

Optimisation of Artificial Neural Network (ANN) model design for prediction of macroinvertebrate communities in the Zwalm river basin (Flanders, Belgium)

A. Dedecker, P. Goethals, W. Gabriels & N. De Pauw

Laboratory of Environmental Toxicology and Aquatic Ecology,

Ghent University, J. Plateaustraat 22, B-9000 Ghent, Belgium (andy.dedecker@rug.ac.be)

Abstract: To meet the requirements of the EU Water Framework Directive, models are useful to predict communities in watercourses based on the abiotic characteristics of their aquatic environment. For that purpose back-propagation artificial neural network algorithms were used to induce predictive models on a dataset of the Zwalm river basin (Flanders, Belgium). This dataset consisted of 120 samples, collected over a two year period. Fifteen environmental variables including temperature, percentage of dissolved oxygen, water depth, stream velocity, presence/absence of hollow beds, ... were measured at each site, as well as the abundance of the aquatic macroinvertebrate taxa. Different neural networks were developed and optimised to obtain the best model configuration for the prediction of the presence/absence of macroinvertebrate taxa. The best performing number of hidden layers and neurons, transfer functions in the hidden and output layer and training algorithms have been searched for. The different options were theoretically and practically validated and assessed. The theoretical validation was based on cross-validation. For the practical validation, potential applications of the neural network models were analysed, and the predictive performance of the models was assessed using ecological expert knowledge. The obtained results indicate that the number of times a taxon was found in the whole river basin influences the architecture of the network. The presence of the very rare taxon *Aplexa* and the absence of the very common taxon Tubificidae are predicted better by the Levenberg-Marquardt algorithm while Asellidae which are moderately frequent are predicted better by the gradient descent algorithm. One may also conclude that not all network models result in a relevant relation between a variable and a specific taxon. For Gammaridae for example, a rather small ANN structure gave a better idea of the impact of dissolved oxygen on its presence than a larger one. More reliable predictions and ecological interpretations for river ecosystem management would thus be possible provided the best configuration could be found.

Keywords: Macroinvertebrates; Habitat Preferences; Ecological Modelling

1. INTRODUCTION

The prediction of aquatic communities by means of ANN models has recently been discussed by several authors (Wagner *et al.*, 2000). It is stressed that the ANN architecture is a problem related one. For this reason, it is necessary to develop and optimise the ANNs to obtain the best model configuration. If predictions are made for several macroinvertebrate taxa, simultaneously another problem could emerge because the number of sites on which a taxon was found could influence the ANN architecture. The optimal ANN for a very rare taxon might be different from the one for a taxon that appears at fifty percent of the sites. This

could be an advantage for vulnerable taxa, because most of these are found at a limited number of sites. If reliable predictions would be possible for these rare taxa, conclusions regarding ANN model design for practical use in ecological river management could be drawn.

The aim of this paper was to discuss the development and optimisation of different neural networks to obtain the best model configuration for the prediction of macroinvertebrate taxa. The best performing number of hidden layers and neurons, transfer functions in the hidden and output layer and training algorithms were searched

for. The different options were validated both from a theoretical and a practical point of view. Finally, an ecological interpretation of the constructed models was made.

2. MATERIAL AND METHODS

2.1 Study sites and collected data

The Zwalm river basin which is part of the hydrographical basin of the Upper-Scheldt (Carchon & De Pauw, 1997) was selected as study area (Figure 1). The basin has a total surface of 11,650 ha, the Zwalm River itself has a length of 22 km. The river has an irregular flow regime, with low values during summer (minima lower than $0.3 \text{ m}^3\text{s}^{-1}$) and relatively high values in rainy periods mainly during winter time (maxima up to $4.7 \text{ m}^3\text{s}^{-1}$). Although Flanders is in general a rather flat region, the Zwalm river basin is characterized by a number of differences in altitude, making it a quite unique ecosystem. In the unpolluted headwaters a sensitive and vulnerable fauna is found (e.g. the bullhead (*Cottus gobio*) the brook lamprey (*Lampetra planeri*) and the mayfly Heptageniidae). Since 1999, the water quality in the Zwalm river basin has considerably improved due to investments in sewerage and wastewater treatment plants during the preceding years. Several parts of the river are however still polluted by untreated urban wastewater and by diffuse pollution originating from agricultural activities. Besides, still numerous structural and morphological disturbances exist (e.g. weirs for water quantity control, artificial embankments, etc.) (Carchon & De Pauw, 1997).

In total, sixty sites were selected in the Zwalm river basin (Figure 2) at which physical and chemical samples were taken. Besides also observations about the structural characteristics were made. Each site was examined twice over a two year period (2000-2001). In this way, 120 sets of observations were available. Certain structural characteristics (meandering, substrate type, ...) were visually monitored (Dedecker *et al.*, 2002). Flow velocity was determined by timing the transport of a float over a distance of 10 m. Field measurements were made for temperature and dissolved oxygen (OXI 330/SET), pH (Jenway 071) and conductivity (WTW LF 90). Suspended solids were measured spectrophotometrically in the laboratory (Dedecker *et al.*, 2002). Macroinvertebrates were collected by means of a standard handnet during five minute kick sampling within a river stretch of 10 m and by *in situ* exposure of artificial substrates (De Pauw *et al.*, 1994). The objective of the sampling was to collect the most representative diversity of the

macroinvertebrates at the examined site (De Pauw & Vanhooren, 1983).

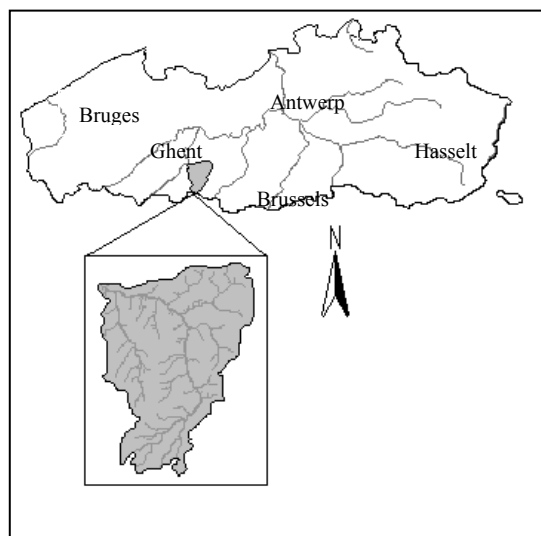


Figure 1. Location of the Zwalm river basin in Flanders, Belgium.

The structural characteristics and physico-chemical variables (Table 1) were used as inputs for the neural network models to predict the presence or absence (respectively represented by 1 and 0) of macroinvertebrate taxa in the headwaters and brooks of the Zwalm river basin.

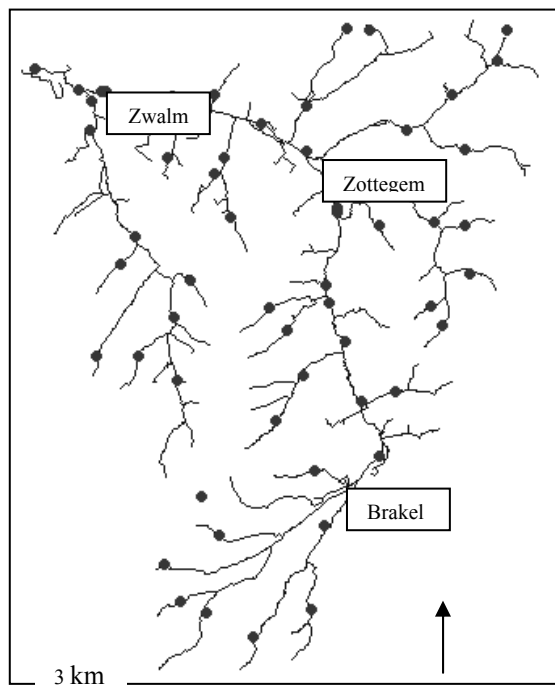


Figure 2. Selected sampling sites in the Zwalm River basin.

Table 1. Abiotic input variables and units used in the ANN model.

Variables	Units
Temperature	°C
pH	
Conductivity	µS/cm
Suspended solids	mg/l
Dissolved oxygen	mg/l
Water level	cm
Fraction of pebbles	% of river bed
Shade	%
Water plants	Present / absent
Width	cm
Flow velocity	m/s
Meandering	6 classes (1=well developed to 6 = absent)
Hollow river beds	6 classes (1 = well developed to 6 = absent)
Deep/shallow variation	6 classes (1 = well developed to 6 = absent)
Artificial embankment structures	3 classes (0 = absent ; 1 = moderate ; 2 = intensive)

2.2 Data processing

Because the input variables have very different orders of magnitude it is recommended to rescale the data. In this way, more reliable predictions can be made. The variables are rescaled to be included within the interval [-1, 1] by using the following equation:

$$V_n = 2 \times \frac{(V_0 - V_{\min})}{(V_{\max} - V_{\min})} - 1 \quad (1)$$

in which V_0 and V_n are respectively the old and new value of the variable for a sampling point, V_{\min} and V_{\max} are the minimum and maximum values of that variable in the original dataset.

2.3 Artificial Neural Networks

In this study, different neural network models were tested and optimized to obtain the best model configuration for the prediction of the presence/absence of macroinvertebrate taxa. The modelling method was based on the principles of the backpropagation algorithm (Rumelhart *et al.*, 1986). The ANN model constructs a model based on examples of data with known outputs. A backpropagation network typically comprises three types of neuron layers: an input layer, one or more

hidden layers and an output layer each including one or several neurons. As shown in Figure 3, nodes from one layer are connected to all nodes in the following layer, but no lateral connections within any layer, nor feed-back connections are possible. Fifteen input neuron are used, each representing an environmental variable. The output layer comprises one neuron, indicating the presence or absence of a macroinvertebrate taxon. With the exception of the input neurons, which only connect one input value with its associated weight values, the net input for each neuron is the sum of all input values x_n , each multiplied by its weight w_{jn} , and a bias term z_j which may be considered as the weight from a supplementary input equalling one:

$$a_j = \sum w_{ji}x_i + z_j \quad (2)$$

The output value, y_j , can be calculated by feeding the net input into the transfer function of the neuron:

$$y_j = f(a_j) \quad (3)$$

Many transfer functions can be used. In this study, two types of sigmoid functions have been compared: the tangential and logarithmic sigmoid transfer function.

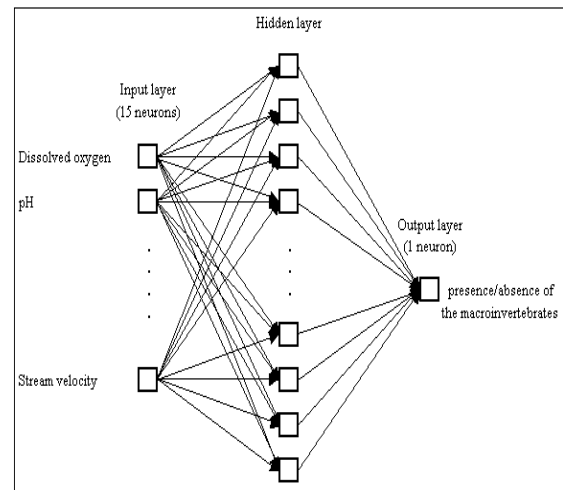


Figure 3. Illustration of a three-layered neural network with one input layer, one hidden layer and one output layer.

Before training, the values of the weights and biases are initially set to small random numbers. Subsequently, a set of input/output vector pairs is presented to the network. For each input vector, the output vector is calculated by the neural network model, and an error term is calculated for

the outputs of all hidden and output neurons, by comparing the calculated output vector and the actual output vector (the "target"). Using this error term, the weights and biases are updated in order to decrease the error, so future outputs are more likely to be correct. This procedure is repeated until the errors become small enough or a predefined maximum number of iterations is reached. This iterative process is termed "training". After the training, the ANN can be validated using independent data.

In this study, two variations of the basic backpropagation algorithm have been compared to train the models: the gradient descent algorithm and the Levenberg-Marquardt algorithm (Hagan *et al.*, 1996). The gradient descent algorithm updates the network weights and biases in the direction of the negative of the gradient. One iteration of this algorithm can be written as:

$$x_{k+1} = x_k - \alpha_k g_k \quad (4)$$

in which x_k is a vector of current weights and biases, g_k is the current gradient, and α_k is the learning rate. The Levenberg-Marquardt algorithm is similar to the quasi-Newton method in which a simplified form of the Hessian matrix (second derivatives) is used. The Hessian matrix can be approximated as

$$H = J^T J \quad (5)$$

and the gradient can be computed as

$$g = J^T e \quad (6)$$

in which J is a matrix which contains first derivatives of the network errors with respect to the weights and biases, and e is a vector of network errors. One iteration of this algorithm can be written as

$$x_{k+1} = x_k - [J^T J + \mu J]^{-1} J^T e \quad (7)$$

The model validation was based on tenfold cross-validation (Witten & Frank, 2000). For tenfold cross-validation the data is split into ten folds or partitions. Each fold in turn is used for validation while the rest is used for training. That is, use nine-tenth for training and one-tenth for validation, and repeat the procedure ten times so that in the end, every instance has been used exactly once for validation. The neural network models were implemented with the neural network extension of the software package MATLAB 5.3 for MS Windows™.

To obtain the best model configuration for the prediction of the presence/absence of several macroinvertebrate taxa, occurring in different numbers of sites, two types of sigmoid functions were compared: the tangential and logarithmic sigmoid transfer function. For each taxon two training algorithms were also compared: the gradient descent algorithm and the Levenberg-Marquardt algorithm. For both training algorithms different network architectures were analysed: five three-layered and five four-layered networks with respectively [2], [5], [10], [20], [25] and [5 5], [10 5], [10 10], [20 10] neurons in the hidden layer(s).

3. RESULTS AND DISCUSSION

3.1 Development and optimization of the ANN model configuration

The dataset consisted of 120 sets of observations. In a former study within the Zwalm river basin, a dataset of only 60 sites was used (Dedecker *et al.*, 2002). When the percentage of Correctly Classified Patterns (CCP) was compared for both studies, one could notice that the predictions are better for taxa which are moderately present. The CCP for Asellidae and Gammaridae for example was respectively 74.2 % and 68.2 % while Dedecker *et al.* (2002) found a CCP of only 60 % for both taxa. For very rare and very common taxa (e.g. *Aplexa* and Tubificidae) the differences were not significant. One can conclude that better predictions are possible when a bigger dataset is used.

Layers of neurons with non-linear transfer functions allow the network to learn non-linear and linear relationships between input and output vectors. Thus they are ideally suited for the modelling of ecological data which are known to be often non-linear. Therefore two types of sigmoid transfer functions were compared. In general, when the logarithmic sigmoid transfer function was used in the output layer, the model predicted the taxa as present. In this way, the tangential sigmoid function was selected as transfer function in the output layer, the logarithmic function was further used in the hidden layer(s).

The CCP for *Aplexa* (Mollusca) is shown in Figure 4. *Aplexa* was found at only 4.2 percent of the sites. The problem with rare taxa is that there is little information to allow the neural network model to learn when these taxa are present. In this way the models tend to "learn" that very rare taxa are always absent. The same difficulty occurs with very common taxa, for instance Tubificidae (Dedecker *et al.*, 2002). Here the models "learn"

that very common taxa are always present. In this way, the percentage of Correctly Classified Patterns was high for both training algorithms (Figure 4). For all the analysed network architectures, the gradient descent algorithm resulted in a higher percentage of Correctly Classified Patterns. As stressed by Manel *et al.* (2001) it is important to look at the predictions of the sites where the rare taxa are present. Otherwise the evaluation of these models could be misleading. When the sites were analysed where *Aplexa* was found, then the gradient descent algorithm never predicted *Aplexa* as present at these sites. On the contrary, the Levenberg-Marquardt algorithm was able to predict *Aplexa* as present at these sites. For this reason, a network with the Levenberg-Marquardt algorithm is to be preferred, although the CCP for the gradient descent algorithm was slightly better. When the architecture of the network models with the Levenberg-Marquardt algorithm however becomes more complex, the predictions for these sites were better. For the very common taxon Tubificidae, similar conclusions could be drawn. In this way, one could conclude that for the very rare taxon *Aplexa* and the very common taxon Tubificidae a more complex network model with the Levenberg-Marquardt algorithm is to be preferred. A disadvantage of the complex network models were their long calculation time.

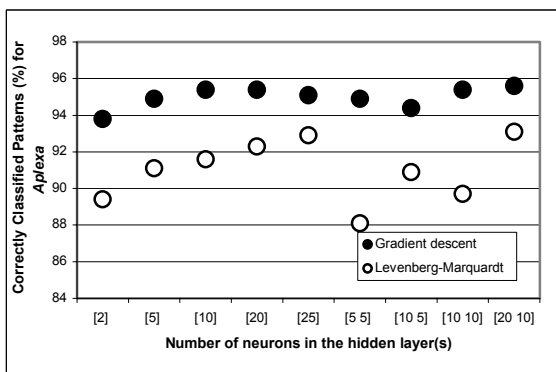


Figure 4. Comparison of the percentage Correctly Classified Patterns for *Aplexa* (Mollusca) with the gradient descent and the Levenberg-Marquardt algorithm in different ANN architectures.

Asellidae (Crustacea) on the other hand are more frequent in the Zwalm river basin than *Aplexa*. They were found in 45.4 percent of the sites. For this taxon the best prediction was made by a less complex network model. A network with the gradient descent algorithm and two hidden layers [5 5] provided the highest percentage of Correctly Classified Patterns (Figure 5).

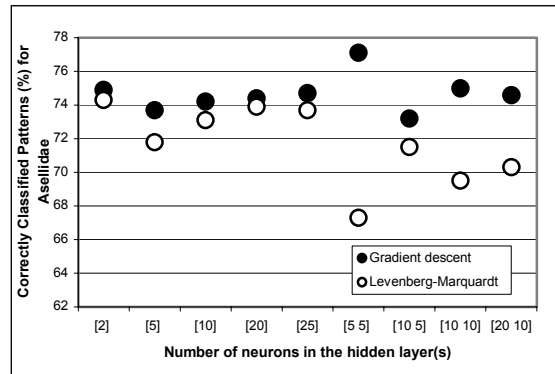


Figure 5. Comparison of the percentage Correctly Classified Patterns for Asellidae (Crustacea) with the gradient descent and the Levenberg-Marquardt algorithm in different ANN architectures.

3.2 Ecological interpretation of induced models

In ecology, it is useful to know the magnitude of impacts of a variable. Therefore, an experimental approach could be used to determine the response of the model to each of the input variables separately (Laë *et al.*, 1999; Dedecker *et al.*, 2002). A range of variation of a single independent variable to the model is applied while the others are held constant. In this way, one is able to determine the impact of the variable on the presence or absence of a specific taxon. Different network architectures with the gradient descent algorithm were applied for each variable. Figure 6 shows the effect of dissolved oxygen on Gammaridae (Crustacea).

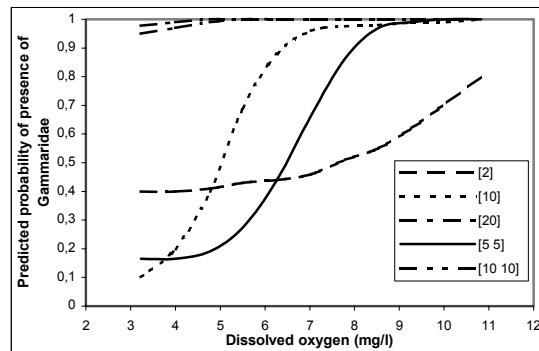


Figure 6. The impact of dissolved oxygen on the probability of presence of Gammaridae (Crustacea). The predictions have been made for different network architectures.

As mentioned in the literature, Gammaridae prefer relatively high levels of dissolved oxygen. This relationship can also be derived from the induced models (Figure 6). However, not all the networks gave this relationship. The most complex networks [20] and [10 10] predicted Gammaridae always as present, which is ecologically inappropriate. When the dataset of the Zwalm river basin was analysed

Gammaridae was present in 75 percent of the cases if the amount of dissolved oxygen was higher than 7 mg/l. Thus the best approach is given by the network with two hidden layers [5-5]. However when the best network model was used, sensitivity analyses provided useful insight in the habitat preference of that taxon, which means important information for river ecosystem management. Laë *et al.* (1999) illustrated the influence of six independent environmental variables on the fish yield in the ANN modelling. For most variables the authors found ecological relevant relations. This in contrast to this research where the relations between some variables and the presence/absence of the macroinvertebrate taxa were difficult to interpret although the predictive performance of the ANN models was in general good. Further optimisation of the ANN models can be obtained by the selection of more appropriate input variables (D'heygere *et al.*, 2002).

4. CONCLUSIONS

Artificial neural network models are efficient tools to predict the occurrence of macroinvertebrate taxa based on the abiotic characteristics of their aquatic environment. Several authors proved that ANN models are good alternatives for Multiple Regression (Lek *et al.*, 1996). As mentioned before, the network structure to be used is very problem depending. The results of this research indicate that also the frequency of occurrence of a taxon in the whole river basin influences the architecture of the network. The presence of very rare taxa (e.g. *Aplexa*) and the absence of very common taxa (e.g. Tubificidae) are better predicted with the Levenberg-Marquardt algorithm while taxa which are moderately present (e.g. Asellidae) are better predicted with the gradient descent algorithm. One might also conclude that not all network models are capable of finding a relevant relation between a variable and a specific taxon. For the Gammaridae for example, a rather small network structure gave a better idea of the impact of dissolved oxygen than a larger one. The challenge will be to build the best model configuration, if more reliable predictions are to be expected. This is essential for a correct ecological interpretation, needed for ecosystem management.

5. ACKNOWLEDGEMENTS

The first author is a recipient of a grant of the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

6. REFERENCES

Carchon, P., and N. De Pauw, Development of a methodology for the assessment of surface

waters. Study by order of the Flemish Environment Agency (VMM), Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent, Belgium, 55p., 1997 (in Dutch).

Dedecker, A.P., P.L.M. Goethals, and N. De Pauw, Comparison of artificial neural network (ANN) model development methods for prediction of macroinvertebrate communities in the Zwalm river basin in Flanders, Belgium., *TheScientificWorldJOURNAL*, 2, 96-104, 2002.

D'heygere, T., P. Goethals, and N. De Pauw, Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates, *Ecological Modelling*, 2002 (submitted).

De Pauw, N., V. Lambert, A. Van Kenhove, and A. Bij De Vaate, Performance of two artificial substrate samplers for macroinvertebrates in biological monitoring of large and deep rivers and canals in Belgium and the Netherlands, *Environmental Monitoring and Assessment*, 30, 25-47, 1994.

De Pauw, N., and G. Vanhooren, Method for biological assessment of watercourses in Belgium. *Hydrobiologia*, 100, 153-168, 1983.

Hagan, M.T., H.B. Demuth, and M. Beale, Neural network design, PWS Publishing Company, 712 p., Boston, 1996.

Laë, R., S. Lek, J. Moreau, Predicting fish yield of African lakes using neural networks, *Ecological Modelling*, 120, 325-335, 1999.

Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, and S. Aulagnier, Application of neural networks to modelling nonlinear relationships in ecology, *Ecological Modelling*, 90, 39-52, 1996.

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

Rumelhart, D.E., G.E. Hinton, and R.J. Williams, Learning representations by back-propagation errors, *Nature*, 323, 533-536, 1986.

Wagner, R., T. Dapper, and H.H. Schmidt, The influence of environmental variables on the abundance of aquatic insects: a comparison of ordination and artificial neural networks, *Hydrobiologia*, 422-423, 143-152, 2000.

Witten, I.H., and E. Frank, Data mining: practical machine learning tools and techniques with Java implementations, Morgan Kaufmann Publishers, 369 p., San Francisco, 2000.