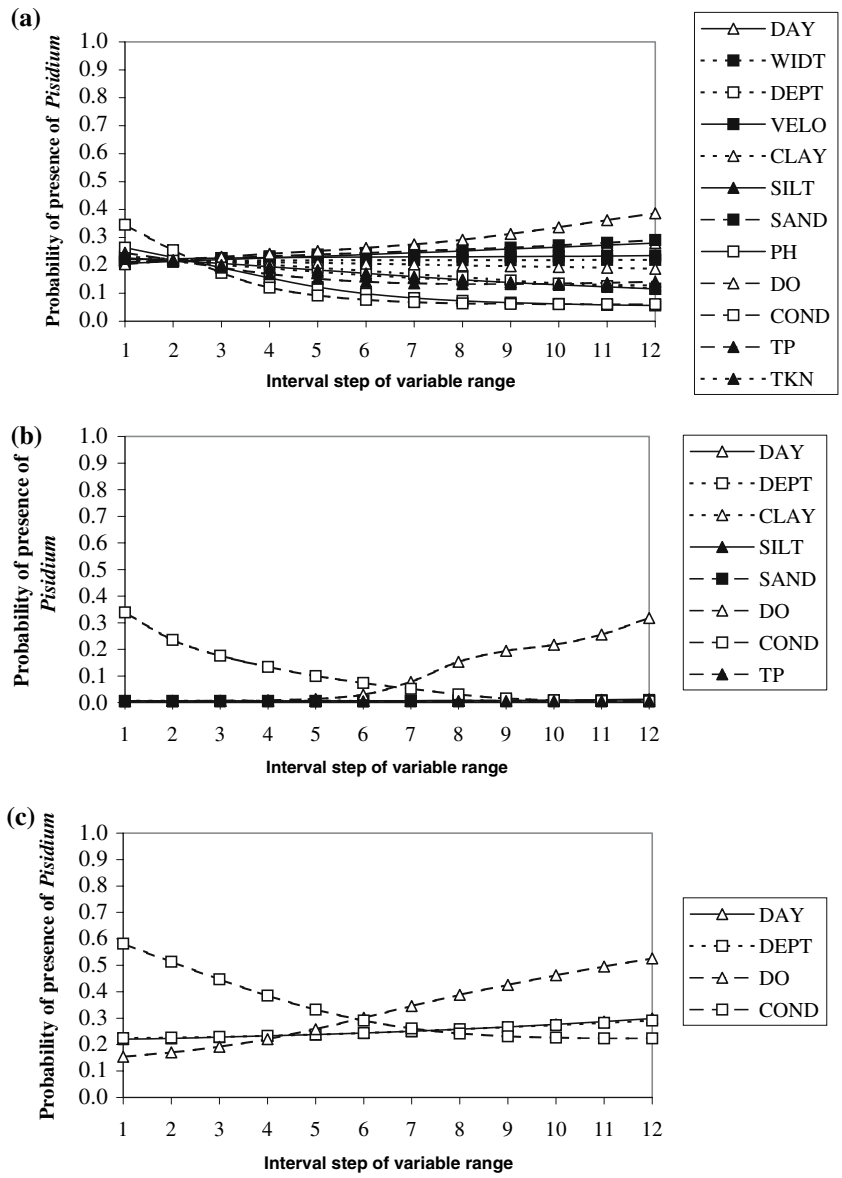

Fig. 4 The impact of the input variables on the ANN-generated probability of presence of *Pisidium*. Only the curves for 12 (**a**), 8 (**b**) and 4 (**c**) variables are shown. See text for further explanation and Table 1 for variable abbreviations

oxygen were best expressed when 12, 8 and 4 input variables were plotted. An increase of conductivity resulted in a decrease of *Pisidium* occurrence, while an increase of dissolved oxygen led to an increase. When the number of input variables becomes smaller, these effects become more distinct. Although Julian day is the most diagnostic variable based on the SLOO selection procedure, it is not expressed well with this sensitivity analysis (Fig. 4). The two most important input variables for *Erpobdella*, based on the SLOO selection procedure were conductivity and dissolved oxygen. Sensitivity analysis confirmed their importance, except for the case where all input variables were used (Fig. 5a). In that case, only conductivity and Kjeldahl nitrogen concentration showed a substantial influence. A decrease of conductivity induced an increase in the probability of presence of *Erpobdella*. An increase of dissolved oxygen resulted in an increase of *Erpobdella* occurrence. For Chironomidae, group *thummi-plumosus*, all input variables were expressed relatively well. When only four variables were used, the impact of pH becomes most

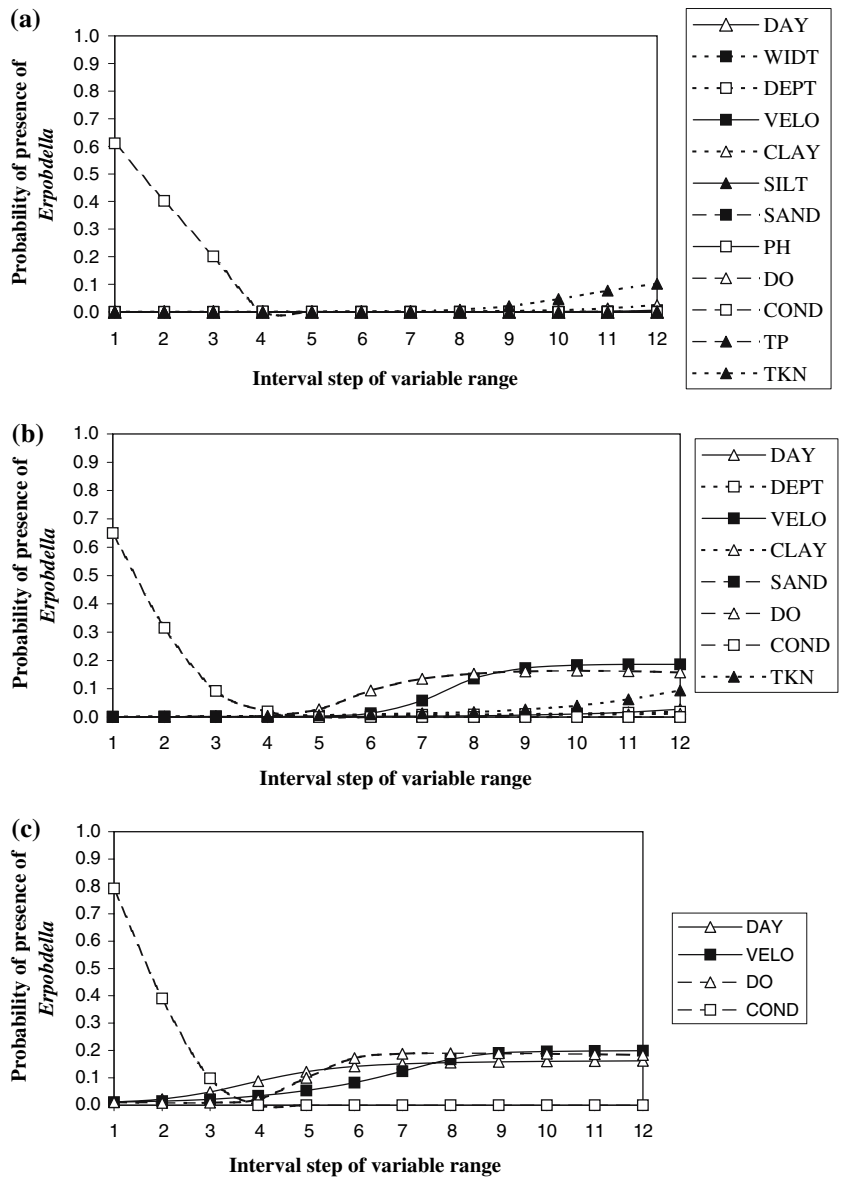

Fig. 5 The impact of the input variables on the probability of presence of *Erpobdella*. Only the curves for 12 (**a**), 8 (**b**) and 4 (**c**) variables are shown. See text for further explanation and Table 1 for variable abbreviations

important, resulting in low probabilities of presence for this taxon at low pH values and high probabilities at high pH values (Fig. 6).

Discussion

Performance of the initial ANN model

The predictive success of the initially constructed ANN, including all 12 variables was moderate (CCI = 95%, κ = 0.537). Gabriels et al. (2002) obtained a slightly lower CCI (92.6%) using the same dataset. Hoang et al. (2001) obtained CCIs between 75% and 95% when independently testing 37 ANNs each predicting a specific stream macroinvertebrate taxon. Based on a dataset of the Zwalm river basin (Flanders, Belgium), Dedecker et al. (2002) found CCIs between 59% and 99% when ANN models were tested for 10 river macroinvertebrate taxa. In contrast to CCI, κ indicates to what extent models correctly

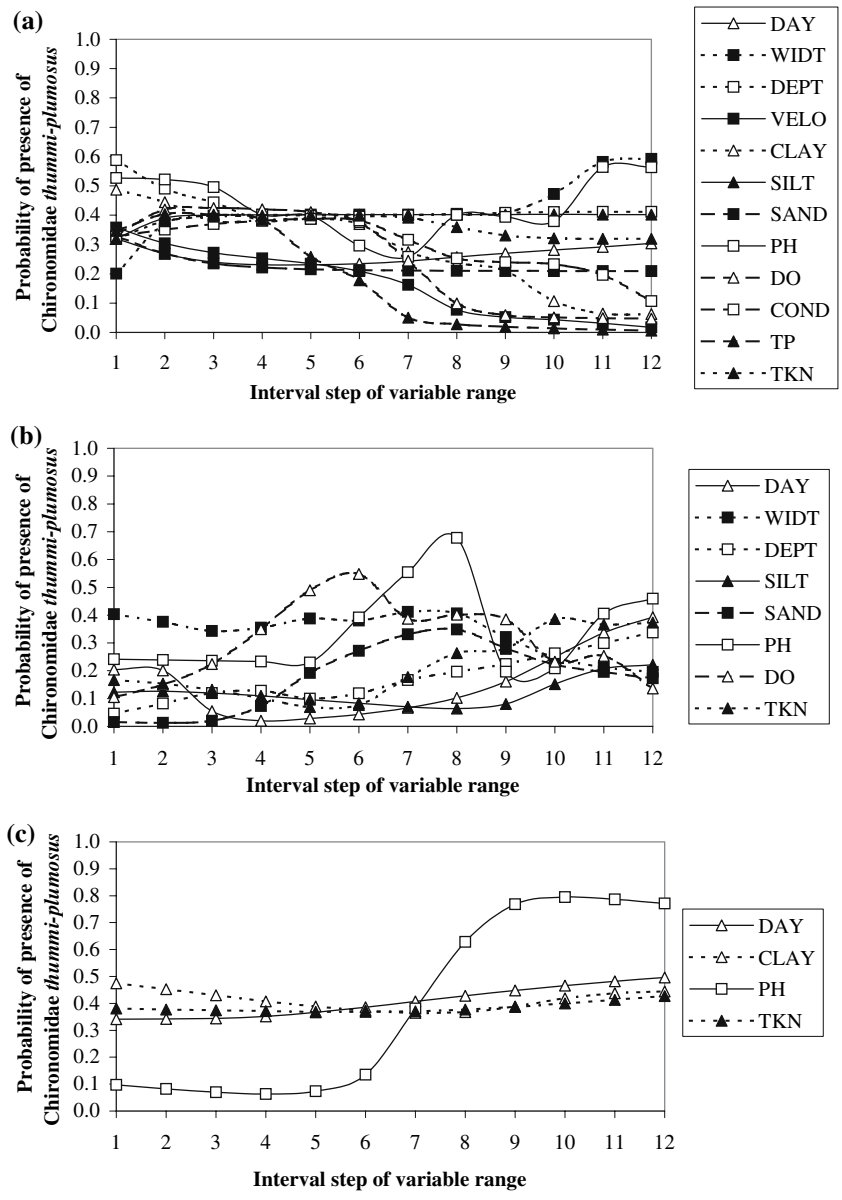

Fig. 6 The impact of the input variables on the probability of presence of Chironomidae, group *thummi-plumosus*. Only the curves for 12 (**a**), 8 (**b**) and 4 (**c**) variables are shown. See text for further explanation and Table 1 for variable abbreviations

predict occurrence at rates that are better than chance expectation (Fielding and Bell 1997; Manel et al., 2001). Therefore, in this study, κ was preferred over CCI for assessing model performance.

Variable selection for all taxa simultaneously

Performance of the neural networks remained virtually unchanged when the number of input variables was reduced from 12 to 4. When the number of input variables was further reduced, performance decreased, although not dramatically, as can be seen in Fig. 3. Surprisingly, a moderate κ value was obtained even with only one input variable. These results are more or less in agreement with Walley and Fontama (1998), who observed an unaffected performance when five or eight (depending on the output variable) out of 13 variables were removed and only a slightly reduced performance when ten out of 13 variables were excluded. It should be noted however that the target variable, the assessment of performance and the selection procedure were different. In the cited study, assessment was based on the correlation coefficient between predicted and actual value, a parameter that would not be suitable for the present study since categoric (presence/absence) variables were compared here.

All input variables could be ranked in order of importance, based on the smallest set of variables in which each variable still appeared. A different ranking can be set up by comparing the performance of all models with only one input variable (not shown). In both rankings, three variables appeared among the four variables ranked highest: Julian day, conductivity and dissolved oxygen. Many macroinvertebrate taxa occurrences are characterised by an annual cycle (e.g. Dolédec 1989; Rosillon 1989; Linke et al. 1999; Reece et al. 2001). Thus, Julian day is evidently a key variable. Dissolved oxygen is also known as an important factor regulating benthic macroinvertebrate community composition (e.g. Ruse 1996; Weigel et al. 2003; Chaves et al. 2005) and low values may indicate organic pollution. Conductivity integrates several variables like natural mineral content of the water due to geology, but also minerals from pollutant degradation (effluents of wastewater treatment plants) and inorganic pollutants. D'heygere et al. (2003) applied genetic algorithms to select input variables in decision tree models for eight taxa, based on the same dataset. They found that dissolved oxygen and conductivity were the most important predictor variables for their models.

In the present study, a total of 153 neural networks were trained and validated, in order to select the input variables. If all possible combinations of input variables were to be tested, one would have to train and validate $2^{12} - 1$, or 4095 ANNs. The method developed in this study results in a drastic reduction of calculation time.

Variable selection for three individual taxa

A trend towards higher κ values for higher numbers of input variables was observed for the three individual taxa, especially in the case of *Pisidium* and *Erpobdella* (Fig. 3). However, this trend was reduced and even inverted when the number of input variables approached 12. A possible explanation is that for 12 input variables the only possible combination of variables was tested, whereas for a smaller number of variables the best out of a number of possible combinations was selected. When comparing the maximum κ value obtained during the selection process, an increase from 0.17 to 0.32 was found for *Pisidium*, from 0.33 to 0.42 for *Erpobdella*, and from 0.07 to 0.19 for Chironomidae, group *thummi-plumosus*. D'heygere et al. (2006), applying genetic algorithms to select input variables for ANN models using the same dataset, obtained an increase in κ only for *Pisidium* (from 0.34 to 0.37) and *Erpobdella* (from 0.28 to 0.33). For Chironomidae, group *thummi-plumosus*, a decrease from 0.16 to 0.14 in model performance was found.

The sensitivity analysis provided a direct interpretation of the effect of river characteristics on the probability of presence of the three focus taxa. However, slightly different curves were obtained when not all variables were used for the sensitivity analysis. Due to the exclusion of some variables of minor importance for model performance, relationships between the dominant variables and the macroinvertebrates became

more distinct (Figs. 4–6). Analysis of the sensitivity curves can thus enhance insight in the effects of various impact types on individual taxa (Marshall et al. 2002). As such, this method would enable impact-specific indicator taxa to be readily identified and would enhance the capacity to monitor and mitigate the effects of human activities on river ecosystems (Dedecker et al. 2005, 2007).

The similarity of the most important variables among the three focus taxa partially results from the range of environmental features represented by the dataset, because it can be expected that variables covering a wide range of values will more likely be selected. Many of the most important variables, such as dissolved oxygen and conductivity, are associated with organic pollution. Organic wastewater pollution is an important problem in Flemish surface waters today (e.g. VMM 2003) and is clearly correlated with variation in macrobenthic communities.

Autecological relationships of macroinvertebrates are described in the literature (e.g. Tachet et al. 2002). Knowledge of tolerance of certain taxa to particular environmental conditions may help in deciding which environmental variables should be measured or preselecting input variables for predictive models. Thus, knowledge of the autecology of taxa can be complementary to automated input variable selection. A narrow tolerance interval for an environmental characteristic for a certain taxon can be expected to be an important variable in predictive models for that taxon. However, the detected range of tolerance interval for a certain characteristic is highly dependent on the range of the sampling sites visited. Furthermore, the effect of one variable can be confounded by interactions between different characteristics (e.g. Gevrey et al. 2006).

General comments and further research

Identification of key variables is important for enhancing knowledge of river ecology and supporting river management. Input variable selection can also improve the efficiency of data collection since some variables may be irrelevant to the problem being examined. Improvement of river water quality may result in other, previously ignored, variables becoming essential. For this reason, expert knowledge remains crucial when it comes to the construction of generalised and robust models (Goethals 2005).

Although application of ANNs, in combination with the SLOO input variable selection procedure, are well accepted, other methods for input variable selection and/or comparison of input variable importance are available, such as correspondence analysis (e.g. Ruse 1996) or principal component analysis (Roadknight et al. 1997), genetic algorithms (Goldberg 1989), senso-nets (Schleiter et al. 2001), sensitivity analysis (Schleiter et al. 1999; Hoang et al. 2001) and progressive elimination of the least important variables (Walley and Fontama 1998). Nonetheless, the simple selection method tested in this article provided useful results. The added value of more advanced techniques such as genetic algorithms could be insignificant when, as in this case, the available set of initial variables is small, and consequently the calculation time for the procedure used here will not become excessively long. Stepwise input variable selection procedures were previously tested for prediction of macroinvertebrate taxa (e.g. Schleiter et al. 1999; Obach et al. 2001; Schleiter et al. 2001; Beauchard et al. 2003), but these were unidirectional procedures. A reversed procedure, in which variables are added stepwise starting from one variable, has not yet been tested. Due to the small number of variables needed to obtain acceptable model performance, the calculation effort could be reduced substantially with a reversed method.

ANN architecture is generally highly problem dependent (Maier and Dandy 2000). For this reason, it is necessary to develop and optimise the ANNs to obtain the best model configuration that gives the lowest error during training. However, throughout our selection procedure, all characteristics of the ANN were unaltered, except for the number of input variables. A more refined procedure could include optimisation of neural network architecture for each number of input variables, although this would involve a substantial increase in calculation time.

The taxonomic levels of identification used in the present study are those defined within the

TRIAD assessment method (Ministry of the Flemish Community 2000). Although they are commonly used in biological water quality assessment systems (e.g. De Pauw and Vanhooren 1983; Hawkes 1997; Gabriels et al. 2005), these levels may be insufficient from the perspective of biodiversity and conservation.

We recommend further work on optimisation of the tested approach as well as comparison with other variable selection techniques, such as the already cited ones. Using abundance values, or a rescaling of abundance values, instead of presence/absence data and the use of taxon-specific models instead of one model for the whole community might enhance model reliability, and as a result possibly optimise the selection as well. In order to determine which variable selection method is the most appropriate for which problem, an extensive comparison should be elaborated using the different methods with different scenarios. Key considerations are calculation time restrictions, data collection costs and required model reliability, all dependent on the studied problem.

Conclusions

ANNs were developed to predict absence or presence of benthic macroinvertebrate taxa in unnavigable watercourses in Flanders. A SLOO procedure was followed to detect those river characteristics which are most significant for macrobenthic communities, resulting in simplified models with only slightly reduced predictive performance. For the three taxa considered, the major input variables included Julian day, conductivity and dissolved oxygen concentration. One may conclude that the presence/absence of organic wastewater discharges had a major influence on the macrobenthic communities in Flemish watercourses during the period of sampling. The sensitivity analysis illustrated that in general the ecological relations were stable during the selection procedure, in particular for *Erpobdella* and *Pisidium*. For Chironomidae, group *thummi-plumosus*, many input variables had a complex relation with the probability of presence. When only four variables were used, the impact of pH becomes most important for this taxon. This demonstrates that pruning predictive models can illuminate ecological relations that remain hidden in more complex models. In conclusion, a combination of input variable selection with sensitivity analyses can contribute to the development of reliable and ecologically relevant ANN models.

Acknowledgements The authors wish to thank ir. Tom D'heygere and Lic. Steven Heylen (Research Unit Aquatic Ecology, Ghent University), Lieven Detemmerman and ir. Ward De Cooman (Flemish Environment Agency), and AMINAL, Division Water, Brussels for providing useful data on the sediment samples in unnavigable watercourses in Flanders, Belgium, and two anonymous reviewers for valuable remarks. Andy Dedecker is a recipient of a grant of the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). The international cooperation within this research was financed by the scientific exchange program Tournesol (project T2003.01). The applied neural network sensitivity analysis toolbox was developed within the European PAEQANN project (EVK1-CT1999-00026).

References

Beauchard O, Gagneur J, Brosse S (2003) Macroinvertebrate richness patterns in North African streams. J Biogeogr 30:1821–1833

Chapman PM, Power EA, Dexter RN, Andersen HB (1991) Evaluation of effects associated with an oil platform, using the Sediment Quality Triad. Environ Toxicol Chem 10:407–424

Chaves ML, Chainho PM, Costa JL, Prat N, Costa MJ (2005) Regional and local environmental factors structuring undisturbed benthic macroinvertebrate communities in the Mondego River basin, Portugal. Arch Hydrobiol 163:497–523

Chen S, Billings SA, Grant PM (1990) Non-linear system identification using neural networks. Int J Control 51:1191–1214

Clarke RT, Wright JF, Furse MT (2003) RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecol Model 160:219–233

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

Davies PE (2000) Development of a national river bioassessment system (AUSRIVAS) in Australia. In: Wright JF, Sutcliffe DW, Furse MT (eds) Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria

De Cooman W, Florus M, Vangheluwe M, Janssen C, Heylen S, De Pauw N, Rillaerts E, Meire P, Verheyen R (1999) Sediment characterisation of rivers in

Flanders. In: De Schutter G (ed) CATS4. PIH, Antwerp, Belgium
Dedecker AP, Goethals PLM, De Pauw N (2002) Comparison of artificial neural network (ANN) model development methods for prediction of macroinvertebrate communities in the Zwalm river basin in Flanders, Belgium. The Scientific World J 2:96–104
Dedecker AP, Goethals PLM, Gabriels W, De Pauw N (2004) Optimisation of Artificial Neural Network (ANN) model design for prediction of macroinvertebrate communities in the Zwalm river basin (Flanders, Belgium). Ecol Model 174:161–173
Dedecker AP, Goethals PLM, D'heygere T, Gevrey M, Lek S, De Pauw N (2005) Application of artificial neural network models to analyse the relationships between *Gammarus pulex* L. (Crustacea, Amphipoda) and river characteristics. Environ Monit Assess 111:223–241
Dedecker AP, Goethals PLM, D'heygere T, Gevrey M, Lek S, De Pauw N (2007) Selecting variables for habitat suitability of *Asellus* (Crustacea, Isopoda) by applying input variable contribution methods to Artificial Neural Network models. Environ Model Assess (in press)
De Pauw N, Vanhooren G (1983) Method for biological quality assessment of watercourses in Belgium. Hydrobiologia 100:153–168
De Pauw N, Heylen S (2001) Biotic index for sediment quality assessment of watercourses in Flanders, Belgium. Aquat Ecol 35:121–133
D'heygere T, Goethals PLM, De Pauw N (2003) Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecol Model 160:291–300
D'heygere T, Goethals PLM, De Pauw N (2006) Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecol Model 195:20–29
Dolédec S (1989) Seasonal dynamics of benthic macroinvertebrate communities in the Lower Ardèche River (France). Hydrobiologia 182:73–89
Dzeroski S, Grbovic J, Walley WJ, Kompare B (1997) Using machine learning techniques in the construction of models. II. Data analysis with rule induction. Ecol Model 95:95–111
Eu (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. Official J Eur Communities L327:1–72
Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49
Gabriels W, Goethals PLM, De Pauw N (2002) Prediction of macroinvertebrate communities in sediments of Flemish watercourses based on artificial neural networks. Verh Internat Verein Limnol 28:777–780
Gabriels W, Goethals PLM, De Pauw N (2005) Implications of taxonomic modifications and alien species on biological water quality assessment as exemplified by the Belgian Biotic Index method. Hydrobiologia 542:137–150
Gevrey M, Dimopoulos I, Lek S (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Model 160:249–264
Gevrey M, Rimet F, Park YS, Giraudel JL, Ector L, Lek S (2004) Water quality assessment using diatom assemblages and advanced modelling techniques. Freshwater Biol 49:208–220
Gevrey M, Dimopoulos I, Lek S (2006) Two-way interaction of input variables in the sensitivity analysis of neural network models. Ecol Model 195:43–50
Goethals PLM (2005) Data driven development of predictive ecological models for benthic macroinvertebrates in rivers. PhD thesis, Ghent University
Goethals P, De Pauw N (2001) Development of a concept for integrated river assessment in Flanders, Belgium. J Limnol 60:7–16
Goldberg DE (1989) Genetic algorithms in search, optimisation and machine learning. Addison-Wesley Publishing Company, Reading, Massachusetts
Guégan JF, Lek S, Oberdorff T (1998) Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391:382–384
Hagan MT, Demuth HB, Beale M (1996) Neural network design. PWS Publishing Company, Boston
Hawkes HA (1997) Origin and development of the biological monitoring working party score system. Water Res 32:964–968
Heylen S, De Pauw N (2003) Mentum deformations in Chironomus larvae for assessment of freshwater sediments in Flanders, Belgium. Verh Internat Verein Limnol 28:781–785
Hoang H, Recknagel F, Marshall J, Choy S (2001) Predictive modelling of macroinvertebrate assemblages for stream habitat assessments in Queensland (Australia). Ecol Model 146:195–206
Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Networks 2:359–366
Landis JR, Koch GC (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
Lek S, Guégan JF (1999) Artificial neural networks as a tool in ecological modelling, an introduction. Ecol Model 120:65–73
Lek S, Belaud A, Dimopoulos I, Lauga J, Moreau J (1995) Improved estimation, using neural networks, of the food consumption of fish populations. Mar Freshwater Res 46:1229–1236
Lek S, Belaud A, Baran P, Dimopoulos I, Delacoste M (1996a) Role of some environmental variables in trout abundance models using neural networks. Aquat Living Resour 9:23–29
Lek S, Delacoste M, Baran P, Dimopoulos I, Lauga J, Aulagnier S (1996b) Application of neural networks to modelling nonlinear relationships in ecology. Ecol Model 90:39–52
Linke S, Bailey RC, Schwindt J (1999) Temporal variability of stream bioassessments using benthic macroinvertebrates. Freshwater Biol 42:575–584
Logan P, Furse M (2002) Preparing for the European Water Framework Directive – making the links

between habitat and aquatic biota. Aquat Conserv 12:425–437

Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resource variables: a review of modelling issues and applications. Environ Modell Softw 15:101–124

Manel S, Williams HC, Ormerod SJ (2001) Evaluating absence-presence models in ecology: the need to account for prevalence. J Appl Ecol 38:921–931

Marshall J, Hoang H, Choy S, Recknagel F (2002) Relationships between habitat properties and the occurrence of macroinvertebrates in Queensland streams (Australia) discovered by a sensitivity analysis with artificial neural networks. Verh Internat Verein Limnol 28:1415–1419

Ministry of the Flemish Community (2000) Manual for the characterisation of sediments in Flemish watercourses through the TRIAD approach, second revised print (in Dutch). Administration Environment, Nature, Land and Water management (AMINAL), in cooperation with the Flemish Environment Agency (VMM), Brussels

Obach M, Wagner R, Werner H, Schmidt HH (2001) Modelling population dynamics of aquatic insects with artificial neural networks. Ecol Model 146:207–217

Olden JD, Joy MK, Death RG (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol Model 178:389–397

Paruelo JM, Tomasel F (1997) Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. Ecol Model 98:173–186

Randin CF, Dirnböck T, Dullinger S, Zimmermann NE, Zappa M, Guisan A (2006) Are niche-based species distribution models transferable in space? J Biogeogr 33:1689–1703

Reece PF, Reynoldson TB, Richardson JS, Rosenberg DM (2001) Implications of seasonal variation for biomonitoring with predictive models in the Fraser River catchment, British Columbia. Can J Fish Aquat Sci 58:1411–1417

Roadknight CM, Balls GR, Mills GE, Palmer-Brown D (1997) Modeling complex environmental data. IEEE T Neural Networ 8:852–862

Rosillon D (1989) The influence of abiotic factors and density-dependent mechanisms on between-year variations in a stream invertebrate community. Hydrobiologia 179:25–38

Ruse LP (1996) Multivariate techniques relating macroinvertebrate and environmental data from a river catchment. Wat Res 30:3017–3024

Schleiter IM, Borchardt D, Wagner R, Dapper T, Schmidt KD, Schmidt HH, Werner H (1999) Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. Ecol Model 120:271–286

Schleiter IM, Obach M, Borchardt D, Werner H (2001) Bioindication of chemical and hydromorphological habitat characteristics with benthic macro-invertebrates based on artificial neural networks. Aquat Ecol 35:147–158

Tachet H, Richoux P, Bournaud M, Usseglio-Polatera P (2002) Invertébrés d'eau douce. Systématique, biologie, écologie. CNRS Editions, Paris

VMM (2003) Water quality – water discharges 2002. Flemish Environment Agency, Aalst, Belgium

Walley WJ, Fontama VN (1998) Neural network predictors of average score per taxon and number of families at unpolluted sites in Great Britain. Water Res 32:613–622

Weigel BM, Wang L, Rasmussen PW, Butcher JT, Stewart PM, Simon TP, Wiley MJ (2003) Relative influence of variables at multiple spatial scales on stream macroinvertebrates in the Northern Lakes and Forest ecoregion, U.S.A. Freshwater Biol 48:1440–1461

Witten IH, Frank E (2000) Data mining. Practical machine learning tools and techniques with Java implementations. Academic Press, San Diego

Wright JF (2000) An introduction to RIVPACS. In: Wright JF, Sutcliffe DW, Furse MT (eds) Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK