Aquatic Ecology (2007) 41:413–425
DOI 10.1007/s10452-005-9016-0

Macro-invertebrates in a dynamic river environment: analysis of time series from artificial substrates, using a 'white box' neural network modelling method

N. G. Jaarsma[1,*], M. Bergman[2,3], F. H. Schulze[2] and A. Bij de Vaate[4]
*[1]Ecology group, Witteveen + Bos consulting engineers, 7400, AE, Deventer, The Netherlands; [2]Statistics and neural networks group, Witteveen + Bos consulting engineers, 7400, AE, Deventer, The Netherlands; [3]Uni Network Consultancy B.V. (UNC), 1019, PD, Amsterdam, The Netherlands; [4]Ministry of transport and public Works, Institute for Inland Wastewater Management and Wastewater Treatment, 8200, AA, Lelystad, The Netherlands; *Author for correspondence (e-mail: n.jaarsma@witbo.nl)*

Received 17 March 2004; accepted in revised form 14 September 2005

Key words: *Corophium curvispinum*, *Dikerogammarus villosis*, Invasive taxa, Peak discharge, Product unit networks, RF5 algorithm, Rhine

Abstract

New statistical modelling methods, such as neural networks (NNs), allow us to take a step further in the understanding of complex relations in aquatic ecosystems. In this paper the results from the analysis of macro-invertebrate communities in a complex riverine environment are presented. We attempted to explain observed changes in species composition and abundance with neural network modelling methods and compared the results to linear regression. The NN method used is an improved form of the RF5 algorithm, developed to effectively discover numeric laws from data. RF5 uses Product Unit Networks (PUNs), which are in effect multivariate non-discrete power functions. The data set consisted of a 10-year time series of monthly samples of macro-invertebrates on artificial substrates in the rivers Rhine and Meuse in the Netherlands. During this period the invertebrate community has largely changed coinciding with the invasion of Ponto-Caspian crustaceans. We used physical–chemical data and data on the abundance of the invasive taxa *Corophium curvispinum* and *Dikerogammarus villosis* to explain the observed changes in the resident invertebrate community. The analyses showed temperature, abundance of invasive taxa and peak discharges as important factors. Comparison of the results from NN modelling to linear regression revealed that the factors temperature and abundance of *Dikerogammarus villosis* explained equally well in both cases. Only the neural network was able to use information on peak discharge and timing of the peak in the previous winter to improve model performances. Neural networks are known to yield excellent modelling results, a drawback however is their lack of transparency or their 'black box' character. The use of relatively easy interpretable (white box) PUNs allows us to investigate the extracted relations in more detail and can enhance our understanding of ecosystem functioning. Our results show that peak discharges might be an important factor structuring invertebrate communities in rivers and hint on the existence of interacting effects from invasive species and discharge peaks. They finally show the value of biological data sets that are collected over a long period and in a highly standardised way.

Introduction

Aquatic communities in Dutch rivers have been subject to human influences like river regulation and pollution for centuries (Bij de Vaate 2003). More recently, a large-scale human induced impact of these ecosystems is the invasion by non-indigenous species. Important vectors for these species are ships' ballast water and invasions from the Ponto-Caspian area after the construction of the Main-Danube canal (Bij de Vaate et al. 2002). Some of these species have had major impacts on autochthonous fauna (Van den Brink et al. 1993; Van der Velde et al. 1994). Understandably, this has implications for ecological rehabilitation of rivers. A side effect is that it influences ecological water quality assessment as required by the EU Water Framework Directive (Directive 2000/60/EC). The aim of ecological assessment is to show the effects of human impacts on the aquatic ecosystem. However to quantify these anthropogenic effects, knowledge is needed of the ways in which the environment influences ecological processes and ultimately species composition. Much work still has to be done in this field, especially in complex situations where combined effects of multiple factors play a role. New computational techniques like neural networks can help us qualify and quantify these complex relations.

We used linear and neural network modelling methods to analyse the factors influencing the changes in invertebrate community composition on artificial substrates from 1992 to 2001. During this period major invasions of Ponto-Caspian invertebrate species have occurred. Some of these species have been so successful that they dominate the native invertebrate community. The most spectacular invasions in recent years are by the crustacean's *Corophium curvispinum* and *Dikerogammarus villosis*. These species have had a great impact on the resident taxa; the filter feeding *C. curvispinum* by covering hard substrata in the river bed with muddy tubes (van den Brink et al. 1993) and *D. villosis* because of its predatory behaviour (Dick and Platvoet 2000). Data on the abundance of these taxa and the whole invertebrate community was used to compare linear models to non-linear models that were fitted by neural network methods. There are already quite a few examples of applications of NN to data on macro-invertebrates (e.g. Cereghino et al. 2001; Park et al. 2003a; Park et al. 2003b) and fish (e.g. Lek et al. 1996; Guégan et al. 1998; Brosse et al. 1999; Brosse et al. 2001; Reyjol et al. 2001; Ibarra et al. 2003). Most studies use feed-forward neural networks or multi-layer perceptrons (MLPs). These are supervised NNs, both input as well as targets are presented to the network. An often-mentioned drawback of the use of MLPs is that these methods are basically considered 'black box' methods. For this reason there have been some attempts to find methods to gain explanatory insight into the contributions of each variable (e.g. reviews by Olden and Jackson 2001; Olden and Jackson 2002; Gevrey et al. 2003). However, the contributions of the explanatory variables remain rather implicit, lacking interpretable estimated relations such as with linear regression models.

We used an improved form of the RF5 algorithm (Saito and Nakano 1997a; Saito and Nakano 1997b; Oost et al. 2002) for NN modelling. This uses so-called Product Unit Networks (PUNs) for equation discovery. The resulting functions are multivariate non-discrete power functions that are reasonably comprehensible and can hence be called 'white box'. PUNs distinguish themselves from traditional NNs such as MLPs because of the relatively easy interpretation of the extracted relations. The relations extracted from the data can help us identify and understand complex relations between species and their environment. Therefore, throughout this paper we focus mainly on the ecological interpretation of the output from both linear and NN methods. The main question is: can we understand what these methods come up with? In the discussion we present a brief synthesis of the results and focus on the added value and applicability of the PUNs with respect to the interpretability of estimated mathematical functions.

Methods

Macro-invertebrates

The data set consists of samples from artificial substrates of four sites in the rivers Rhine and Meuse in the Netherlands. Figure 1 shows the locations of the two sampling sites in the lower river Rhine, at

Figure 1. The four sampling sites of macro-invertebrates in the rivers Rhine (Lobith and Kampen) and Meuse (Borgharen and Grave) in the Netherlands.

the towns of Lobith and Kampen, and the two sites in the river Meuse, at the towns of Borgharen and Grave. The artificial substrates consisted of iron cases containing marbles that were left suspended in the river for colonisation during a period of 1 month. After a month the samples were taken out, washed and the invertebrates collected and fixed in formalin. Invertebrates were identified to the lowest taxonomic level possible, species level in most cases. The samples were collected monthly (two samples per site) from spring to autumn (April–October) from 1992 to 2001. In total this yielded between 120–130 samples per site. For the analysis the abundance of taxa has been transformed using $\ln(x+1)$ transformation.

Physical–chemical measurements

At different sampling sites in the rivers, physical and chemical measurements were taken routinely on different time intervals, ranging from hourly for discharge to weekly for chlorophyll-*a*. For the analysis the measurements have been transferred to the average, minimum and maximum of the previous month. In the case of discharge also the peak discharge of the previous year has been used (this is the highest discharge in previous 365 days, usually in winter). Since not only the peak discharge but also the moment it occurred might be of importance, the number of days between the peak discharge and the sampling date has been taken into account. Table 1 gives an overview of the variables that were used for the analysis, their measuring unit and the number of samples or measurements per year.

Statistical methods

Clustering and ordination

We used the software program FLEXCLUS (Van Tongeren 1986) to group samples into clusters by calculating similarities between samples containing information on abundance or presence/absence of taxa. For ordination we used the software package CANOCO version 4.02 (Ter Braak and Smilauer 1999). Indirect ordination was used to investigate the major source of variation in community composition and direct ordination to relate it to environmental characteristics. We used (canonical) correspondence analysis, which assumes unimodal (Gaussian) responses of species to environmental

Table 1. Average values of selected physical and chemical variables at the sampling sites 'Lobith' in the river Rhine and 'Eijsden' in the river Meuse during the period 1990–2001. The table also shows the number of measurements per year, which is set at 365 in the case of multiple daily measures.

Variable	Unit	Rhine	Meuse	Number of measurements per year
Cadmium	µg/ll	0.07	0.34	25–52
Chlorophyll-*a*	µg/ll	9.3	12.8	25–52
Discharge	m^3/s	2354	272	365
Mercury	µg/ll	0.03	0.03	13–52
NH4–N	mg/l	0.16	0.46	25–52
o-PO4-P	mg/l	0.09	0.30	25–52
Oxygen	%	96	81	52–340
Oxygen	mg/l	10.2	8.7	320–362
Pentachlorophenol	µg/l	0.01	0.02	12–52
Secchi-depth	dm	5.4	7.4	1–52
Silicate	mg/l	2.24	2.39	25–52
Sodium	mg/l	72	27	13–27
Sulfate	mg/l	61	42	25–52
Total organic carbon	mg/l	4.3	5.2	12–52
Total-P	mg/l	0.21	0.44	25–52
Water temperature	°C	14.0	14.5	365

factors. To test for unimodality we determined gradient length by running a Detrended Correspondence Analysis (DCA), in case of a gradient length < 3 SD generally a linear model is used (Ter Braak and Smilauer 1998). Both direct (DCCA) and indirect (DCA) ordinations have been carried out. With direct ordination environmental factors are directly related to species composition, thereby influencing ordination scores. With indirect ordination sample scores are calculated on species composition alone, environmental factors are afterwards related to the scores of samples on the ordination axes. For this we used multiple techniques (see the analysis section).

Multiple linear regression modelling

Multiple linear regression or linear regression is a common statistical modelling method. Applications of linear regression in ecosystems are useful if the ecosystem is adequately understood and can, hence, be described in (transformed) linear relations. For linear regression we used the software package SPSS11.0.

Neural network modelling

The algorithm used for neural network modelling is an improvement of the existing RF5 (Rule extraction from Facts version 5) algorithm, which was developed to effectively discover numeric laws from numeric data. The original RF5 algorithm (Saito and Nakano 1997a, 1997b) consists of a combination of three techniques;

–Using Product Unit Networks (PUNs) for function approximation, resulting in multivariate non-discrete power functions
–Training them with the BPQ optimisation algorithm (a second-order learning algorithm) and
–Selecting the number of hidden nodes with the MDL (Minimum Description Length) metric.

This procedure iteratively continues, until an optimal PUN is estimated based on the MDL score. A product unit (neuron) is defined as

$$\prod_{i=1}^{i=I} X_i^{p_i} \text{ instead of the regular}$$

$$\text{MLP summation unit} \sum_{i=1}^{i=I} w_i \cdot X_i$$

in which p is a power weight, w a multiplicative weight and $i = I$ the number of inputs. In general a Product Unit Network (PUN) can be described as

$$Y = \sum_{j=1}^{j=J} w_j \prod_{i=1}^{i=I} X_i^{p_i}$$

in which $j = J$ is the number of units. PUNs can approximate many relatively comprehensible non-linear relations, if-then-else constructions and interactions directly related to input variables. For more details we refer to Saito and Nakano (1997a and 1997b).

The RF5 algorithm as described above has been successfully applied to small data sets. To optimise the comprehensibility of the results and to be able to use it on larger data sets, some improvements have been made (Oost et al. 2002). First, an estimated PUN can be simplified using a pruning algorithm to reduce irrelevant connections (parameters). Since most pruning algorithms have been created for MLPs, a new pruning algorithm is used that is specifically designed for pruning single weights from PUNs (Oost et al. 2002). This method is called the Enhanced Sensitivity-based Pruning (ESP) method (Moody and Utans 1992). Other improvements were introduced as well, for instance the use of Levenberg–Marquardt as a search algorithm instead of BPQ. For more details we refer to (Oost et al. 2002). The improved existing RF5 algorithm is programmed in the software package Matlab R13.

Analysis

The aim of our study was to identify the major factors influencing invertebrate community composition in Dutch rivers and in doing so to compare neural networks to linear models. For this we used the methods and data that were mentioned in the previous sections. From the invertebrate data set we removed the recently invaded crustaceans *Corophium curvispinum* and *Dikerogammarus villosis* and used the $(\ln(x+1))$ transformed abundances of these species as explanatory factors instead. The reason for this is that they are thought to have a major impact on community composition, using them as explanatory factors allows us to investigate the strength of this impact. For the preliminary analysis of the full data set

(four locations) we also used the variables *river* (Rhine or Meuse), *location* (distance from country border), *day* (day of sampling from January 1) and *year* (year of sampling). These were used to provide us with information on the influence of river specific, location specific and seasonal variations and trends in time.

For the analysis we used multiple combinations of techniques. For our preliminary analysis we used clustering and ordination of the full data set to identify the major source of variation and the most important factors explaining this variation. For the analysis of the abundance of a single species we used linear regression and neural network modelling. For the analysis of the whole community we used combinations of linear regression and neural networks with detrended correspondence analysis.

Results

Preliminary analysis of the full data set

First we analysed the full data set (4 locations, 504 samples) by clustering and ordination. Clustering yielded three groups of samples (Table 2). The results show that the major variation in the invertebrate community can be attributed to location specific differences, all samples taken from sampling station 'Borgharen' clearly differ from the other three sampling stations. The next largest variation can be attributed to the year of sampling. For the three remaining stations the samples taken after 1995 to 1997 (depending on location) clearly differ in species composition from the earlier ones. This implies that on all three sampling sites drastic changes have occurred around 1995–1997. Direct ordination (not presented here) revealed that parameters identifying river (Rhine or Meuse), location (distance from country border), year of sampling, the abundance of invasive taxon *D. villosis* and temperature could explain the major variation in species composition. Since the major source of variation could be attributed to location specific differences, we chose to use the data from sampling station 'Lobith' (Rhine) for further analyses. This allows us to focus on changes in the invertebrate community in time.

Table 2. Results of the clustering of the macro-invertebrate data set. For each sampling site the total number of samples (n) and the number of samples per cluster is given. The clusters that emerged from the analysis are characterised by sampling site and year of sampling.

Sampling site	cluster 1: Borgharen all years	cluster 2: other sites 1992–1995/6	cluster 3: other sites 1996/7–2001
Borgharen ($n=130$)	130		
Grave ($n=120$)		63	57
Kampen ($n=128$)		66	62
Lobith ($n=126$)		46	80

Application of linear models and neural networks to single species

The changes in the abundance of the invasive taxon *C. curvispinum* reflect the changes of the whole invertebrate community on the location Lobith in the river Rhine. This species can be seen as a key stone species. It dominated the invertebrate community since 1988 until after 1995 a sudden drop in the abundance of this taxon and in the total invertebrate community occurred. More or less at the same time new taxa invaded the river-ecosystem. *D. villosis* probably represents the most important of these, being a large and voracious predator (Dick and Platvoet 2000). We tried to link the changes in the abundance of the former dominant *C. curvispinum* to the abundance of *D. villosis* and a number of environmental factors. Table 3 shows the results of the analysis by both linear and NN modelling for the training set ($n=100$) and a small test set ($n=6$) of randomly chosen samples. Because the improved RF5 algorithm yields multiple solutions, from simple to complex models with low to high performance, we only present a selection of the results. This selection is made on the basis of simplicity and performance (r^2) and ranges from single factor models to models with four factors. We used the same factors to construct linear regression models. The high r^2 for the test in Table 3 set is, in this case, a result of the small number of samples ($n=6$) and the absence of extreme values.

The analysis reveals that the most important factor explaining the abundance of *C. curvispinum* is temperature. This factor alone explains roughly 60% of the total variation in taxon abundance. Temperature reflects the seasonal variation in

Table 3. Results for the linear model (LIN) and the product unit network (PUN) from the analysis of *Corophium curvispinum* abundance in relation to temperature, abundance of *Dikerogammarus villosis*, peak-discharge and time (number of days since peak). The table shows the percentage of variance explained (r^2) and root mean square error (RMSE) for the training set ($n = 100$) and test set (between brackets, $n = 6$).

Model	r^2		RMSE	
	PUN	LIN	PUN	LIN
temperature	59 (91)	57 (91)	1.30 (0.64)	1.33 (0.64)
added *Dikerogammarus*	78 (88)	75 (88)	0.96 (0.74)	1.02 (0.73)
added peak discharge	79 (90)	75 (87)	0.95 (0.66)	1.01 (0.74)
added time	83 (94)	76 (88)	0.83 (0.52)	1.00 (0.72)

production and decomposition of the community and is in that respect an indicator of ecological processes and a number of relevant variables like light and food availability. The fact that for both techniques temperature explains equally well shows that the relation is approximately linear. The next important factor explaining the variation in abundance of *C. curvispinum* is the abundance of *D. villosis*. This species was first recorded in the river near Lobith around 1994 and became abundant in 1995, corresponding to a dramatic decline in *C. curvispinum* and overall taxon abundance. Peak discharges in the winter of 1994/1995 caused severe flooding of large areas along the rivers Rhine and Meuse. These events have had a catastrophic effect on the invertebrate community by way of a major washout of individuals. It seems that in this – after disturbance – climate, *D. villosis* was able to quickly colonise and establish itself seemingly at the expense of *C. curvispinum* and other species. That this might be the case is supported by the observations of voracious predatory behaviour of *D. villosis* in the Rhine system, whereas this species is mainly detrivorous in its original habitat (Dick and Platvoet 2000). The importance of peak discharge itself is finally tested by adding the variables peak discharge in the previous year and time elapse since this peak occurred. Table 3 shows the results of adding these variables to the model and concludes that only PUNs can use this information to explain another

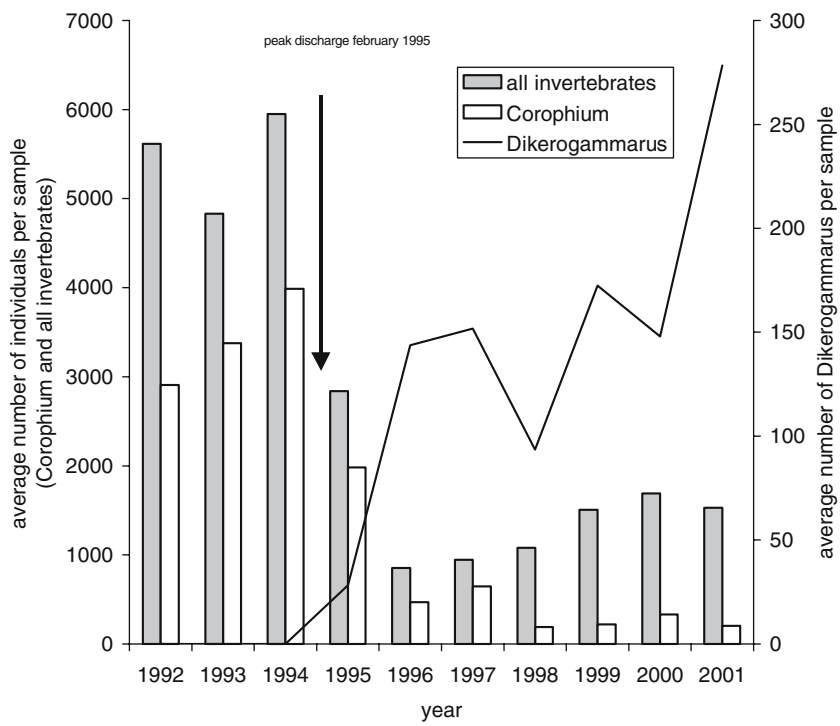

Figure 2. Variation in the average number of all invertebrates and *Corophium curvispinum* (left axis) and the average number of *Dikerogammarus villosis* (right axis) from 1992 to 2001. After the peak-discharge in 1995 a dramatic drop in overall abundance of invertebrates takes place, coinciding with a steady increase in *D. villosis* abundance.

part (±5%) of the remaining variation. Figure 2 shows the developments graphically. In the figure the drop in overall taxon abundance and abundance of *C. curvispinum* after the floodings in early 1995 are plotted against the increase in abundance of and *D. villosis*. Figure 3 compares the results of the linear model to the PUN, arrows indicating discharges larger than 9000 m^3/s. The largest deviations of the linear model occur after or around these high discharges, indicating that a PUN is better able to explain these from data on peak discharges. Note that the deviation of the linear model after a high discharge is not always the same. This might be caused by combined effects of peak discharge, timing and *D. villosis*.

Application of linear models and neural networks to the whole community

In order to apply the modelling techniques to data from whole invertebrate communities it is convenient to capture variation in community composition into one or a few parameters. The relations between environment and these parameters – as abstractions of the whole community – can then be identified. A method that is widely used for such a purpose is ordination. This technique is able to extract hypothetical gradients from large data sets containing many taxa with their abundances (Ter Braak and Smilauer 1998). Ordination yields species and sample scores that can be plotted in ordination diagrams. Recently the Kohonen Self Organising Map (SOM), an unsupervised neural network, has been used for the purpose of community ordination. Application of SOM on ecological data sets has shown results similar to those obtained with conventional statistical community ordination methods (Brosse et al. 2001; Cereghino et al. 2001; Giraudel and Lek 2001; Park et al. 2003a, 2003b). However, for purposes of directly relating environmental variables to communities, traditional ordination is a proven and straightforward method.

To calculate the ordination scores we ran a DCA on the invertebrate data from Lobith (excluding *C. curvispinum* and *D. villosis*). Figure 4 shows the ordination diagram depicting sample scores and environmental variables. Again we used temperature, peak discharge, time (number of days since peak) and the abundance of *D. villosis* as explanatory variables and added the ($\ln(x+1)$ transformed) abundance of *C. curvispinum*. Due to some missing values for environmental data, fig-

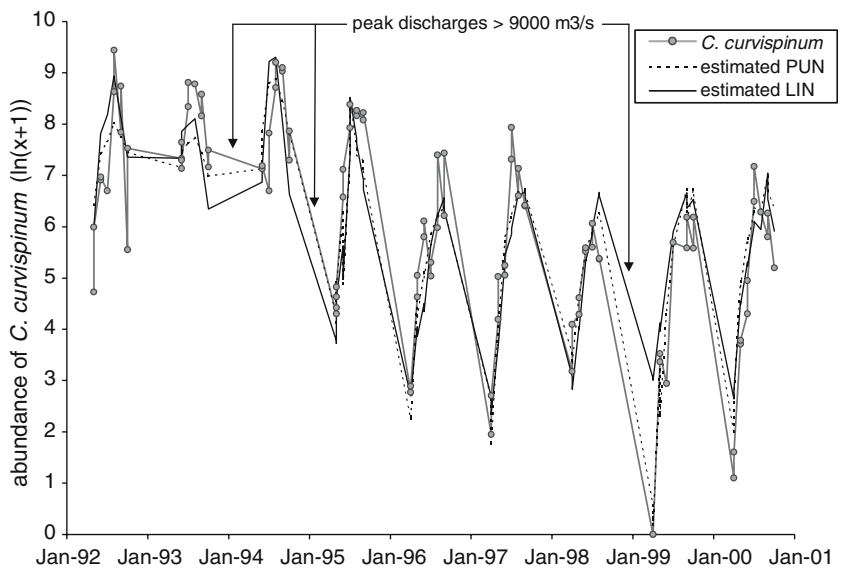

Figure 3. Graphical comparison of the results from the linear model (LIN) to the product unit network (PUN). The graph shows the ($\ln(x+1)$ transformed) real *Corophium curvispinum* abundance and the modelled abundance of *C. curvispinum* as a function of temperature, abundance of *Dikerogammarus villosis*, peak-discharge and time (number of days since peak). In the graph, discharges larger than 9000 m^3/s are indicated with arrows. These correspond to the largest deviations of the linear model from the observed abundance of *C. curvispinum*.

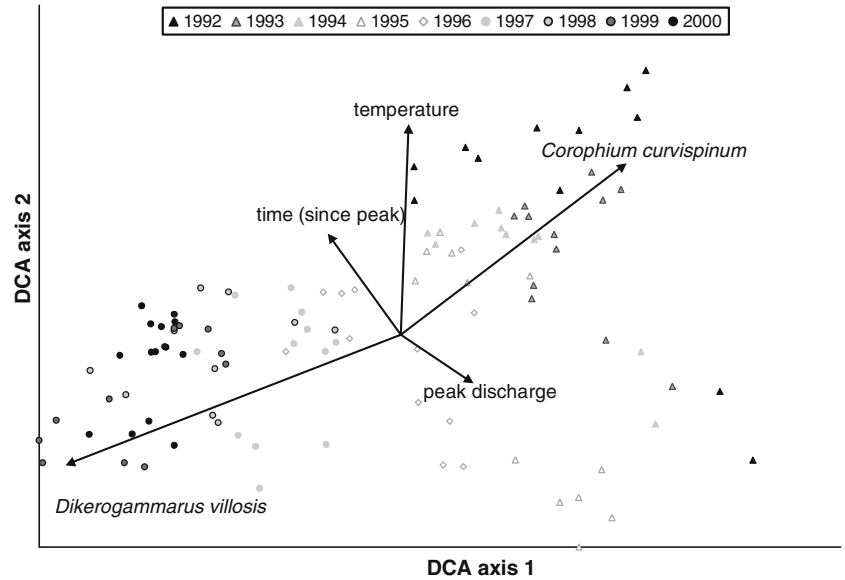

Figure 4. Indirect ordination (DCA) of the macro-invertebrates from sampling site Lobith. The dots showing the sample scores on the first and second ordination axis in relation to the environmental factors (arrows) that were used to explain variations in species composition. The year of sampling is indicated by the shape and colour of the symbol.

ure 4 shows a somewhat smaller subset ($n = 106$) than the original ordination ($n = 126$).

To compare linear models to PUNs we related the environmental factors to the calculated scores on the ordination axes. Table 4 summarises the results. In this case the major source of variation is explained by the abundance of *D. villosis*, this alone explains approximately 50% (training set) and 38–39% (test set) of the total variation on the first and most important axis. Adding the time of the peak and peak discharge itself gives clearly better results for the PUN (57% training, 47% test set) but hardly for the linear model (52 and 38% respectively). This suggests that peak discharge in the previous winter is an important, non-linear factor determining community composition and abundance on artificial substrates in the River Rhine near Lobith. When temperature and the abundance of *Corophium curvispinum* are added, this results in improved performance for both models (73 and 61% respectively). The results presented here are for the PUNs that yielded the highest r^2 and lowest RMSE for both the training-set and the test set. Models that performed better on the training set (highest r^2 = 89%) predicted the test set poorly due to a few (extreme) low discharge events in this set. The total number of observations that the model is based on (95) limits the complexity (risk of over-fitting), more data are needed to properly fit more complex models. The overall results however show that the PUNs are better able to explain the variation in invertebrate community composition from the explanatory variables.

Table 5 summarises the formulas from linear and NN models. Comparison reveals that the simplest models, that only take into account the

Table 4. Percentage of variance explained (r^2) and root mean square error (RMSE) for linear (LIN) and product unit network (PUN) models from the analysis of the ordination-scores of the first axis in relation to temperature, abundance of *Dikerogammarus villosis and Corophium curvispinum*, peak-discharge and time of the peak (number of days since peak). The table shows the percentage of variance explained (r^2) and root mean square error (RMSE) for the training set ($n = 95$) and test set (between brackets, $n = 11$).

Model	r^2		RMSE	
	PUN	LIN	PUN	LIN
Dikerogammarus	50 (39)	50 (38)	0.53 (0.60)	0.53 (0.61)
added time	50 (38)	52 (38)	0.53 (0.61)	0.52 (0.61)
added peak and time	57 (47)	52 (38)	0.49 (0.56)	0.52 (0.61)
added temperature	58 (51)	53 (41)	0.48 (0.54)	0.52 (0.59)
added *Corophium*	73 (65)	61 (51)	0.39 (0.46)	0.47 (0.54)

Table 5. Linear models and product unit networks from the analysis of the ordination-score of the first axis in relation to temperature [temp], abundance of *Dikerogammarus villosis* [DV] and *Corophium curvispinum* [CC], peak-discharge [peak] and number of days since peak [time]. For modelling purposes, values for zero abundance of *D. villosis* and *C. curvispinum* are 0.1 for all models.

Model	Linear	
Dikerogammarus	2.04−0.24*[DV]	
added time	2.33−0.25*[DV]−0.0013*[time]	
added peak	2.32−0.25*[DV]−0.0013*[time] + 0.0000019*[peak]	
added temperature	2.67−0.26*[DV]−0.0010*[time] + 0.0000041*[peak]-0.021*[temp]	
added *Corophium*	2.54−0.17*[DV]−0.0014*[time]−0.00000182*[peak]−0.092*[temp] + 0.22*[CC]	
Model	Product unit network	Nodes
Dikerogammarus	$2.02-0.17*[DV]^{1.23}$	1
added time	$2.02-0.14*[DV]^{1.22}*[time]^{0.034}$	1
added peak	$8.09-4.28*[DV]^{0.25}*[peak]^{0.0063}-0.12*[DV]^{-0.48}*[time]^{0.31}*[peak]^{0.066}$	2
added temperature	$8.52-5.36*[DV]^{0.424}*[time]^{-0.12}*[peak]^{-0.020}-0.57*[DV]^{-0.16}*[time]^{0.22}*[peak]^{0.043}*[temp]^{0.12}$	2
added *Corophium*	$36667-56315*[temp]-^{0.0074}*[CC]^{0.0032}*[DV]^{0.0011}*[peak]^{0.0012}*[time]^{0.0027} + 19178*[temp]^{-0.021}*[CC]^{0.0092}*[DV]^{0.0032}*[peak]^{0.0045}*[time]^{0.0076} + 586*[peak]^{-0.085}$	3

abundance of *D. villosis* and timing of peak discharge, are approximately linear. Subsequent models also taking peak discharge, temperature and *C. curvispinum* into account are more complex. However, examining each node (product unit), for example by plotting it as a separate response to the factors incorporated in the model, makes them easier to interpret and reveals the influence of separate (combinations of) factors. In many cases the response of a node is largely determined by only one or two factors.

Further interpretation of Table 5 reveals that the full model, incorporating all five variables, seems to distinguish between situations where *D. villosis* is present and cases where this species is absent. Figure 5 shows the response of the inver-

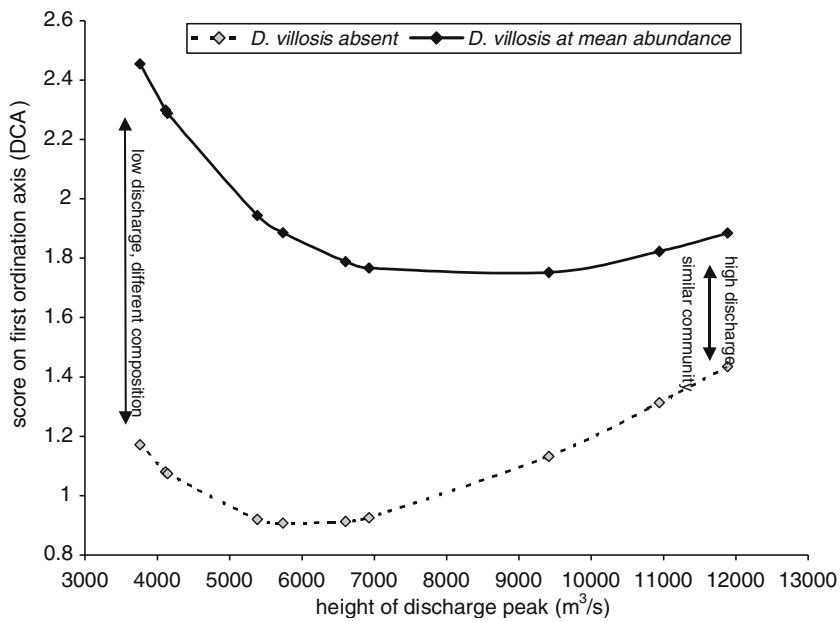

Figure 5. Graphical representation of the modelled effect of the height of discharge peaks on community composition (axis score) in the absence of *Dikerogammarus villosis* (dotted line) and when this species is present at mean abundance (solid line). The figure is derived from the full model that uses all variables from Table 5b.

tebrate community (score on first ordination axis) as a function of the highest discharge in the previous year, while all other factors are kept constant at their mean values. Dependent on the presence (mean value) or absence of *D. villosis*, the curves have a different shape. Interestingly the outcome of the model differs quite a lot in the case of low and intermediate discharges, whereas the model predicts a more similar community composition (ordination score) at high discharges.

Application of linear regression and NN techniques to the scores on the second ordination axis (not further presented here) revealed temperature as the most important factor (linear models explain about 70%, PUNs about 80% of total variation on the second axis).

Discussion

The analysis of the time-series of macro-invertebrates in Dutch rivers yielded interesting results. The frequency and time span of sampling allowed us to gain insight in seasonal variations as well as trends in time. The seasonal variation in single species and whole communities is well explained by temperature as an indicator of ecological processes and other season-related variables. The major trend in time is illustrated by the invasion of Ponto-Caspian invertebrate taxa. The results suggest that some of these taxa have had an enormous impact on the resident (autochthonous and allochthonous) invertebrate taxa. These findings are supported by literature, e.g. Van den Brink et al. (1993) and Dick and Platvoet (2000) who describe the invasions of *C. curvispinum* and *D. villosis* and their possible impacts on river food webs. Apart from these factors a third and probably very important factor that emerged from the analysis, is peak discharge. Both the height of the peak and the moment at which it occurred seem to play an important role in this. After washout of invertebrates during a disturbance event, recolonisation and inter- and intraspecific interactions such as predation and competition determine community composition. In our example we found different effects of peak discharge in the presence or absence of *D. villosis*. This is an important result that also makes sense ecologically; the general effect of a peak discharge is washout of invertebrates causing an overall decline in taxon abundance. The effect of a peak discharge in the presence of *D. villosis* might be different because after a disturbance and subsequent washout, the development of the community might be strongly influenced by this voracious predator. The peak discharge resets the community and the pathway to a new (stable) situation is determined by the relative success of a single species to colonise a habitat, escape predation and compete for resources and space. These findings are supported by earlier work that describes the effects of hydrological disturbance and the effects of the introduction of new species. Townsend et al. (1998), Jaarsma et al. (1998) and Townsend and Riley (1999) investigated the effects of disturbances, and timing of disturbance (Townsend et al. 1997), on stream food webs. They found significant effects on the web complexity, the most frequently disturbed sites being the least taxon rich and having the simplest webs. Lancaster (1996) mentions changes in the competitive strength of two invertebrate predators after a disturbance. Wootton et al. (1996) found a 77% reduction in the abundance of invertebrate taxa after a brief spate. Mulholland et al. (1991) found that the response of periphyton communities to a disturbance could be explained by interactions between disturbance, nutrient availability and grazing. Finally, Fausch et al. (2001) found a relation between the flood disturbance regime and the invasion success of rainbow trout. Success was highest in rivers with a disturbance regime that matched those in their native range. Thus, the disturbance regime might affect habitat availability, inter-specific competition and predation, invasion success and ultimately the river-community on different temporal and spatial scales. The effects of disturbance might be enhanced in an environment that is under severe stress of human and human-induced impacts like river regulation and invasions of non-indigenous species. These interacting effects have to be taken into account when taking measures for river restoration or designing methods for ecological assessment.

In this paper we present the results from an improved RF5 algorithm, which is designed to obtain understandable rules from PUNs. To our knowledge the use of supervised neural networks in aquatic ecology is mostly restricted to feed

forward networks like MLPs. These often yield very good results in the amount of variance explained but are often considered 'black box' models because they provide little explanatory insight in the relative contribution of each variable. Attempts to 'illuminate' the black box have been made (e.g. Olden and Jackson 2001; Olden and Jackson 2002; Gevrey et al. 2003). These authors used various approaches like neural interpretation diagrams or sensitivity analysis to assess the contribution of each variable. Although they might improve explanatory power greatly, none of these methods provides directly estimated, interpretable formulas. PUNs are more 'user friendly' in that they give more transparent (though still complex) functions. For instance, non-linearity is incorporated at explanatory variable scale, using (non-discrete) power weights, instead of transfer functions at unit scale in the case of MLPs. This allows for a more detailed analysis of the extracted relations, which will help us understand ecosystem functioning better.

For instance, in our analysis of the whole invertebrate community at Lobith, two separate nodes emerged describing the effects of a peak discharge. One node (interaction of variables) describes the general effect, the other describes the effect in the presence of *D. villosis*. In our opinion PUNs provide a valuable tool for analysing these kinds of complex interactions between explanatory variables.

Like MLPs, PUNs are able to describe all kinds of different relations, from linear to logistic, quadratic etc. The flexibility of these methods ensures that complex relations can be extracted, given the fact that the right explanatory variables are offered to the NN. The other way around, when NNs predict poorly this might be because important factors are overlooked. NN modelling may provide hints where to look for these factors.

Combinations of traditional and new techniques, for example community ordination (PCA or CA) and NNs, offer possibilities to gain a more subtle understanding of ecosystems than a 'one-way' approach. Knowledge of the factors underlying ecological processes is always necessary to decide whether results are valid. If some of these relations are already known, this knowledge can be combined with the use of NNs.

Conclusions

In this paper we have shown that supervised NNs, that are trained using the improved RF5 algorithm, can provide us with a valuable tool to qualify and quantify relations between species and their environment. In the examples shown, the effects of a dynamic environment (discharge variations) and the impact of invasive taxa on the invertebrate community are complexly interwoven. It is promising that even in such a dynamic environment, empirical models, such as the estimated PUNs, are able to extract the major factors underlying the observed changes in community composition. The flexibility of the improved RF5 algorithm combined with the possibilities to interpret the extracted relations makes it both a useful method to explore ecological data and to quantify relations. Therefore we believe that the improved RF5 algorithm can help us to gain new insights in ecosystem functioning, in particular in the complex pathways in which the environment affects species composition. To successfully use analysing techniques in general, large data sets are needed that have been sampled in a standardised way. Problems often encountered using large data sets are differences in the taxonomic determination level, faults in taxon identification or habitat sub-sampling. These differences introduce noise, thereby obscuring the real variation. It remains to be seen if such data sets are or will become available. For the successful application of conventional or new techniques and a further unravelling of ecological processes, the availability of high quality data might very well prove to be the bottleneck. This is at the same time a justification for putting effort into standardised sampling and a challenge to develop highly standardised sampling methods.

Acknowledgements

We thank two anonymous reviewers for their constructive remarks on our original manuscript, this has greatly improved our paper. Funding for this project was largely provided by Witteveen + Bos consulting engineers.

References

Bij de Vaate A. 2003. Degradation and recovery of the freshwater fauna in the lower sections of the rivers Rhine and Meuse. Thesis, Wageningen University. Wageningen, The Netherlands.

Bijde Vaate A., Jazdzewski K., Ketelaars H.A.M., Gollasch S. and van der Velde G. 2002. Geographical patterns in range extension of Ponto-Caspian macro-invertebrate species in Europe. Can. J. Fish. Aquat. Sci. 59: 1159–1174.

Brosse S., Giraudel J.L. and Lek S. 2001. Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. Ecol. Model. 146: 159–166.

Brosse S., Guégan J.F., Tourenq J.N. and Lek S. 1999. The use of artificial neural network to assess fish community structure in a mesotrophic lake. Ecol. Model. 120: 299–311.

Cereghino R., Giraudel J.L. and Compin A. 2001. Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self organizing maps. Ecol. Model. 146: 167–180.

Dick J.T.A. and Platvoet D. 2000. Invading predatory crustacean *Dikerogammarus villosis* eliminates both native and exotic species. Proc. Roy. Soc. London B 267: 977–983.

Directive 2000/60/EC. Official Journal of the European Communities, L327. (22 December 2000, pp. 1–72).

Fausch K.D., Taniguchi Y., Nakano S., Grossman G.D. and Townsend C.R. 2001. Flood disturbance regimes influence rainbow trout invasion success among five Holarctic regions. Ecol. Appl. 11: 1438–1455.

Gevrey M., Dimopoulos L. and Lek S. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Model. 160(3): 249–264.

Giraudel J.L. and Lek S. 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecol. Model. 146(1–3): 329–339.

Guégan J.F., Lek S. and Oberdorff T. 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391: 382–384.

Ibarra A., Gevrey M., Park Y.-S., Lin P. and Lek S. 2003. Modelling the composition of fish guilds in the Garonne basin (France) with a backpropagation network: an aid to decision-making. Ecol. Model. 160(3): 281–290.

Jaarsma N.G., De Boer S.M., Townsend C.R., Thompson R.M. and Edwards E.D. 1998. Characterising food webs in two New Zealand streams. N. Zeal. J. Mar. Freshwater Res. 32: 271–286.

Lancaster J. 1996. Scaling the effects of predation and disturbance in a patchy environment. Oecologia 107: 321–331.

Lek S., Delacoste M., Baran P., Dimopoulos I., Lauga J. and Aulagnier S. 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90: 39–52.

Moody J. and Utans J. 1992. Principled architecture selection for neural networks: Application to corporate bond rating prediction. In: Moody J.E., Hanson S.J. and Lippmann R.P. (eds), Advances in Neural Information Processing Systems, (4), pp. 683–690.

Mulholland P.J., Steinman A.D., Palumbo A.V., Flum T. and DeAngelis D.L. 1991. Influence of nutrients and grazing on the response of stream periphyton communities to a scour disturbance. J. N. Am. Benthol. Soc. 10: 127–142.

Olden J.D. and Jackson D.A. 2001. Fish-habitat relationships in lakes: Gaining predictive and explanatory insight using artificial neural networks. Trans. Am. Fish. Soc. 130: 878–897.

Olden J.D. and Jackson D.A. 2002. Illuminating the 'black box': Understanding variable contributions in artificial neural networks. Ecol. Model. 154: 135–150.

Oost E.M., Ten Hagen S. and Schulze F.H. (2002). Extracting multivariate power functions from complex data sets. In: Proceedings of the Belgian-Dutch AI Conference (BNAIC-02).

Park Y.-S., Céréghino R., Compin A. and Lek S. 2003a. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecol. Model. 160(3): 265–280.

Park Y.-S., Verdonschot P.F.M., Chon T.-S. and Lek S. 2003b. Patterning and predicting aquatic macroinvertebrate diversities using artificial neural network. Water Res. 37(8): 1749–1758.

Reyjol Y., Lim P., Belaud P. and Lek S. 2001. Modelling of microhabitat used by fishes in natural and regulated flows in the Garonne river (France). Ecol. Model. 146(1–3): 131–142.

Saito K. and Nakano R. 1997a. Law discovery using neural networks. Proceedings of the 15th Int. Joint Conference on Artificial Intelligence (IJCAI-97), pp. 1078–1083, San Francisco.

Saito K. and Nakano R. 1997b. Numeric law discovery using neural networks. Proceedings of the 4th Int. Conference on Neural Information Processing (ICONIP-97), pp. 843–846.

Ter Braak C.J.F. and Smilauer P. 1998. CANOCO Reference Manual and User's Guide to Canoco for Windows (version 4). Centre for Biometry, Wageningen, The Netherlands.

Ter Braak C.J.F. and Smilauer P. 1999. CANOCO for Windows (version 4.02) – a FORTRAN Program for Canonical Community Ordination. Centre for Biometry, Wageningen, The Netherlands.

Townsend C.R. and Riley R. 1999. Assessment of river health: accounting for perturbation pathways in physical and ecological space. Freshwater Biol. 41: 393–405.

Townsend C.R., Scarsbrook M.R. and Dolédec S. 1997. The intermediate disturbance hypothesis, refugia and biodiversity in streams. Limnol. Oceanogr. 42: 938–949.

Townsend C.R., Thompson R.M., McIntosh A.R., Kilroy C., Edwards E. and Scarsbrook M.R. 1998. Disturbance, resource supply and food-web architecture in streams. Ecol. Lett. 1: 200–209.

Vanden Brink F.W.B., van der Velde G. and Bijde Vaate A. 1993. Ecological aspects, explosive range extension and impact of a mass invader, Corophium curvispinum Sars, 1895 (Crustacea, Amphipoda), in the Lower Rhine (The Netherlands). Oecologia 93: 224–232.

Vander Velde G., Paffen B.G.P., Vanden Brink F.W.B., Bijde Vaate A. and Jenner H.A. 1994. Decline of zebra mussel populations in the Rhine: Comparison between two mass

invaders ''(Dreissena polymorpha and Corophium curvispinum)''. Naturwissenschaften 81(1): 32–34.
Van Tongeren O. 1986. ''FLEXCLUS, an interactive flexible cluster program''. Acta Bot. Neerl. 35: 137–142.
Wootton J.T., Parker M.S. and Power M.E. 1996. The effect of disturbance on river food webs. Science 273: 1558–1560.