ELSEVIER

Ecological Modelling 160 (2003) 265–280

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters

Young-Seuk Park *, Régis Céréghino, Arthur Compin, Sovan Lek

CESAC, UMR 5576, CNRS, University Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse cedex, France

Abstract

Two artificial neural networks (ANNs), unsupervised and supervised learning algorithms, were applied to suggest practical approaches for the analysis of ecological data. Four major aquatic insect orders (Ephemeroptera, Plecoptera, Trichoptera, and Coleoptera, i.e. EPTC), and four environmental variables (elevation, stream order, distance from the source, and water temperature) were used to implement the models. The data were collected and measured at 155 sampling sites on streams of the Adour–Garonne drainage basin (South-western France). The modelling procedure was carried out following two steps. First, a self-organizing map (SOM), an unsupervised ANN, was applied to classify sampling sites using EPTC richness. Second, a backpropagation algorithm (BP), a supervised ANN, was applied to predict EPTC richness using a set of four environmental variables. The trained SOM classified sampling sites according to a gradient of EPTC richness, and the groups obtained corresponded to geographic regions of the drainage basin and characteristics of their environmental variables. The SOM showed its convenience to analyze relationships among sampling sites, biological attributes, and environmental variables. After accounting for the relationships in data sets, the BP used to predict the EPTC richness with a set of four environmental variables showed a high accuracy ($r = 0.91$ and $r = 0.61$ for training and test data sets respectively). The prediction of EPTC richness is thus a valuable tool to assess disturbances in given areas: by knowing what the EPTC richness should be, we can determine the degree to which disturbances have altered it. The results suggested that methodologies successively using two different neural networks are helpful to understand ecological data through ordination first, and then to predict target variables.

Keywords: Self-organizing map; Backpropagation algorithm; Ordination; Classification; Prediction; Aquatic insects; Species richness

1. Introduction

Understanding communities with respect to environmental features is a fundamental basis for ecosystem management. Especially in aquatic ecosystems, the species composition of benthic communities depends on the diversity and stability of stream habitats (Cummins, 1979; Ward and Stanford, 1979) which provide the possibilities of development (Malmqvist and Otto, 1987). Therefore, benthic macroinvertebrates are widely used as indicators of short- and long-term environmental changes in running waters (Hellawell, 1978; Lenat, 1988; Smith et al., 1999; Hawkins et al.,

* Corresponding author. Tel.: +33-5-61-558687; fax: +33-5-61-556096.

E-mail addresses: park@cict.fr (Y.-S. Park), cereghin@cict.fr (R. Céréghino), compin@cict.fr (A. Compin), lek@cict.fr (S. Lek).

0304-3800/02/$ - see front matter
PII: S0304-3800(02)00258-2

2000). Species richness (i.e. the number of species occurring in a given area) is commonly used as an integrative descriptor of the community (Lenat, 1988), as it is influenced by a large number of environmental factors, such as environmental stability (Cummins, 1979; Ward and Stanford, 1979), ecosystem productivity (Lavandier and Décamps, 1984) and heterogeneity (Malmqvist and Otto, 1987), and biological factors (MacArthur, 1965; Feminella and Resh, 1990). The interactions of these factors can determine gradients in stream species richness (Vannote et al., 1980; Minshall et al., 1985). The species richness of aquatic invertebrates is also strongly influenced by natural and/or anthropogenic disturbances (Rosenberg and Resh, 1993), which may lead to spatial discontinuities of predictable gradients (Ward and Stanford, 1979, 1983) and losses of taxa (Brittain and Saltveit, 1989). Resh and Jackson (1993) observed that species richness measures were sensitive to the impact of human activities on stream ecosystems, and this was particularly true of some aquatic insects, e.g. Ephemeroptera, Plecoptera or Trichoptera (EPT), which can be considered as good biological indicators of disturbance in streams. Thus, the species richness of a restricted number of selected taxonomic groups is a good descriptor of the influence of disturbance upon the biota (Lenat, 1988).

Such ecological data are bulky, non-linear and complex, showing noise, redundancy, internal relations and outliers (Gauch, 1982; Jongman et al., 1995). There are also wide variability in variables and complex interactions between explanatory and response variables (Jongman et al., 1995). Traditionally, conventional multivariate analyzes have been applied to solve these problems (Bunn et al., 1986; Ludwig and Reynolds, 1988; Legendre and Legendre, 1998). In ecosystem management the River Invertebrate Prediction And Classification System (RIVPACS) was developed for assessing the biological quality of fresh waters. The RIVPACS and its derivatives are the primary ecological assessment analysis techniques for Great Britain (Wright et al., 1993) and Australia (Norris, 1995). They are empirical (statistical) models that predict the aquatic macroinvertebrate fauna that would be expected to occur at a site in the absence of environmental stress (Barbour et al., 1999; Coysh et al., 2000).

The models are based on a stepwise progression of multivariate and univariate analyzes. With these non-linear and complex ecological data, however, non-linear analyzing methods should be preferred (Blayo and Demartines, 1991). One of these methods is artificial neural networks (ANNs), which are versatile tools to extract information out of complex data, and which could be effectively applicable to classification and association.

In ecological modelling, ANNs have been implemented in diverse aspects (Lek and Guégan, 1999, 2000): classifying groups (Chon et al., 1996; Levine et al., 1996), patterning complex relationships (Lek et al., 1996; Tuma et al., 1996), predicting population and community development (Tan and Smeins, 1996; Recknagel et al., 1997; Chon et al., 2000), and modelling habitat suitability (Paruelo and Tomasel, 1997; Özesmi and Özesmi, 1999). Most of these studies used one of two ANNs: a self-organizing map (SOM) (Kohonen, 1982) for clustering input vectors, and a backpropagation algorithm (BP) (Rumelhart et al., 1986) for predicting biotic attributes with biotic and/or abiotic variables.

Explaining the variation of ecological data can be considered in two steps: ordination methods to summarize the variability of the data as a first step, and exploration for possible relationships between biological and environment variables as a second step (Jongman et al., 1995). From this point of view, this study focuses on practical approaches to present how two different ANNs, SOM and BP, can be applied to understand complex, non-linear ecological data.

2. Materials and methods

2.1. Artificial neural networks (ANNs)

In this study we used two different ANNs: a self-organizing map (SOM) (Kohonen, 1982, 2001), an unsupervised neural network, and a BP (Rumelhart et al., 1986), a supervised neural network. The SOM is an approximation to the probability density function of the input data, and a method

for clustering, visualization, and abstraction, the idea of which is to show the data set in another, more usable, representation form (Kohonen, 2001). The BP is a mathematical algorithm to extract relationships between explanatory and response variables, and offers an effective approach to the computation of the gradients (Kung, 1993).

We tried to pattern and to predict ecological data following two steps, by using two different ANNs (Fig. 1). First, the SOM as an ordination method was applied to summarize the variability of the data. Thus, sampling sites could be arranged on the reduced dimensions, so that these arrangements optically summarize the spatial variability of their biological and environmental features. At the second step of the analysis, the BP was used to predict the arrangements obtained with environmental variables, and to know the characteristics of their biological attributes.

2.1.1. Self-organizing map (SOM)

2.1.1.1. SOM algorithm. At first the biological data were used to train the SOM, When an input vectors **x** is sent through the network, each neuron k of the network computes the distance between the weight vector **w** and the input vector **x**, The output layer consists of D output neurons which usually are arranged into a two dimensional grid in order to better visualization. The best arrangement for the output layer is a hexagonal lattice, because it does not favor horizontal and vertical directions as much as the rectangular array (Kohonen, 2001). Among all D output neurons, the best matching unit (BMU), which has minimum distance between weight and input vectors, is the winner. For the BMU and its neighborhood neurons, the weight vectors **w** are updated by the SOM learning rule.

The training is usually done in two phases: at first a rough training for ordering with a large neighborhood radius, and then a fine tuning with a small radius. This results in training the network to classify the input vectors by the weight vectors they are closest to. The detailed algorithm of the SOM can be found in Kohonen (1989, 2001) for theoretical considerations, and Chon et al. (1996) and Giraudel et al. (2000) for ecological applications.

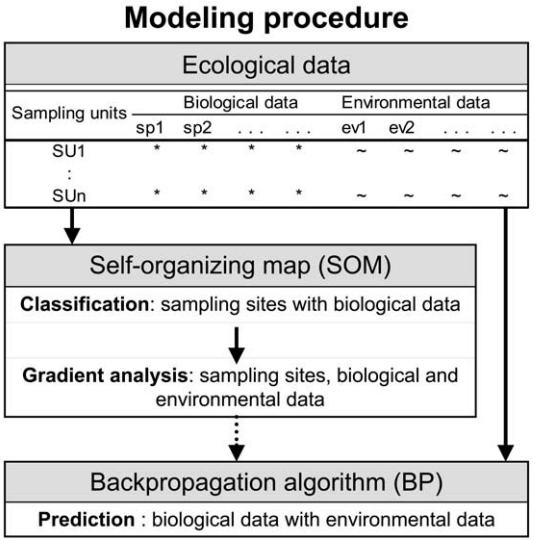

Fig. 1. Schematic diagram of the modeling procedure. SOM were applied to analyze ecological data sets in the first step, then BP was used to predict biological attributes. The solid arrow represents a direct relationship between modeling steps, while the dotted arrow displays indirect relationships.

2.1.1.2. Map quality measures. After the SOM has been trained, it is important to know whether it has been properly trained or not, because an optimal map for the given input data should exist. Although several map quality measures have been proposed (Kohonen, 2001; Zupan et al., 1993; Villman et al., 1994), the SOM quality is usually measured with two evaluation criteria: resolution and topology preservation. In this study, we first computed a quantization error (Kohonen, 2001) which is the average distance between each data vector and its BMU for measuring map resolution. Topographic error was also calculated. This error represents the proportion of all data vectors for which first and second BMUs are not adjacent for the measurement of topology preservation (Kiviluoto, 1996). Thus, this error value is used as an indicator of the accuracy of the mapping in the preserving topology (Kohonen, 2001). The topographic error ε_t can be computed in the map as follows (Kiviluoto, 1996):

$$\varepsilon_t = \frac{1}{N}\sum_{k=1}^{N} u(x_k) \quad (1)$$

where N is the number of input samples, and $u(x_k)$ is 1 if the first and second BMUs of x_k are not next to each other, otherwise $u(x_k)$ is 0.

2.1.1.3. Map size. The number of output neurons (i.e. the map size) is important to detect the deviation of the data. If the map size is too small, it might not explain some important differences that should be detected. Conversely, if the map size is too big, the differences are too small (Wilppu, 1997). Thus, we trained the network with different map sizes, and chose the optimum map size based on the minimum values for quantization and topographic errors.

2.1.1.4. Clustering SOM units. On the trained SOM map, it is difficult to distinguish subsets because there are still no boundaries between possible clusters. Therefore, it is necessary to subdivide the map into different groups according to the similarity of the weight vectors of the neurons. We used two methods to divide the trained SOM units into several subgroups. First, the unified distance matrix algorithm (U-matrix; Ultsch and Siemon, 1990; Ultsch, 1993) was applied. The U-matrix calculates distances between neighboring map units, and these distances can be visualized to represent clusters using a grey scale display on the map (Kohonen, 2001). A k-means method (Jain and Dubes, 1988) also was applied to the trained SOM map to confirm the subgroups divided by the U-matrix. To select the best patterning among partitions with different numbers of clusters, the Davies–Bouldin index (DBI; Davies and Bouldin, 1979), a relative index of cluster validity, was calculated. The smaller DBI, the better the clustering. Small values of DBI occur for a solution with low variance within clusters and high variance between clusters. Therefore, a choice is made concerning the number of clusters at which this index attains its minimum value (Hruschka and Natter, 1999).

2.1.1.5. Component planes. During the learning process, neurons that are topographically close in the array will activate each other to learn something from the same input vector. This results in a smoothing effect on the weight vectors of neurons (Kohonen, 2001). Thus, these weight vectors tend to approximate the probability density function of the input vector. Therefore, the visualization of elements of these vectors for different input variables is convenient to understand the contribution of each input variable with respect to the clusters on the trained SOM. This visualization method is related to a principal component analysis (PCA), and more directly describes the discriminatory powers of input variables in mapping (Kohonen, 2001). Therefore, to analyze the contribution of variables to cluster structures of the trained SOM, each input variable (component) calculated during the training process was visualized in each neuron on the trained SOM map in grey scale. Based on the component planes, the correlation coefficients were calculated between component pairs in both observed and calculated data.

2.1.1.6. Relationships between biological and environmental variables. It is necessary to understand the relationships between biological and environmental variables, because natural distributions of organisms are primarily determined by their environment (Huntley, 1999). To understand relationships between biological and environmental variables, we tried to introduce environmental variables into the SOM trained with biological variables. At first, we submitted each environmental variable to the trained SOM, and then we calculated the mean value (E_v) of each environmental variable in each output neuron of the trained SOM. The mean value can be computed as follows:

$$E_v = \frac{1}{n}\sum_{i=1}^{n} e_i \quad (2)$$

where n is the number of input vectors assigned (e.g. sampling sites) to each output neuron of the trained SOM, and e_i is the value of each environmental variable of input vector i. If the output

neuron was not occupied by input vectors, the value was replaced with the mean value of neighboring neurons. These mean values of environmental variables assigned on the SOM map were visualized in grey scale, and then compared with maps of sampling sites as well as biological attributes.

2.1.2. Backpropagation algorithm (BP)

In order to pattern relationships between biological and environmental variables, the BP was used as a non-linear predictor (Haykin, 1994). The BP is most popular and more used than other neural network types in various fields of investigation. This is a supervised learning algorithm and an interactive algorithm designed to minimize the mean square error between the computed output of the network and the desired output. The network normally consists of three layers: input, hidden, and output layers. It requires input vectors in the input layer, as well as target (or desired) values in the output layer corresponding to each input vector. The learning algorithm of the BP is very popular and common, and the detailed description will not be given here. A description of the learning rules can be found, for instance, in Rumelhart et al. (1986), Kung (1993), Lek and Guégan (2000).

After the learning process, a dataset not used in the training process was applied to test the reliability of the trained BP. The correlation coefficients were calculated to verify the predictability of the network in both learning and testing phases. A sensitivity analysis was carried out to evaluate the contribution of each input variable to the neural network. Sensitivity analysis is a method to study the behavior of a model, and to assess the importance of each input variable on the values of the output variable of the model (Rcotti and Zio, 1999; Salvador et al., 2001). There are many ways to perform the sensitivity analysis (Helton, 1993; Zurada et al., 1994; Hamby, 1994; Yao et al., 1998). In this study, random values ranging from ± 10 to $\pm 50\%$ were added to each input variable (Jørgensen, 1994; Opara et al., 1999; Scardi and Harding, 1999).

2.2. Ecological data

From the database of the Center for Research in Aquatic Ecosystems in Toulouse (CESAC) laboratory, 155 sampling sites were selected to implement the model. The sampling sites belonged to the Adour–Garonne stream system (116 000 km^2, South-western France), and ranged from 10 to 2500 m a.s.l., i.e. representing high Pyrenean mountains to plain areas. They were characterized using four environmental variables: elevation a.s.l. (m), stream order, distance from the source (km), and maximum water temperature (°C) in summer. These variables were chosen because they relate the location of sampling sites within the stream system without a priori consideration of any disturbance, and they are easy to collect using a geographical map and a min–max thermometer.

We focused on four insect orders, i.e. Ephemeroptera, Plecoptera, Trichoptera, and Coleoptera (EPTC), which are commonly identified at the species level in freshwater studies. The EPTC richness (i.e. the number of species occurring in a given area) was thus recorded at each sampling site. The EPTC richness was correlated to the overall macroinvertebrate richness in Adour–Garonne streams according to a highly significant linear relationship ($r = 0.91$, $P < 0.01$, see Céréghino et al., 2001), and was thus a good estimator of the overall community richness. A detailed description of these ecological data was also given in Cayrou et al. (2000), Céréghino et al. (2001).

In the modelling process with ANNs, the SOM was first applied to classify the sampling sites according to EPTC richness. The network consisted of an input layer with four input neurons, and an output layer on a two-dimensional hexagonal lattice. The data were proportionally normalized between 0 and 1 in the range of the minimum and maximum values.

To predict overall EPTC richness as output, the four environmental variables (i.e. stream order, elevation, distance from the source, and maximum water temperature) were used as input for the BP. The data sets of input and output were proportionally normalized between 0 and 1 in the range of the minimum and maximum values. One

hundred and thirty sampling sites out of 155 were used to train the network, whereas the remaining sites were used to test the feasibility of the trained BP.

3. Results and discussion

3.1. EPTC richness patterns

3.1.1. Map qualify measures and map size

Two types of quantities evaluated the quality of the trained SOM: quantization and topographic errors. The error values were computed at different map sizes (Table 1). Based on these error values the grid (map) size was selected as 140 (14×10) units. The trained SOM had a quantization error of 0.11, and a topographic error of 0.01. The result showed that only one pair of the first- and second-BMUs was not adjacent in the trained hexagonal map, so the SOM was smoothly trained in topology.

It is clear that an optimal map exists for a given input dataset. However, there are no rules to define the optimal map size. In this study we used two measures to evaluate the quality of the trained maps. Another possibility is to use different architectures like the 'growing self-organizing maps' (Fritzke, 1996). The basic idea is to estimate correct initial values for a map that has plenty of units (Kohonen, 2001).

The size of the SOM map has a strong influence on the quality of the classification. Increasing the map size brings more resolution into the mapping. The stiffness and smoothness of the map can be controlled independently of the map size by changing the final width of the neighborhood radius in the learning process. Setting the number of nodes approximately equal to the number of the input samples seems to be a useful rule-of-thumb for many applications when the data sets are relatively small (Kaski, 1997). However, attention should be paid to overfitting problems when a large map size is used. This may happen when the number of map units is as large or larger than the number of training samples. For the form of the array, the hexagonal lattice is to be preferred because it does not favor horizontal and vertical directions as much as rectangular array, and the shape of the grid (or the edges of the array) ought to be rectangular rather than square because the elastic network formed of the weight vectors must be oriented along with probability density function and be stabilized in the learning process (Kohonen, 2001).

3.1.2. Clustering sampling sites and contribution of input variables

After training the SOM with EPTC richness, the U-matrix algorithm was applied to cluster the units in the trained map. The results showed boundaries for four large clusters on the map, although that was at first not clear. The k-means also showed four clear clusters based on the minimum DBI (DBI = 0.91) (Table 2). The clusters defined by the U-matrix and k-means methods agreed with each other. Thus, sampling sites were classified into four groups (I–IV) on the trained SOM map (Fig. 2). The obtained groups corresponded to geographic regions of the drainage basin.

Fig. 3 displays component planes of each EPTC richness in each neuron on the trained map in grey scale. Dark represents high richness, while light reveals low richness. The map units in the lower areas of the SOM map showed the highest richness values for Ephemeroptera, Trichoptera and Coleoptera. The units on the lower left corner showed the highest richness for Plecoptera, whereas the units on the upper right corner corresponded to

Table 1
Map quality measures at different map sizes of the trained SOM

Map size	40	56	81	120	140	160	200
Quantization error	0.17	0.15	0.14	0.12	0.11	0.11	0.10
Topographic error	0.01	0.01	0.03	0.02	0.01	0.02	0.04

Table 2
Davies–Bouldin index (DBI) of k-means clustering at different number of clusters on the trained SOM

Number of clusters	2	3	4	5	6	7
DBI	0.97	1.10	0.91	1.31	1.10	1.14

the lowest richness values. According to these characteristics of distribution of EPTC richness on the map, sampling sites in the lower areas of the SOM map had the highest EPTC richness, whereas sampling sites in the upper areas had the lowest richness. Sites in cluster 1 had high EPTC richness,

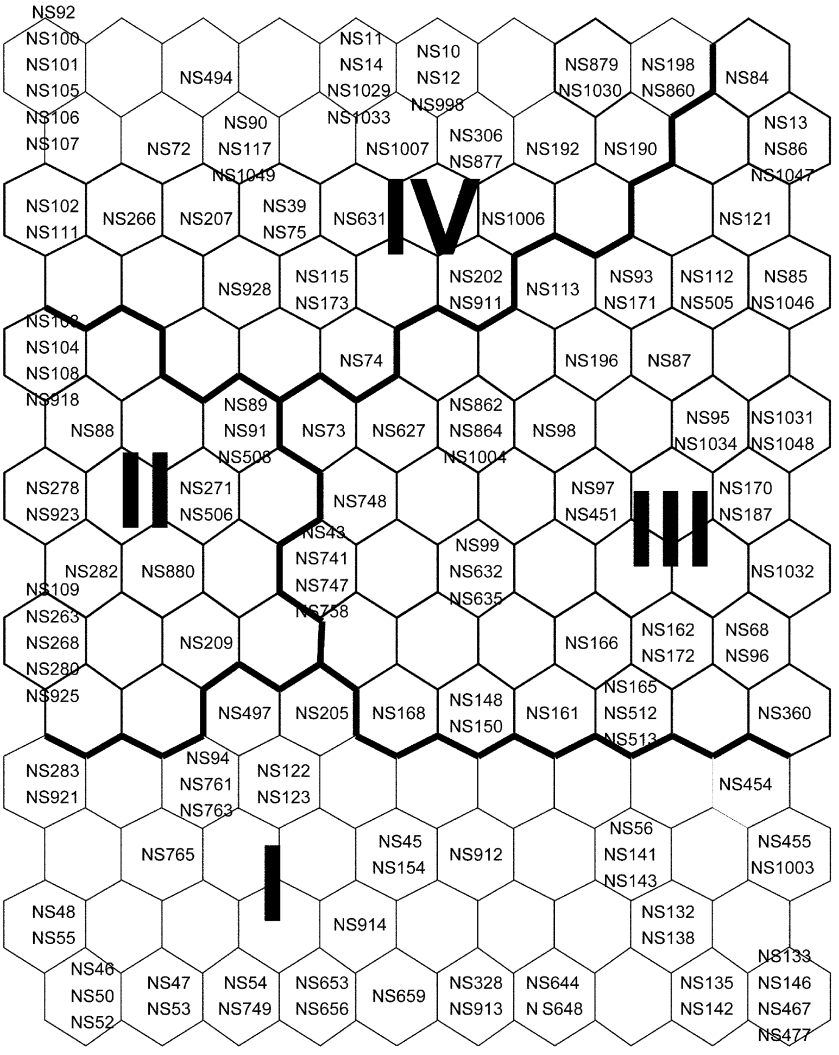

Fig. 2. Clustering of the trained SOM units. The *U*-matrix and k-means methods were applied to set boundaries on the SOM map. The Latin numbers (I–IV) in different grey scales display clusters, and the codes in each unit of the map represent the sampling sites, and can be referred to the map of Cayrou et al. (2000).

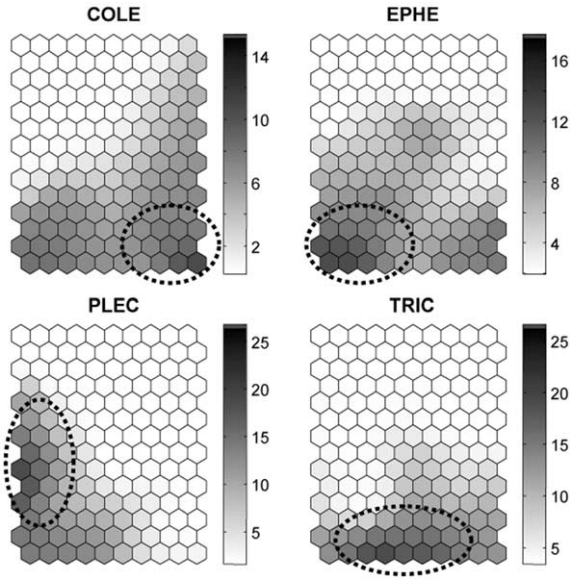

Fig. 3. Visualization of EPTC richness calculated in the trained SOM in grey scale. The values of EPTC richness were calculated during the learning process. Dark represents high value richness, while light is low value richness. The area with the highest values is circled with dotted line. COLE, Coleoptera; EPHE, Ephemeroptera, PLEC, Plecoptera; and TRIC, Trichoptera.

while sites in cluster IV had the lowest EPTC richness. Clusters II and III were separated by the species richness of Plecoptera. Sites in cluster II had high Plecoptera richness with moderate richness for other insects. Sites in cluster III had low numbers of Plecoptera species, along with moderate richness for other taxa.

Fig. 4 displays the relationships among EPTC richness variables in both observed and calculated values in the trained SOM. The scattergrams in the upper triangle of the figure show the relationships among taxonomic groups in observed data, while the plots on the lower triangle in calculated data. Species richness relationships were highly significant for both observed and estimated data among Ephemeroptera, Trichoptera and Coleoptera, but the correlation was relatively low when Plecoptera were plotted against other insect orders in both observed and calculated data sets. The correlation coefficients were relatively higher in calculated data than in observed data. The bar charts on the diagonal represent the histograms of observed and calculated values of each taxonomic group: black for the input data and grey for calculated output data. There were significantly high correlations between observed and estimated (i.e. calculated from the output neurons of SOM) EPT richness in each taxonomic group ($r > 0.65$, $P < 0.05$), while the correlation was low in Coleoptera ($r = 0.38$, $P > 0.1$).

U-matfix and *k*-means clustering methods were applied to separate subsets on the trained SOM in this study. However, other methods can be considered, because each method has advantages and disadvantages for clustering. Sometimes it is not easy to detect clear boundaries on the grey scale map of the *U*-matrix, although this method has become popular recently (Kohonen, 2001). To subdivide the trained SOM map into several groups, the fuzzy *c*-means (FCM, Bezdek, 1981; Giraudel et al., 2000) and the hierarchical agglomerative clustering and partitive clustering using *k*-means (Vesanto and Alhoniemi, 2000) have been also applied. Vesanto and Alhoniemi (2000) reported that agglomerative clustering and partitive clustering showed clear clusters, although the *U*-matrix could not find clear boundaries on the trained SOM. Thus, when we separate subgroups after training the SOM, it is necessary to compare several clustering methods if there are no clear clusters

3.1.3. *Relationships between biological and environmental variables*

The SOM has shown its high performance for visualization and abstraction for our non-linear and complex ecological data. However, it was not easy to include environmental variables in the SOM trained with biological variables. Thus, we suggest a method to introduce (or include) environmental variables into the SOM map trained with biological variables, in order to understand their effects on biological variables and on the classification of sampling sites in the trained SOM. To do this, the mean value of each environmental variable was calculated in each output neuron of the trained SOM, then each variable was visualized on the trained SOM map (Fig. 5). Dark represents high values, while light represents low values. The areas with the highest values were

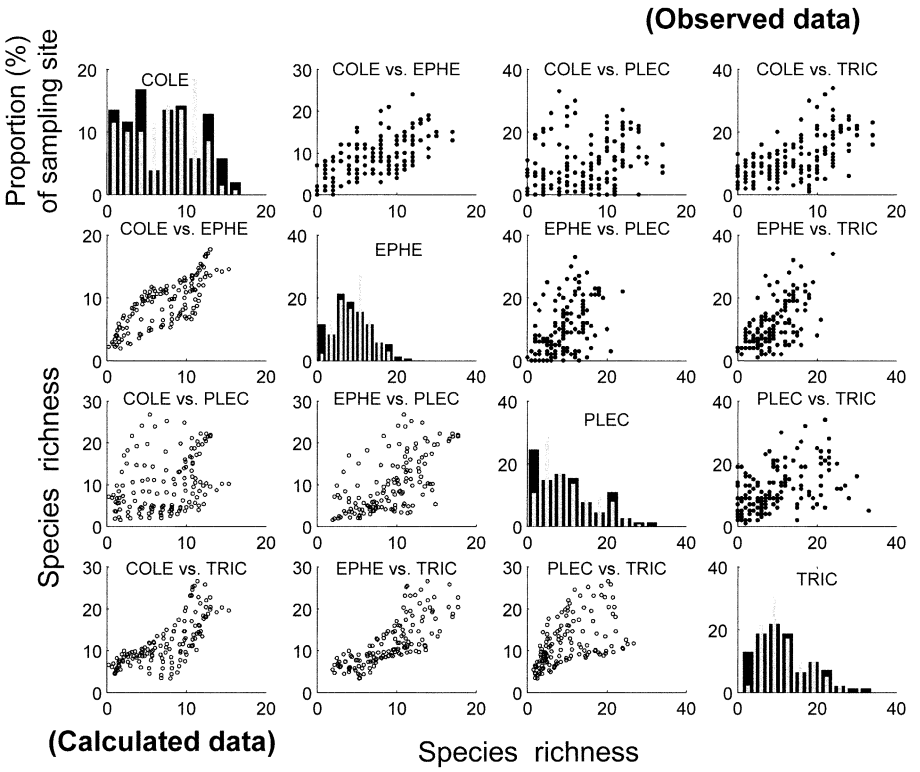

Fig. 4. Relationships among input variables in observed and calculated data in the trained SOM, The bar charts on the diagonal represent the proportion (%) of the sampling sites of observed and calculated values for each taxonomic group: black for the input and grey for calculated output. Acronyms of taxa are given in Fig. 3.

marked with a circle for each variable. Environmental variables showed gradient distributions on the SOM map. Stream order increased from the left to the right side of the map. Elevation was the highest in upper left area, and showed the clearest gradient among environmental variables. Distance from the source was lower in left areas, and higher in right areas of the SOM map. Maximum water temperature did not show a clear gradient in its distribution on the map. These results revealed that elevation was the most important factor in patterning sampling sites according to EPTC richness, while the effect of the maximum water temperature was the lowest.

At this point, we have three types of parameters (sampling sites, biological and environmental data) on the trained SOM map. Using these data, we superimposed each parameter on the same SOM map (Fig. 6). We can compare the relationships among clusters (and/or sampling sites), EPTC richness, and environmental variables. Sampling sites on the lower area of the SOM map have the highest species richness (Fig. 5). Sites in cluster I, which was characterized by high EPTC richness, belonged to the 1st–2nd order streams, sites in cluster IV having the lowest EPTC richness were located at high elevations. Sites in cluster II had high Plecoptera richness with moderate richness for other insects (Ossau valley, stream order 3–4). Sites in cluster III had low numbers of Plecoptera species with moderate richness for other taxa, and belonged to the 5th–7th order streams. Furthermore, we can see that the distribution of Plecoptera was affected by the elevation as well as stream order and the distance from the source. Coleoptera were affected by stream order and distance from the source, and Ephemeroptera and Trichoptera were also related to stream

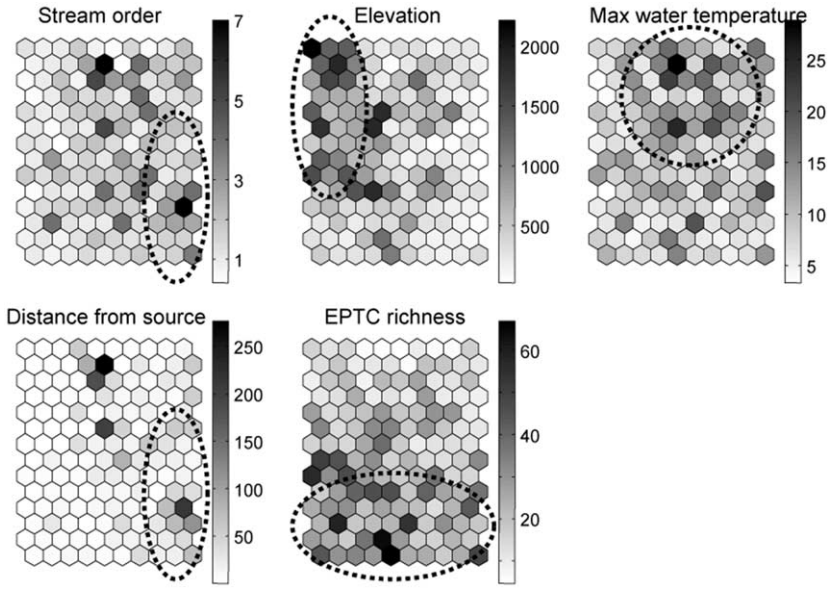

Fig. 5. Visualization of environmental variables and overall EPTC richness on the trained SOM map. The mean value of each variable was calculated in each output neuron of the trained SOM. Dark represents a high value, while light is low. The areas with the highest values are marked with a dotted circle.

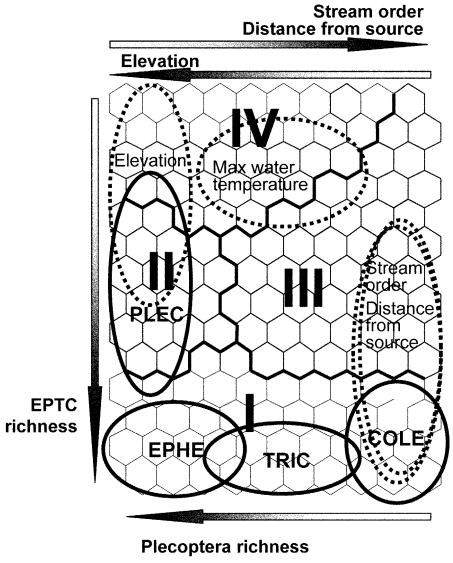

Fig. 6. Comparison of relationships among clusters (and/or sampling sites), EPTC richness, and environmental variables. Each parameter from Figs. 2, 3 and 5 is overlaid on the trained SOM map. The symbols COLE, EPHE, PLEC, and TRIC were explained in Fig. 3.

order and distance from the source. Although the SOM visualization is an indirect gradient analysis like a PCA (ter Braak, 1995), the analysing technique presented above showed the relationships between sampling sites, environmental variables, and biological variables. Thus, this approach is a much more practical tool to analyze the relationships between variables than general indirect gradient analysis.

When environmental variables are projected on the trained SOM, the distribution gradient of each variable is not clear if its variation is large, especially in relatively large maps. In this case, reducing the map size can help to find the gradient of variable distribution more clearly. On the SOM map, a clear gradient in the distribution of a variable represents a high contribution to the classification. If there is no clear gradient, the effect of the variable may be relatively low. In further studies, it is necessary to quantify both the gradients in the distribution of each variable, and relationships between biological and environmental variables.

Ordination and cluster analysis are often used in the early exploratory phase of an ecological

investigation, and the results may suggest relationships that deserve to be studied in more detail in subsequent research (Jongman et al., 1995), whereas a regression analysis could be helpful to study more specific questions in the later phases of research. This order in the analysis procedure—ordination and/or cluster analysis first, and regression analysis later—has been also used in this study.

3.2. Prediction of EPTC richness

The SOM showed how the EPTC richness was highly correlated with environmental variables. Thus, the BP was also applied to predict EPTC richness as an output variable, using the four environmental variables as input. The convergence of the learning process was generally reached after 3000 iterations with a sum of square error of 1.91.

The trained BP showed a high accuracy in predicting the overall EPTC richness on the basis of the environmental variables ($r = 0.91$, $P < 0.001$; Fig. 7a). There was, however, an underestimation of some high EPTC richness and an overestimation of around 30 EPTC richness values. The trained BP was tested with new data not used during the training phase, and the accuracy of the predictions was also very high ($r = 0.61$, $P < 0.01$; Fig. 7).

Residuals of the model have an average of 0.013 and a standard deviation of 7.83. The residuals were well distributed near the horizontal line representing the residual mean ($r = -0.01$, $P > 0.5$; Fig. 8). The histogram of residuals revealed that most values were centered near zero. To test the normality of model residuals, the statistical test of Lilliefors (1967) was applied.

The test did not reject the null hypothesis that the residuals are normally distributed ($P = 0.2$; Fig. 8b). The relationships showed no obvious sign of dependence of residuals, showing that the BP fitted the data with no bias.

A sensitivity analysis was carried out to evaluate the effect of each input variable to the network adding small random values to each input variable. To measure a response in output values, mean square errors were calculated at different levels of the perturbation of input variables. The sensitivity analysis showed that elevation and stream order provided the highest contributions among the four input variables when predicting EPTC richness, whereas maximum water temperature provided the lowest contribution (Fig. 9). This is in agreement with the results of the ordination based on the trained SOM (Figs. 3, 5 and 6).

Several methods were proposed to explain of the contribution of variables in the ANN models. These algorithms allow illustrating the role of variables in ANN models. Among these algorithms, we can classify in several categories using: (i) the connection weights (Garson 1991; Goh 1995), (ii) the connection weights and a fictitious

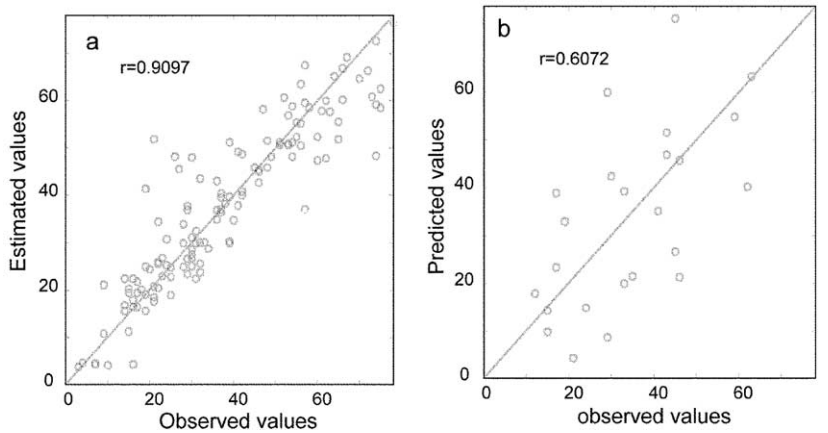

Fig. 7. Scatter plots of correlations between observed and estimated (or predicted) values by the trained BP. The diagonal lines represent perfect prediction values (predicted and observed values (a) learning, (b) testing.

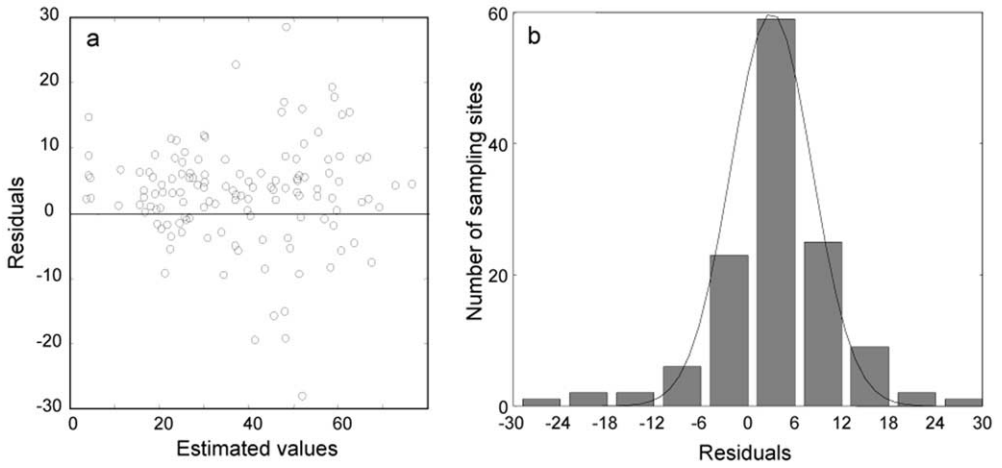

Fig. 8. Relationships between residuals and estimated values (a), and histogram of residuals (b).

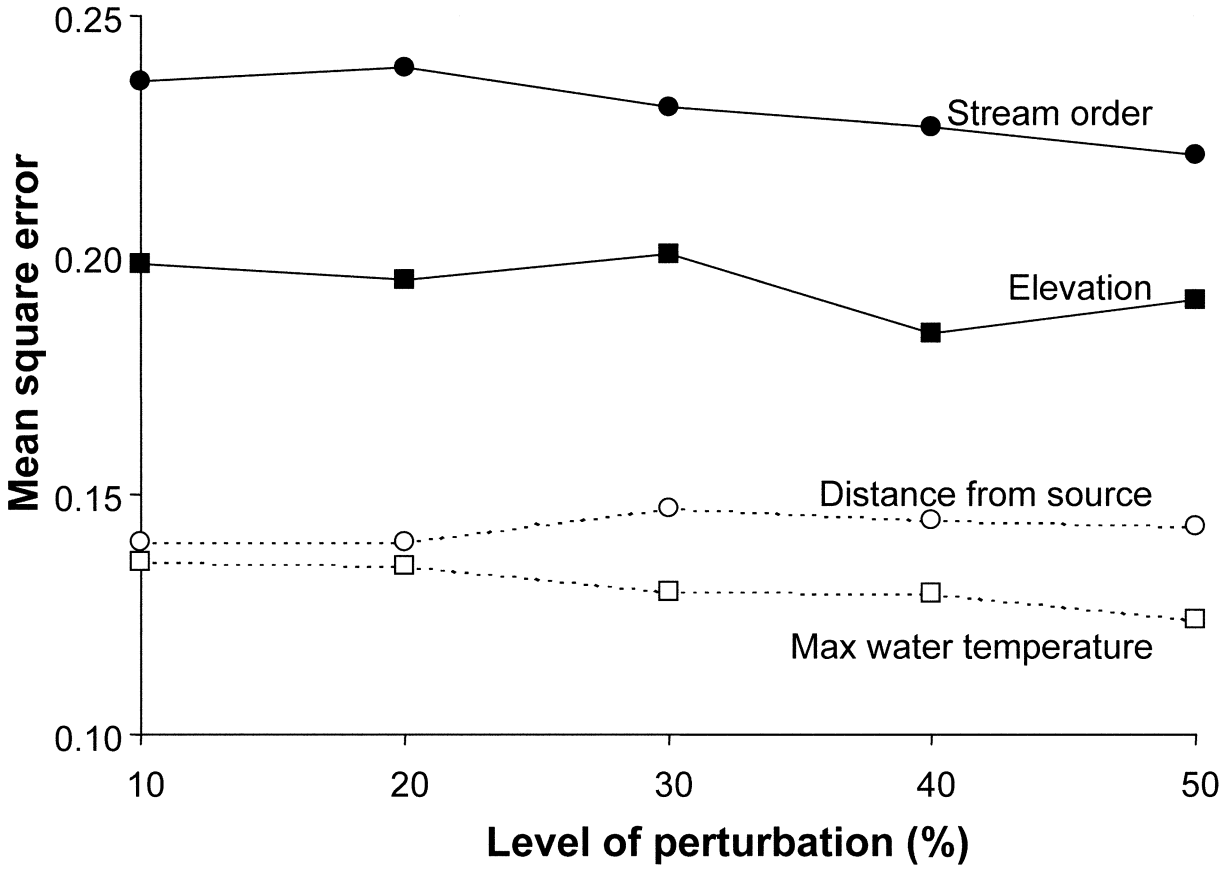

Fig. 9. Sensitivity analysis of the BP. Mean square error values were measured at different levels of pertubation of the input variables.

input matrix considering a successive variation of one input variable while the others are kept constant at a fixed value (Lek et al., 1995, 1996), (iii) the connection weights and a perturbation of the input variables (Scardi and Harding, 1999), and (iv) the partial derivatives of the output according to the input variables using the connection weights of the ANN (Dimopoulos et al., 1995, 1999; see Gevrey et al. (2002) for details).

It is recognized that BP is able to make better predictions than regression models (Lek et al., 1996; Paruelo and Tomasel, 1997). However, there are disadvantages with BP. One of them is that BP cannot explain causalities in the network because it provides a 'black-box' approach to describe the relationships between input and output variables, although a sensitivity analysis can be applied to explain the contribution of input variables to output variables. However, sometimes it is not sufficient to explain the relationships between two variable sets in terms of causality. The ordination approach proposed in this study could be helpful to explain the relationships between input and output variables.

4. Conclusions

Two different neural networks, supervised and unsupervised, have been applied to suggest practical approaches for understanding ecological data. The SOM showed a high performance for visualization and abstraction of ecological data. The trained SOM efficiently classified sampling sites according to gradients of input variables, and displayed a distribution of each component (input variable). The component planes helped to interpret the contribution of each component to the classification. Additionally, by introducing new variables (i.e. environmental variables) not used in its training phase, the SOM showed high performance in analyzing the relationships among sampling sites, biological variables and environmental variables. This method could be used as a tool to extract relationships between sampling sites, communities, and environmental variables, although the algorithm is theoretically an indirect gradient analysis. However, it remains necessary to quantify the relationships among variables.

After understanding the relationships between biological and environmental variables using the SOM, the BP, used as a nonlinear predictor, showed high accuracy in predicting EPTC richness on the basis of a set of four environmental variables. Thus, this prediction could be a valuable tool to assess disturbances in given areas: by knowing what the EPTC richness should be, we can determine the degree to which disturbances have altered it.

Finally, approaches using two different ANNs (first, understanding data sets using visualization and abstraction methods with SOM and second, prediction for target variables with BP) showed that they could take into account the variability of ecological data efficiently. Therefore, this procedure could be preferred when ecological modeling is applied to understand non-linear and complex ecological data.

Acknowledgements

We are grateful to Michele Scardi for useful comments on the manuscript. This work was supported by the EU project PAEQANN (EVK1-CT 1999-00026).

References

Barbour, M.T., Gerritsen, J., Snyder, B.D., Stribling, J.B., 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, second ed.(FPA 841-B-99-002). US Environmental Protection Agency; Office of Water, Washington, DC.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Blayo, F., Demartines, P., 1991. Data analysis: how to compare Kohonen neural networks to other techniques. In: Prieto, A. (Ed.), Proceedings of IWANN'91, International Workshop on Artificial Neural Networks. Springer, Berlin, Heidelberg, pp. 469–476.

Brittain, J.E., Saltveit, S.J., 1989. A review of the effect of river regulation on mayflies (Ephemeroptera). Regul. Rivers: Res. Manage. 3, 191–204.

Bunn, S.E., Edward, D.H., Loneragan, N.R., 1986. Spatial and temporal variation in the macroinvertebrate fauna of

streams of the northern jarrah forest, Western Australia: community structure. Freshwater Biol. 16, 67–91.

Cayrou, J., Compin, A., Giani, N., Céréghino, R., 2000. Species associations in lotic macroinvertebrates and their use for river typology: example of the Adour–Garonne drainage basin (France). Ann. Limnol. 36, 189–202.

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

Chon, T.S., Park, Y.S., Moon, K.H., Cha, E.Y., 1996. Patternizing communities by using an artificial neural network. Ecol. Model. 90, 69–78.

Chon, T.S., Park, Y.-S., Park, J.H., 2000. Determining temporal pattern of community dynamics by using unsupervised learning algorithms. Ecol. Model. 132, 151–166.

Coysh, J., Nichols, S., Ransom, G., Simpson, J., Norris, R., Barmuta, L., Chessman, B., 2000. AUSRIVAS Predictive modelling manual, http://ausrivas.canberra.edu.au/.

Cummins, K.W., 1979. The natural stream ecosystem. In: Ward, J.V., Stanford, J.A. (Eds.), The Ecology of Regulated Streams. Plenum Press, New York, pp. 7–24.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell. 1, 224–227.

Dimopoulos, Y., Bourret, P., Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Lett. 2, 1–4.

Dimopoulos, I., Chronopoulos, J., Chronopoulou Sereli, A., Lek, S., 1999. Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece). Ecol. Model. 120, 157–165.

Feminella, J.W., Resh, V.H., 1990. Hydro logic influences, disturbance, and intraspccific competition in a stream caddisfiy population. Ecology 71, 2083–2094.

Fritzke, B., 1996. Growing self-organising networks-Why. In: Verleysen, M. (Ed.), FSANN'96 European Symposium on Artificial Neural Networks. D-Facto Publishers, 1996, Brussels, pp. 61–72.

Garson, G.D., 1991. Interpreting neural network connection weights. Artif. Intell. Expert 6, 47–51.

Gauch, H.G., 1982. Multivariate Analysis in Community Ecology. Cambridge University Press, Cambridge.

Gevrey, M., Dimopoulos, I., Lek, S., 2002. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Model., in press.

Giraudel, J.L., Aurelle, D., Berrebi, P., Lek, S., 2000. Application of the self-organising mapping and fuzzy clustering to microsatellite data: how to detect genetic structure in brown trout (*Salmo trutta*) populations. In: Lek, S., Guégan, J.-P. (Eds.), Artificial Neuronal Networks: Application to Ecology and Evolution. Springer, Berlin.

Goh, A.T.C., 1995. Back-propagation neural networks for modeling complex systems. Artif. Intell. Eng. 9, 143–151.

Hamby, D.M., 1994. A review of techniques for parameter sensitivity analysis of environmental models. Environ. Modelling Assess. 32, 135–154.

Hawkins, C.P., Norris, R.H., Gerritsen, J., Hughes, R.M., Jackson, S.K., Johnson, R.K., Stevenson, R.J., 2000. Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations. J. North Am. Benthol. Soc. 19, 541–556.

Haykin, S., 1994. Neural Networks. Macmillian, New York.

Hellawell, J.M., 1978. Biological Surveillance of Rivers. Water Research Center, Stevenage Laboratory, UK.

Helton, J.C., 1993. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. Reliability Eng. Syst. Saf. 42, 327–367.

Hruschka, H., Natter, M., 1999. Comparing performance of feedforward neural nets and k-means for cluster-based market segmentation. Eur. J. Operational Res. 114, 346–353.

Huntley, B., 1999. Species distribution and environmental change. In: Maltby, E., Holdgate, M., Acreman, M., Weir, A. (Eds.), Ecosystem Management: Auestions for Science and Society. Rlyal Hollway Institute for Environmental Research, University of London, Egham, UK, pp. 115–129.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice-Hall, Englewood Hills, NJ.

Jongman, R.H.G., ter Braak, C.J.F., van Tongerenm, O.F.R. (Eds.), Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge 1995.

Jørgensen, S.E., 1994. Fundamentals of Ecological Modelling. Elsevier, Amsterdam.

Kaski, S., 1997. Data Exploration Using Self-Organizing Maps, Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82. Finnish Academy of Technology, Espoo, Finland.

Kivilnoto, K., 1996. Topology preservation in self-organizing maps. In: Proceedings of ICNN'96, IEE International Conference on Neural Networks. IEEF, Service Center, Piscataway, pp. 294–299.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59–69.

Kohonen, T., 1989. Self-organization and Associative Memory. Springer.

Kohonen, T., 2001. Self-Organizing Maps, third ed.. Springer, Berlin.

Kung, S.Y., 1993. Digital Neural Networks. Prentice Hall, Englewood Cliffs, NJ.

Lavandier, P., Decamps, H., 1984. Estaragne. In: Whitton, B.A. (Ed.), Ecology of European Rivers. Blackwell, London, pp. 237–264.

Legendre, P., Legendre, L., 1998. Numerical Ecology. Elsevier, Amsterdam.

Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120, 65–73.

Lek, S., Guégan, J.F. (Eds.), Artificial Neuronal Networks: Application to Ecology and Evolution. Springer, Berlin 2000.

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

Lenat, D.R., 1988. Water quality assessment of streams using a qualitative collection method for benthic macroinvertebrates. J. North Am. Benthol. Soc. 7 (3), 222–233.

Levine, E.R., Kimes, D.S., Sigillito, V.G., 1996. Classifying soil structure using neural networks. Ecol. Model. 92, 101–108.

Lilliefors, H.W., 1967. The Komogorov–Smirnov test for normality with mean and variance unknown. J. Am. Statist. Assoc. 62, 399–402.

Ludwig, J.A., Reynolds, J.F., 1988. Statistical Ecology: A Primer of Methods and Computing. Wiley, New York.

MacArthur, R.H., 1965. Patterns of species diversity. Biol. Rev. 40, 510–533.

Malmqvist, B., Otto, C., 1987. The influence of substrate stability on the composition of stream benthos: an experimental study. Oikos 48, 33–38.

Minshall, G.W., Petersen, R.C., Nimz, C.F., 1985. Species richness in streams of different size from the same drainage basin. Am. Nat. 125, 16–38.

Norris, R.H., 1995. Biological monitoring: the dilemma of data analysis. J. North Am. Benthol. Soc. 14, 440–450.

Opara, J., Primožič, S., Cvelbar, P., 1999. Prediction of pharmaco kinetic parameters and the assessment of their variability in bioequivalence studies by artificial neural networks. Pharm. Res. 16 (6), 944–948.

Özesmi, S., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecol. Model. 116, 15–31.

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

Rcotti, M.F., Zio, E., 1999. Neural network approach to sensitivity analysis and uncertainty analysis. Reliability Eng. Syst. Saf. 64, 59–71.

Recknagel, F., French, M., Harkonenen, P., Yabunaka, K.-L., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28.

Resh, V.H., Jackson, J.K., 1993. Rapid assessment approaches to biomonitoring using benthic macro in vertebrates. In: Rosenberg, D.M., Resh, V.H. (Eds.), Freshwater Biomonitoring and Benthic Macroinvertebrates. Chapman & Hall, London, pp. 195–223.

Rosenberg, D.M., Resh, V.H., 1993. Freshwater Biomonitoring and Benthic Macroinvertebrates. Chapman & Hall, London.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McCelland, J.L. (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. Foundations, MIT Press, Cambridge, pp. 318–362.

Salvador, R., Piñol, J., Tarantola, S., Pla, E., 2001. Golbal sensitivity analysis and scale effects of a fire propagation model used over Mediterranean shrublands. Ecol. Model. 136, 175–189.

Scardi, M., Harding, L.W., Jr, 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. Ecol. Model. 120, 213–223.

Smith, M.J., Kay, W.R., Edward, D.H.D., Papas, P.J., Richardson, K.S.J., Simpson, J.C., Pinder, A.M., Cale, D.J., Horwitz, P.H.J., Davis, J.A., Yung, F.H., Norris, R.H., Halse, S.A., 1999. AusRivAS: using macroinvertebrates to assess ecological condition of rivers in Western Australia. Freshwater Biol. 41, 269–282.

Tan, S.S., Smeins, F.E., 1996. Predicting grassland community changes with an artificial neural network model. Ecol. Model. 84, 91–97.

ter Braak, C.J.F., 1995. Ordination. In: Jongman, R.H.G., ter Braak, C.J.F., van Tongerenm, O.F.R. (Eds.), Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge, pp. 91–173.

Tuma, A., Haasis, H.-D., Rentz, O., 1996. A comparison of fuzzy expert systems, neural networks and neuro-fuzzy approaches controlling energy and material flows. Ecol. Model. 85, 93–98.

Ultsch, A., 1993. Self-organizing neural networks for visualization and classification. In: Opitz, O., Lausen, B., Klar, R. (Eds.), Information and Classification. Springer, Berlin, pp. 307–313.

Ultsch, A., Siemon, H.P., 1990. Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of INNC'90, International Neural Network Conference. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 305–308.

Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R., Cushing, C.E., 1980. The river continuum concept. Can. J. Fish. Aquat. Sci. 37, 130–137.

Vesanto, J., Alhoniemi, R., 2000. Clustering of the self-organizing map. IEEE Trans. Neural Netw. 11 (3), 586–600.

Villman, T., Der, R., Martinetz, T., 1994. New quantitative measure of topology preservation in Kohonen's feature maps. In: Proceedings of ICNN'94, IEEE International Conference on Neural Networks. IEEE Service Center, Piscataway, pp. 645–648.

Ward, J.V., Stanford, J.A., 1979. Ecological factors controlling stream zoo benthos with emphasis on thermal modification of regulated streams. In: Ward, J.V., Stanford, J.A. (Eds.), The Ecology of Regulated Streams. Plenum Press, New York, pp. 35–55.

Ward, J.V., Stanford, L.A., 1983. The intermediate disturbance hypothesis: an explanation for biotic diversity patterns in

lotic systems. In: Fontaine, T.D., Bartell, S.M. (Eds.), Dynamics of Lotic Ecosystems. Ann Arbor Sciences, Ann Arbor, Michigan, pp. 347–356.

Wilppu, R., 1997. The Visualisation Capability of Self-Organizing Maps to Detect Deviation in Distribution Control. TUCS Technical Report No. 153. Turku Centre for Computer Science, Finland.

Wright, J.F., Furse, M.T., Armitage, P.D., 1993. R1VPACS: A technique for evaluating the biological quality of rivers in the UK. Eur. Water Pollut. Con. 3 (4), 15–25.

Yao, J., Teng, N., Poh, H.-L., Tan, C.L., 1998. Forecasting and analysis of marketing data using neural networks. J. Information Sci. Eng. 14, 843–862.

Zupan, J., Li, X., Gasteiger, J., 1993. On the topology distortion in self-organizing maps. Biol. Cyerbern. 70, 189–198.

Zurada, J.M., Malinowski, A., Cloete, I., 1994. Sensitivity analysis for minimization input data dimension for feedforward neural network. In: Proceedings of IEEE International Symposium on Circuits and Systems. London, May 28–June 2, 1994, vol. 6. IEEE Press, pp. 447–450.