Transactions on Ecology and the Environment vol 2, © 1993 WIT Press, www.witpress.com, ISSN 1743-3541

A neural network predictor of benthic community structure in the Canadian waters of the Laurentian Great Lakes

B.M Ruck,[a] W.J. Walley,[a] T.B. Reynoldson,[b] K.E. Day[b]

[a] *Department of Civil Engineering, Aston University, UK*

[b] *National Water Research Institute, Burlington, Ontario, Canada*

ABSTRACT

A method of predicting benthic community structure from environmental variables using artificial neural networks is described. The input variables represent geophysical, limnological and sedimentological characteristics of sites in Canadian waters of the Laurentian Great Lakes. A single output from the network predicts the number of individuals of a given taxon to be found in a 5.5cm by 10cm deep core sample of lake sediment taken at the site in question. Networks have been trained for four taxa; namely Oligochaeta, Porifera, Chironomidae and Pelecypoda. Three input vector sets were compared: the 28 dimension raw data set, a subset of 9 variables and a 7 dimension eigenvector set. Performance tests were carried out using a 1-fold cross validation technique, which maximises data utility whilst maintaining independence between the training and test sets. Details of the training and testing of the networks are given, together with a brief introduction to neural networks. It is concluded that artificial neural networks have potential for use in biological monitoring systems.

INTRODUCTION

The contamination of sediments occurs in freshwater and marine systems throughout the world. While contaminants such as nutrients, metals and oxygen demanding substances occur naturally, their presence at high concentrations is usually a result of human activity. Chemical methods of classifying the level of contamination of sediments do not take into account the biological stress caused by the contaminants or whether this stress will continue after the primary sources of pollution have been controlled. Concern for the degree of environmental protection afforded by chemical guidelines, together with the lack of uniform international criteria and a failure to introduce plans to remediate degraded areas in the Laurentian Great Lakes, (International Joint Commission [3, 4]), has prompted scientists at the National Water Research Institute to investigate more comprehensive methods of sediment assessment and evaluation criteria. In 1990, they proposed the development of biological sediment guidelines (Reynoldson and Day [6], Reynoldson and Zarull [7]), using two approaches: sediment toxicity and benthic invertebrate community structure.

The overall objectives of the study are to:

a) develop a classification system for unpolluted nearshore sites based on the benthic invertebrate community structure and selected bioassay endpoints;
b) determine the degree to which the site classification can be predicted from physio-chemical variables;
c) establish the relationship between community structure and bioassay assessments;
d) develop procedures for the prediction of key elements of the fauna expected at a site from its environmental features not affected by human activity;
e) select key species and toxicity tests that show the most robust predictive response for the purpose of developing guidelines;
f) establish the sensitivity of selected guidelines at a range of impacted sites.

The fundamental principle underlying the development of sediment guidelines is that it is possible to predict the community structure of species at an unpolluted site from a few physio-chemical variables (Wright, *et al.* [10], Johnson [5]). Comparison of the observed community with the predicted community determines whether or not the site-specific guidelines have been met. This paper describes a novel approach to the prediction process, using neural networks as an alternative to the statistical techniques used by Wright *et al.* [10] in a similar study. The two are not directly comparable, however, because Wright *et al.* [10] were predicting the presence or absence of a wide range of taxa, whereas this study attempts to predict the number of individuals present in a standard sample for a limited number of commonly occurring taxa. Later in this study, a direct comparison will be made between the performance of the neural networks and that of statistical techniques such as canonical correspondence analysis (CCA) and multiple discriminant analysis (MDA).

THE FIELD STUDY

The reference sites being used in the study have been selected on the premise that the guidelines are required for areas where remediation is practical and contamination is likely to be occurring. The reference sites sampled are less than 30m deep, less than 3km from the shore, away from outfalls and anthropogenic development. The substratum is largely composed of fine grained sediments and the sites represent, as far as possible, pre-contamination conditions. A total of 250 "clean" reference sites, sampled over 3 years, will be used in the final study, but this pilot study is based upon the first 50 only.

In order to minimise sampling error a standard methodology has been adopted. Samples are taken with a light weight box core, from which five smaller cores (10cm by 5.5cm) are sub-sampled. A sixth core is taken and used for the physio-chemical analyses. Three categories of environmental matrices were measured; geophysical, limnological and sedimentological variables, details of which are given in Table 1.

Table 1. Variables used for field measurements

GEOGRAPHICAL	LIMNOLOGICAL	SEDIMENTOLOGICAL
Distance from shore	Degree days	Water content
Latitude	Thermocline depth	Particle size
Slope	In bottom 0.5m	Loss on ignition
Depth	Alkalinity	TP, TOC, TON
Shoreline development	Oxygen	AVS
	Nutrients	Metals, major ions in porewater
	pH	
	Temperature, surface and bottom	metals, major ions, nutrients

While identification of benthic taxa is taken to species level where possible, as this leads to greater discrimination and no information loss, for the purpose of this paper family level characteristics were adopted.

NEURAL NETWORKS

The original, biologically inspired, research into neural networks was founded in the 1940's, and was actively pursued until the end of the 1960's when interest and research funding waned. In the late 1980's there was a resurgence of interest due to the development of powerful new learning algorithms, high power computers and some excellent performance claims. Today, the term "neural network" is somewhat misleading as most of the algorithms are biologically implausible. In essence the term is now used to describe a wide assortment of algorithms that are based upon computationally simple units with a high degree of interconnectivity. They are flexible non-linear models that can be "trained" to perform a specific computation. This takes the form of an input-output mapping, and the training is achieved by the adjustment of the network's parameters (or weights).

Networks may be trained in several ways, including supervised, unsupervised and reinforcement learning, with the most common method used for prediction problems being supervised learning. This requires that pairs of input-output patterns be presented to the network, so that the error between the actual and desired outputs for each training input can be established. Training is achieved by a systematic reduction of these errors. The network used in this pilot study was a multi-layer perceptron, a supervised learning model that reduces its errors using the standard back-propagation algorithm, Rumelhart *et al.* [9].

The first perceptron networks had just two layers, input and output, and used a simple learning technique called the delta rule. Unfortunately, they had limited computational capabilities that restricted their use to linear separable problems. In an attempt to overcome this, hidden (i.e. intermediate) layers of units were added, but this created a learning difficulty known as the credit assignment problem. This is the problem of determining which of the weights should be adjusted when an incorrect output occurs. A solution to this was the back-propagation algorithm, Rumelhart *et al.* [9].

Figure 1 shows a typical topology used in this study. There are *n* units in the input layer, *m* in the hidden layer and just one in the output layer, representing the abundance of a taxon.

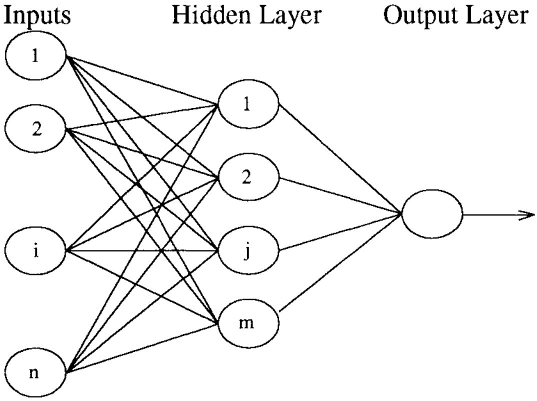

Figure 1. Architecture of a typical multi-layer perceptron.

After initialising the weights the first input-output data pair is presented to the network. Signals from the input layer are fed forward to the units in the hidden layer. These units calculate a weighted sum of their inputs and transform them, by way of a non-linear activation function (the *tanh* function in this case), to output signals. These are similarly processed by the single output unit to produce the network's output signal. The learning procedure records the error between this output and the target output, for each input pattern in the training set, and then tries to minimise:

$$E = \tfrac{1}{2}\sum_{P} (y^{p} - T^{p})^{2} \qquad (1)$$

where y^p is the network's output for input pattern $\underline{x}^p$ and T^p is the target output. The error E is minimised by gradient descent, or by more efficient algorithms such as quasi-Newton or conjugate gradient methods. The input data are presented to the network several times until a suitable error threshold is reached. For a more detailed introduction to neural networks interested readers are referred to Hertz *et al.* [2].

Many neural network algorithms are closely allied to techniques used in nonparametric statistical inference, but their theoretical foundations are not as well developed. However, unlike linear or quadratic discriminant analysis, multi-layer perceptrons are nonparametric, since no prior assumption is made about the mapping relationship. Various comparative studies of statistical, neural network and machine learning techniques, such as the ESPIRIT STATLOG project, are being carried out. The statistical aspects of artificial neural networks have been discussed by Ripley [8].

PREDICTION OF COMMUNITY STRUCTURE

In the field study a total of twenty-three different taxa were used to describe the community structure, of which four were used in this project; namely Porifera, Oligochaeta, Chironomidae, and Pelecypoda. Each of these were present in a large proportion (>76%) of the samples, and represented 93.5%, 3.5%, 0.7%, 0.4% respectively of the total number of animals found. Their recorded populations formed the basis of the target outputs of the training sets. The 28 environmental variables recorded at each sites, as outlined in Table 1, were used to produce three different input sets:

a) a set comprising the 28 raw variables;
b) a set of 7 eigenvectors, derived from a principal component analysis of the correlation matrix of the raw variables; and
c) a subset of the raw data consisting of 9 variables that was considered the best vector for site classification (i.e. pH, alkalinity, V, CaO, Cr, SiO, Co, NaO and the % loss on ignition) from a stepwise multiple discriminant analysis.

The principal component analysis used in set (b) can be viewed as a pre-processing measure used to reduce the amount of noise in the data, as the 7 derived factors explained 80.2% of the variance in the raw data.

Thus three different networks were trained for each of the four taxa. In all cases, the network used was a multi-layer perceptron trained using the standard back propagation algorithm. All input variables were rescaled to zero mean and unit variance, and those which appeared exponential in nature were log transformed using $log_{10}(x+1)$ prior to rescaling. In addition, for the output layer, the log transformation was also applied to the Oligiochaeta and Porifera data so that the resultant scaling reduced the importance of these taxa when they occurred at sites in extremely large numbers.

To make best use of the data, 1-fold cross validation training was used. This is a data resampling scheme that forms a nonparametric assessment of the error, and is generally used for predictive models were there is only a small number training examples. It incorporates the "leave-one-out" philosophy of a jackknife estimator, but uses the omitted sample to test the predictive capabilities.

RESULTS

Figure 2 shows six typical sets of results, expressed in terms of predicted abundance plotted against observed abundance. Graphs (a) and (b) show the networks' ability to reproduce the target outputs of their training sets; that is, the test cases were drawn from the training data. Graphs (c) to (f) show the networks' ability to predict abundance levels from the cross validated, independent, data input sets not previously "seen" by the networks. Graphs (c) and (d) show the results obtained from networks trained on the PCA and 9 component input data sets respectively. All axes have been transformed to a log. scale, and n and m are the number of input and hidden nodes respectively.

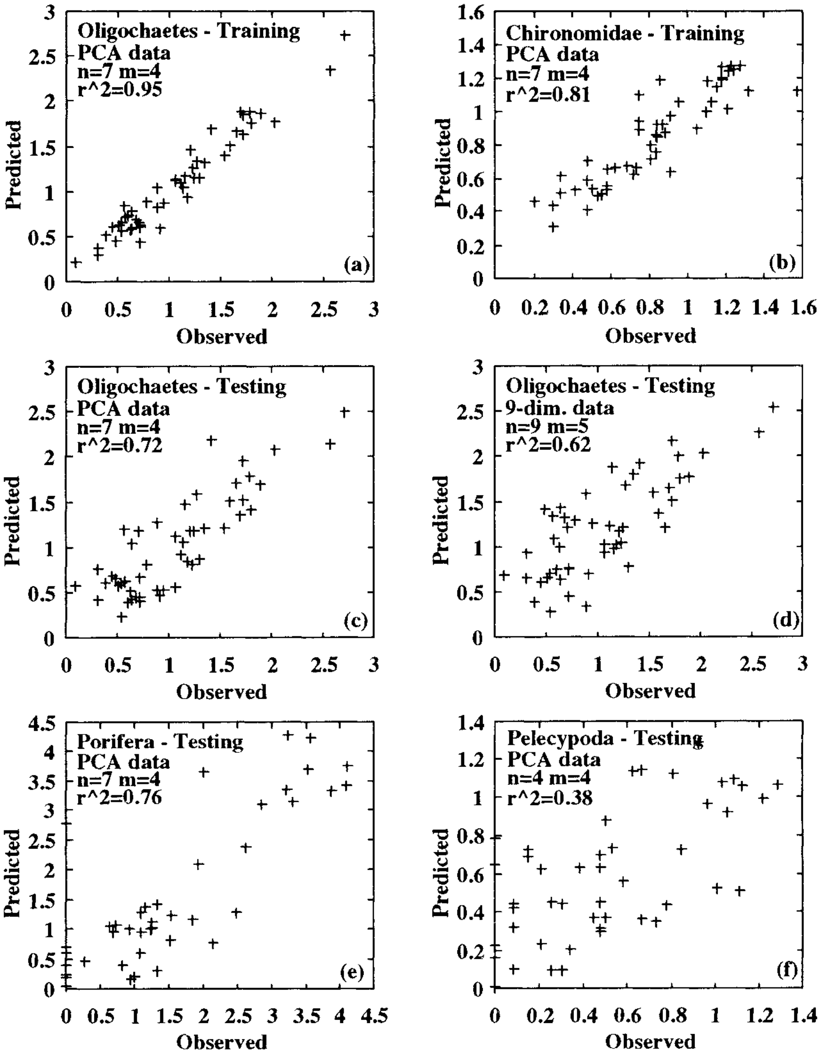

Figure 2. Some examples of the performance tests' results.

DISCUSSION

There is an increasing trend towards the use of biological standards in addition to the more traditional water quality objectives based on chemical concentrations. In the Laurentian Great Lakes the Governments of Canada and the United States committed themselves in the 1987 Great Lakes Quality Agreement to the further advancement of ecosystems development. In the state of Ohio, the Environmental Protection Agency has developed water quality regulations and issued discharge permits that use standards based on characteristics of the fish and the benthic invertebrate communities (Yoder, [11]) and in the United Kingdom the National Rivers Authority has supported the development of a river classification scheme based on the benthic invertebrate communities.

However, one of the difficulties with any regulatory or guideline system based on the biological endpoints is the inherent spatial variability in biological systems, which are by their nature adapted to local conditions. Methods are therefore required that can predict what the expected biological characteristics of a site should be, so that an appropriate guideline can be used. This approach has two necessary components. The first is the establishment of a reference site data technique from which the expected conditions can be defined. The second is a predictive technique that can define what the expected conditions at the test site can be. This approach was pioneered in the United Kingdom by the work of Wright *et al.* [10], the approach being to use classification analyses and multiple discriminant analysis as the prediction tool.

Despite the present availability of only 50 training samples, the results of this pilot study have shown that neural networks have potential as predictors of benthic community populations. Of the four taxa used for the tests, only one (Pelecypoda) produced unsatisfactory results, Figure 2(f). In the other cases, the performance tests carried out on the PCA data sets gave r^2 values of at least 0.72, which is considered quite acceptable at this stage. Improved predictive performance of the networks is expected with the use of a larger set of training data, as problems arising from the under determined nature of the models and data over-fitting will then be reduced. Confidence in their performance can be further increased by introducing a validation filter (Bishop and James [1]), which identifies input data sets that differ significantly from those on which the network was trained. Thus the user knows when the network is interpolating between training points and when it is extrapolating, hence predictions that are possibly unreliable are highlighted.

CONCLUSION

A new method of predicting benthic community structure, based upon artificial neural networks, has been investigated using data from the Canadian side of the Laurentian Great Lakes. The results demonstrate that the method offers a possible alternative to statistical methods of prediction, but that its performance in this pilot study was limited by the small number of data samples available. This non-linear approach to prediction clearly has potential for use in biological monitoring systems, and improved performance is expected as more training data become available and improvements to pre-processing are made.

ACKNOWLEDGEMENTS

The work described in this paper was carried out, in part, under a UK Science and Engineering Research Council studentship.

REFERENCES

1. Bishop, C.M. and James, G.D. (1992) *Analysis of multiphase flows using dual energy gamma densitometry and neural networks.* AEA-InTec-1032 United Kingdom Atomic Energy Authority Report.

2. Hertz J, Krogh, A. and Palmer, R. (1991) *Introduction to the Theory of Neural Computation.* Addison-Wesley.

3. International Joint Commission (1987) *Guidance of characterization of toxic substances problems in Areas of Concern in the Great Lakes basin.* Report from the Surveillance Work Group, pp179.

4. International Joint Commission (1988) *Procedures for the assessment of contaminated sediment problems in the Great Lakes.* Sediment Subcommittee, pp140.

5. Johnson, R.K. and Wiederholm, T. (1989) Classification and ordination of profundal macroinvertebrate communities in nutrient poor, oligo-mesohumic lakes in relation to environmental data. *Freshwater Biology* **21**, pp375-386.

6. Reynoldson, T.B. and Day, K.E. (1991) A study plan for the development of biological sediment guidelines. *NWRI unpublished report,* Burlington, Ont. pp35.

7. Reynoldson, T.B. and Zarull, M.A. (1993) An approach to the development of biological sediment criteria. In: *Ecological Integrity and the Management of Ecosystems.* (eds S.J. Woodley, G. Francis, J. Kay) St Lucie Press, Fl. pp177-200.

8. Ripley, B.D. (1993) Statistical aspects of neural networks. In: *Chaos and Networks - Statistical and Probabilistic Aspects* (eds. O.E. Barndorff-Nielsen, D.R. Cox, J.L. Jensen and W.S. Kendall). London: Chapman & Hall.

9. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986) Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1. Foundations* (eds. D.E. Rumelhart, J.L. McClelland, and the PDP group, pp318-362. Cambridge, MA: MIT Press.

10. Wright, J.F., Moss, D., Armitage, P.D. and Furse, M.T. (1984) A preliminary classification of running water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. *Freshwater Biology* **14**, pp221-256.

11. Yoder, C.O. (1989). The development and use of biological criteria for Ohio surface waters. In: *Proc. Water. Qual. Stand. 21st Century.* U.S. Environ. Prot. Agency, Criteria/Stand. Div Washington D.C., pp139-146.