Benthic Macroinvertebrates Modelling Using Artificial Neural Networks (ANN): Case Study of a Subtropical Brazilian River

D. Pereira[a,c], **M. de A. Vitola** [a], **O. C. Pedrollo** [a], **I. C. Junqueira**[b] **and S. J. De Luca**[a]

[a] *Hydraulic Research Institut, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul State, Brazil.*

[b] *Environment Municipal Department of Porto Alegre.*

[c] *Feevale University Center.*

Abstract: Back-propagation Artificial Neural Networks (ANN) were tested with the aim of modelling the occurrence of benthic macroinvertebrate families in a south Brazilian river. The dataset, consisting of 67 sets of observations of macroinvertebrate abundance (families Hydrobiidae, Tubificidae, Chironomidae, Baetidae and Leptophlebiidae) and water quality variables (pH, temperature, dissolved oxygen, biochemical oxygen demand, nitrate, phosphate, total solids, turbidity and fecal coliforms), was collected at eleven sampling sites in the Sinos River Basin during 1991-1993. Five different ANN architectures, with one hidden layer and 2, 5, 10, 20 and 25 neurons were tested. The ANN models were trained using the gradient descendent and Levenberg-Marquardt (LM) algorithms, with different combinations of sigmoid transfer functions (log-log, tan-log, tan-tan). The percentage of success and the correlation coefficient were used to choose the best network architecture for each taxon. The networks with the LM algorithm provided the best predictions of macroinvertebrate family occurrence, independent of the family's frequency. The same network architecture did not always reproduce all the relationships between the taxon occurrence and the environmental variables. The best model, based on a high correlation coefficient among real and predicted data and a high percentage of successes, was the ANN for a very common taxon (Hydrobiidae).

Keywords: Macroinvertebrates; Artificial Neural Networks; Modelling; Water Quality

1. INTRODUCTION

Artificial Neural Network (ANN) models have been widely used as a tool for modelling biological communities in many European countries (Chon, 2000, 2001; Dedecker et al., 2002; Park et al., 2003). In Brazil, this ecological modelling approach is less developed due the scarcity of datasets of the structure and function of biological communities.

The knowledge of taxonomy, structure and organization of benthic communities in the south of Brazil is incipient. Datasets of benthic macroinvertebrate occurrence in rivers of Rio Grande do Sul State (Brazil) were published by Junqueira (1995), Pereira, (2002) and Pereira & De Luca (2003).

This paper, which uses a back-propagation algorithmic approach to modelling the family occurrence of benthic macro invertebrates in a south Brazilian river based on a dataset obtained by Junqueira (1995), is, in a Brazilian context, a pioneering work. The aims of this paper were test the ability of ANN models to model the presence/absence of macro invertebrates and contribute to the ecological management of Sinos River Basin.

2. MATERIAL AND METHODS

2.1 Study sites and collected data

The Sinos River headwaters is situated at an altitude of 700 m above the sea level, in Serra do Mar. The river, from its headwaters to the confluence with the river Jacuí is 185 km long, with 65 tributary. The total area of Sinos River Basin is 4,328 km^2. In this basin there are 28 municipalities with a total population of 1,185,961 inhabitants. The main forms of water use are for residential consumption, for industrial processes

and for agricultural irrigation. The river also serves as the receiving body for the effluents generated in the cities.

The data used in this study were collected by Junqueira (1995) for the purpose of water quality monitoring at 12 sampling sites in the Sinos River Basin (period 1991-1993). The dataset consists of 64 sets of observations of benthic macroinvertebrate (Hydrobiidae, Tubificidae, Chironomidae, Baetidae and Leptophlebiidae) abundance and water quality measurements (temperature, dissolved oxygen, biochemical oxygen demand, pH, nitrate, phosphate, turbidity, total solids and fecal coliforms).

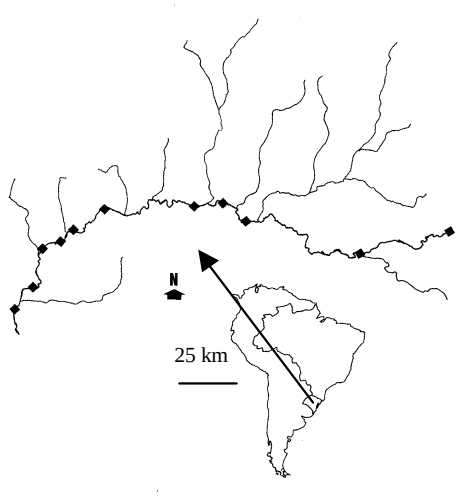

Figure 1. Location of monitoring sampling sites in the Sinos River Basin, Rio Grande do Sul State, South Brazil.

2.2. Data processing

All input variables were analyzed for consistency in order to eliminate any absurd values. The fecal coliforms data were subjected to a logarithmic transformation and the resulting output data were converted to represent either presence or absence (represented by 1 and 0 respectively). After this processing, all input and output variables were rescaled to fall within the interval between –1 and 1 – although neural networks can deal with input of different orders of magnitude, this rescaling leads to more reliable predictions. All variables were rescaled to be included within the interval between -1 and 1; the MATLAB function was used for this.

2.3. Artificial Neural Networks

In this study, different neural network architectures were used based on the supervised training method, in which the value of the target variables are know. This training method was based on the principles of the backpropagation algorithm, which is the generalization of the learning rule of Widrow-Hoff for networks with multiple layers and nonlinear differentiable transfer functions (Rumelhart *et al.*, 1986; Mathworks, 1998).

The backpropagation network normally has three layers: an input layer, one or more hidden layers, and an output layer, each including one or more neurons. The figure 2 shows the topology of the networks used in this study. Each node from one layer is connected to all nodes in the following layer, but neither lateral connections within any layer, nor feed-back connections are used. Nine input neurons were used, each one represent an environmental variable, while the output layer consists of only one neuron indicating the presence or absence of a macroinvertebrate taxon.

The weights and biases should be initialized before training, normally these values are random numbers, the result network should be independent of the initialization, this indicate the ability generalization of the network. The network architectures were train using tow different set of initial values for weights and biases.

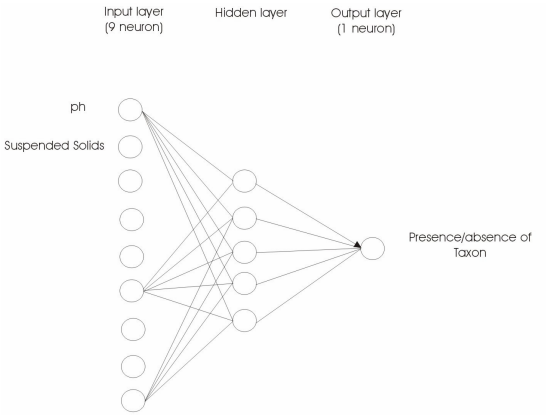

Figure 2. Neural Network topology use in this study.

Other important parameter to be set before training is the transfer function, the neural network can use many different functions; they should be differentiable only. Two type of sigmoid function were used in this study: the tangential and the logarithmic sigmoid transfer function.

The training of an ANN consists in adjusting the weight and the biases using an algorithm of back propagation errors. For each input vector, the output vector is calculated by the neural network,

the error being calculated for the outputs by comparing the output vector with the "target". Using this error, the weights and biases are updated in order to minimize the error. This procedure is repeated until the errors become small enough or a predefined maximum number of interactions is reached.

Many algorithms can be used for the training of the network; most of these are based on optimization techniques. Two different algorithms were used in this study: the gradient descendent with variable learning rate algorithm, which attempts to keep the learning step size as large as possible while keeping learning stable, and the Levenberg-Marquardt algorithm, this algorithm appears to be the fastest method for training moderate-sized feedforward neural networks.

The validation model was based on the methodology described by Dececker *et al.* (2002), which consists in splitting the data set into tenfolds. Each tenfold is used for validation while the others are used for training. This procedure is repeated until all tenfolds are used for validation.

The neural network models were implemented using the toolbox of MATLAB 5.3 for MS Windows™. Three combinations of transfer functions were tested: tangential and logarithmic sigmoid transfer functions. For each taxon, two training algorithms were used: the descendent gradient algorithm with variable learning rate and the Levenberg-Marquardt algorithm. For both training algorithms, two different initializations were tested with five different network architectures with one hidden layer of [2], [5], [10], [20] e [25] neurons.

3. RESULTS AND DISCUSSION

During the training of the ANNs, the gradient descendent algorithm didn't present satisfactory results. Satisfactory results were obtained when the algorithm of Levenberg-Marquardt was used. This algorithm resulted in high percentage of success (86.6-98.5% of the data) and correlation coefficients (r=0.692 and 0.968) among real and predicted data. The acting of the ANNs architectures in the modelling of the simulated families is described below.

3.1 Hydrobiidae

This is a very common taxon, occurring in 62% of the observations. For this taxon the ANNs presented satisfactory results as for the percentage of success and the correlation coefficient among the real and predicted data. These networks were not sensitive to the initialization. The best results (Fig. 3) were obtained with a network of five neurons and logsig-tangsig function of activation with the largest correlation coefficient (r=0.968) and percentage of success (98.51%).

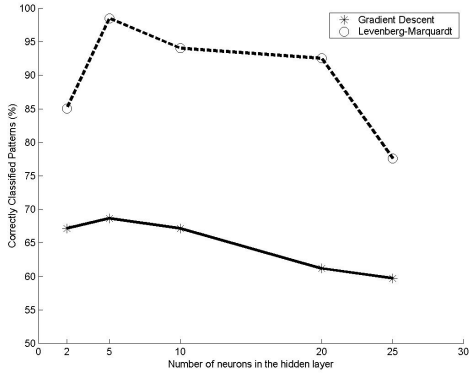

Figure 3. Comparation of the percentage correctly classified patterns for Hidrobiidae with the modified gradient descendent and the Levemberg-Marquardt algorithm with logsig-tansig activation function.

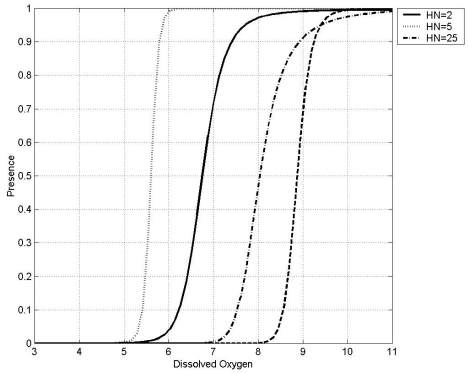

Figure 4. Probability of the presence of Hidrobiidae as a function of dissolved oxygen, by means of sensitivity analysis.

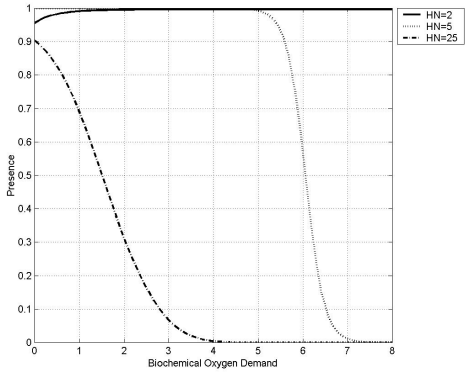

Figure 5. Probability of the presence of Hidrobiidae as a function of dissolved oxygen , by means of sensitivity analysis

The relationship between the occurrence of Hydrobiidae and the nitrate, dissolved oxygen (Fig. 4), DBO_5 (Fig. 5) concentrations, pH and temperature, verified through the sensitivity analysis, demonstrated that this network was effective in the modelling of this family. As mentioned in the literature for Brazilian creeks (Pereira & De Luca, 2003) Hydrobiidae prefer relatively high levels of dissolved oxygen.

3.2 Tubificidae

This is a moderate taxon, occurring in 32% of the observations. For this taxon the ANNs was sensitive to the initialization. A network with two neurons and activation function tangsig-tangsig presented the largest correlation coefficient (r=0.825) and percentage of success (92.54%). However, this network didn't reproduce a real relationship between the tubificidae occurrence and the environmental variables. Just a network with five neurons and activation function logsig-tangsig, with a lower correlation coeficient r=0.692 and 86.57% of success, has been efficient in modelling this family (Fig. 6). This network reproduces a real relationship among the tubificidae occurrence and the nitrate, DBO_5 and dissolved oxygen concentrations and temperature, as demonstrated in the sensitivity analysis (Fig. 7 and 8). The result obtained with a network of 20 neurons indicates unreal relationship among dissolved oxygen, DBO_5 and this taxon. It is known that Tubificidae occur in water with high level of organic material and low level of dissolved oxygen.

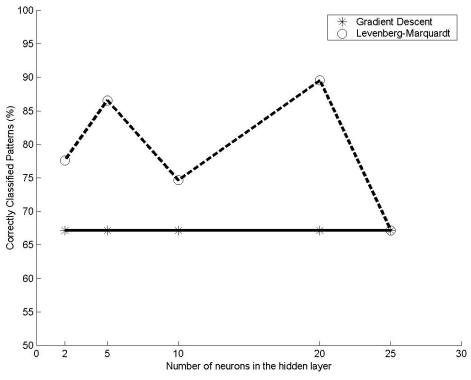

Figure 6. Comparation of the percentage correctly classified patterns for Tubificidae with the modified gradient descent and the Levemberg-Marquardt algorithm with logsig-tansig activation function.

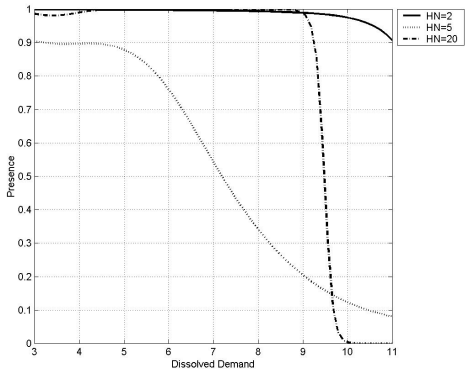

Figure 7. Probability of the presence of Turbificidae as a function of dissolved oxygen, by means of sensitivity analysis.

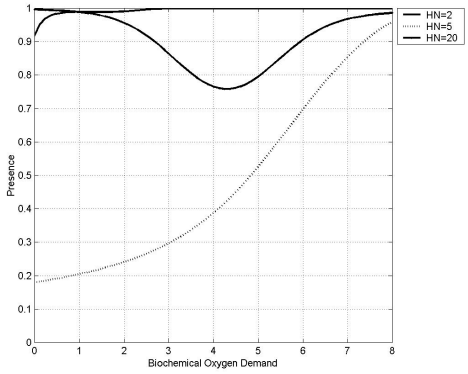

Figure 8. Probability of the presence of idae Tubificidae as a function of dissolved oxygen, by means of sensitivity analysis.

3.3 Chironomidae

This is a very common taxon, occurring in 65% of the observation. For this taxon the ANN was insensitive to the initialization. A network with ten neurons and activation function tangsig-tangsig presented the largest correlation coefficient (r=0,968) and percentage of success (98.51%). The relationship among the occurrence of Chironomidae and the concentrations of DBO_5, fecal coliforms and total solids, was verified through the sensitivity analysis, indicating that this network was effective in modelling of this family. It is known that Chironomidae are resistant to organic pollution occurring in water with high level of organic material and low level of dissolved oxygen.

3.4 Leptophlebiidae

This is a rare taxon, occurring in 25% of the observation. The ANN tested during the modelling of this family presented satisfactory results when

the percentage of success and the correlation coefficient among the real and predicted data was high. The acting of these networks was independent of the initialization, although it had a different behavior in the sensitivity analysis. A network with 20 neurons and activation function logsig-tangsig presented correlation coefficient (r=0.834) and percentage of success (94.03%). The relation among the occurrence of Leptophlebiidae and the concentrations of DBO_5, pH and turbidity, was verified by means of sensitivity analysis. The same network presented the largest correlation coefficient (r=0.876) and number of success (95.52%), under new initialization. This network reproduced the relationship among the occurrence of Leptophlebiidae and total solids.

3.5 Baetidae

This is a very common taxon, occurring in 79% of the observation. For this taxon, the ANN was sensitive to initialization. A network with 25 neurons and activation function tangsig-tangsig presented the largest correlation coefficient (r=0.823) and percentage of success (94.03%). This network reproduced only the relationship between pH and this taxon. A network with 20 neurons and a function of activation logsig-tangsig, with a coefficient of correlation r=0.769 and 91.04% of success, recognized the relationship between the occurrence of Baetidae and the DBO_5, dissolved oxygen concentrations, fecal coliforms and temperature, as demonstrated by the sensitivity analysis. This family and Leptophlebiidae have been shown to be sensitive to organic pollution in Brazilian creeks (Pereira, 2002; Pereira & De Luca, 2003), giving preference to waters with high level of dissolved oxygen.

4. CONCLUSION

In this paper the modified descendent gradient algorithm was not appropriate for training or predicting the presence or absence. The algorithm of Levemberg-Marquardt was the most effective in the training and predicting the occurrence of macroinvertebrate families, independent of the occurrence frequency. According to Dedecker *et al.* (2002) the very rare and very common taxon where better modelled by LM and moderate taxon with GDA.

The best model family was a very common taxon (Hydrobiidae), presenting high correlation coefficient among real and predicted data as well as percentage of success, reproducing real relationship with environmental variables.

The ANN that presented larger correlation coefficient among real and predicted data and high percentage of success was not always the best one to reproduce the relationships between the taxons occurrence and the environmental variables.

Not always the same network architecture reproduces all the relationships between the taxon occurrence and the environmental variables. Similar results were obtained by Dedecker *et al.* (2002). This could be related to with the size of dataset, which influences the generalisation ability of the ANN.

The result could be improved with selection of more appropriate environmental variables, as habitat preferences and hydrodynamics characteristics.

The results presented indicate that ANN could be an appropriate tool to predict the occurrence of macroinvertebrate families based on environmental variables. The principal drawback is to build the best model configuration, which is essential for a correct ecological application of ANNs for ecosystem management.

ACKNOWLEDGEMENTS

The first two authors wish to thanks the National Council of Scientific and Technology Development (CNPq) for the doctoral degree grants.

REFERENCES

Chon, T.-S.; Park,Y.-S. & Park, J.H. 2000. Determining temporal pattern of community dynamics by using unsupervised learning algorithms. *Ecological Modelling* 132:151–166.

Chon, T.S.; Kwak, I.S.; Park,Y.S.; Kim, T.H. And Kim, Y. 2001. Pattern and short-term predictions of macroinvertebrate community dynamics by using a recurrent artificial neural network. *Ecological Modelling* 146:181-193

Dedecker, A.; Goethals, P.; Gabriels, W. & De Pauw, N. 2002. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrate communities in the Zwalm river basin (Flanders, Belgium). In: Proceedings of international Environmental Modelling and Software Society. vol. 2, pag. 142-147, Lugano, Switzerland, junho de 2002.

Junqueira, I. C. 1995. Aplicação de índices biológicos para a interpretação da qualidade da água do rio dos Sinos. Porto Alegre, PUCRS, 115p. (Dissertação).

MATHWORKS. *Neural Networks Toolbox User´s Guide*. The MathWorks Inc., Natick, MA, 1998.

Park,Y.S.; P.F.M. Verdonschot; T.S. Chon and S. Lek., Patterning and predicting aquatic macroinvertebrate diversities using articial neural network, *Water Research 37*:1749-1758, 2003.

Pereira, D., Aplicação de índices ambientais para a avaliação da sub-bacia do arroio Maratá, bacia do rio Caí (RS, Brasil), Master Degree Dissertation, Bioscince Institut of Federal University of Rio Grande do Sul, Porto Alegre, 2002.

Pereira, D. & S. J. De Luca, Benthic macroinvertebrates and the quality of the hydric resources in Maratá Creek basin of (Rio Grande do Sul State, Brazil), *Acta Limnologica Brasiliense 15*(2): 57-68, 2003.