Generic Artificial Neural Network Framework for Habitat Assessment and Prediction of Australian Stream Systems

N. Horrigan* and F. A. Recknagel

School of Earth and Environmental Sciences, University of Adelaide, SA 5000, Australia

*corresponding author: nelli.horrigan@adelaide.edu.au

Abstract: The Stream Decision Support System (SDSS) is taking advantage of both supervised and non-supervised artificial neural networks (ANNs) for stream assessment and prediction by an integrated approach. Non supervised ANNs were applied for patterning the natural variability in stream macroinvertebrate communities in Queensland. Supervised ANNs were developed for the prediction of the occurrence of stream macroinvertebrates in Victoria based on "clean-water" approach. Supervised ANNs were also applied for the prediction of taxonomic richness of native macrophytes and macroinvertebrates in the stream system of NSW by means of multi-layer perceptron ANN. The future development of the SDSS and its applicability for environmental management is discussed.

Keywords: *Stream modeling; Stream Decision Support System; Dirty-water approach*

1. INTRODUCTION

In the last decade supervised as well as non-supervised artificial neural networks (ANN) have been applied successfully to elucidate non-linear relationships between environmental variables (Chon et al., 1996; Lek et al., 1996; Huong and Recknagel, 2003) and predict habitat conditions in stream ecosystems (Walley and Fontama, 1998; Schleiter et al., 1999; Huong et al., 2001). These results have demonstrated that ANN models can improve understanding and prediction of processes in freshwater ecology.

The present paper outlines the structure and functioning of the stream decision support system SDSS that integrates ordination and clustering of complex stream data using non-supervised ANN and prediction of stream habitat conditions by supervised ANN using both "clean-water" ("reference") and "dirty-water" approaches. SDSS is designed to be generic for Australian stream systems. In the context of this paper example applications of the prototype SDSS to the stream databases of Queensland, NSW and Victoria are discussed.

2. DATA

The main database for the development of the SDSS was provided by the Queensland Department of Natural Resources and Mines. It contains presence and absence data of 168 macroinvertebrate families sampled from 1994 to 2002 at 859 sites, and a number of physical and chemical variables selectively used for different parts of the SDSS. In order to test the generic applicability of the framework we used two additional databases provided by the Victorian Environmental Protection Authority and the NSW Department of Land and Water Conservation. The database from Victoria contained occurrence data of 128 macroinvertebrate families sampled at 407 stream sites from 1990 to 1998. The database from NSW was the result of a Multi-Attribute River Assessment (MARA) survey of 122 sites on unregulated streams in 12 sub-catchments within four catchments. All three databases contained slightly different sets of variables which were divided into three subsets:

1. Physical settings and diversity (geographical position, altitude, slope, distance from source, rainfall, substrate heterogeneity, etc.)
2. Biological variables (macrophyte category, number of macrophyte taxa, presence and absence of macroinvertebrate taxa, abundance of macroinvertebrate taxa, number of diatom families, number of native fish species, etc.)
3. Risk factors (flow, water temperature, phosphorus, nitrogen, oxygen, organic matter, etc.)

3. METHODS

Stream Decision Support System SDSS

Figure 1 illustrates the principal structure and the corresponding functionality of the SDSS. The interactive user interface supports both: (1) the access to stream databases and supervised and unsupervised ANN models, and (2) the visualization of modeling results. The stream databases are structured into physical, biological and risk variables. The unsupervised ANN models process the stream data for spatial ordination, clustering and diagnosis of stream sites. The supervised ANN models allow the prediction of the occurrence and abundance of aquatic macroinvertebrates depending on stream habitat and water quality conditions. In addition they can be used for elucidating relationships between physical and biological variables by means of sensitivity analysis and conducting scenario analysis on potential impacts or restoration measures. Application of the sensitivity analysis for elucidation of hypothetical relationships between habitat conditions and macroinvertebrate assemblages in Queensland streams has been described by Hoang et al. (2003). Scenario analysis features are still under development and are not addressed in this paper.

Unsupervised ANN Models

Kohonen (1982) invented unsupervised ANNs (also known as Self-Organising Maps (SOM)) that enable patterns in data to be discovered without learning, cluster the data into a predefined number of classes, and order the classes in a two-dimensional output space such that near neighbours in the data space are near neighbours in the output space. They have proven to be a very useful tool for ordination and clustering of ecological data (Chon et al. 1996).

The ordination and clustering of stream data by SDSS is independent from the predictive modelling by supervised artificial neural networks and improves the understanding of spatial and temporal patterns of various variables. However, provided that the number of spatial classes is properly labelled corresponding with ecological properties of particular stream habitats, SOMs can be used as diagnostic tools for the assessment and continuous monitoring of freshwater habitats. Even though SOM models have been developed for both reference and testing sites of the stream systems of Queensland, Victoria and NSW, only results on the natural variability between reference sites in Queensland are documented in the context of this paper. The Queensland data were sampled from 111 sites of a riffle habitat in spring. The SOM for Queensland streams was generated by an unsupervised artificial neural network supported by Matlab 5.3 and the freeware SOM toolbox from The Laboratory of Computer and Information Science (CIS) at the Helsinki University of Technology. The input layer of the neural network contained 16 physical habitat variables, and presence-absence data of 158 macroinvertebrate families. The resulting SOM consisted of 90 cells and was partitioned into 4 clusters using the k-means algorithm. The 4 clusters were finally visualised in the Queensland map by means of ArcView 3.2.

Supervised ANN Models

Supervised artificial neural networks based on the backpropagation algorithm (Rumelhardt et al., 1986) are preferred tools for predictive modeling in ecology. They are typically characterized by input, hidden and output layers where the hidden layers contain internal connection weights between input and outputs which are steadily modified during training in order to minimize the error between predicted and observed outputs. Supervised artificial neural networks are not

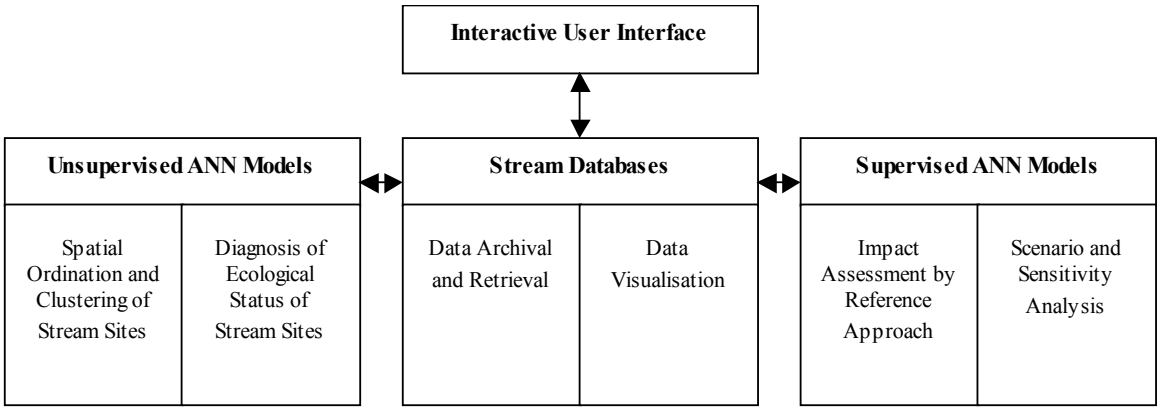

Figure 1. Structure and functioning of the stream decision support system SDSS

limited in complexity and are particularly well suited for multiple nonlinear cross-sectional (Lek et al., 1996) and time-series data (Recknagel et al., 1997).

Supervised artificial neural networks were successfully applied to develop predictive models for specific macroinvertebrate taxa in the Queensland stream system (Hoang et al., 2001) where cross-sectional data based on both the reference and the "dirty-water" approach were utilised. In the context of this paper we demonstrate the predictive ability of ANN using both "reference" and "dirty-water" approaches. We developed predictive models for the occurrence of stream macroinvertebrates in Victoria ("reference" approach), as well as for a number of native macrophyte species and macroinvertebrate families in the stream system of NSW ("dirty-water" approach) by means of multi-layer perceptron neural networks with sigmoid transfer function.

21 variables for physical and biological habitat properties were used as inputs and binary data for the occurrence of 15 macroinvertebrate taxa were used as an output for the Victoria stream model. The 15 output taxa were randomly chosen in order to validate the models' accuracy for common and rare taxa, where 5 were considered to be very common (at more than 70% of sites), 5 to be common (at about 50% of sites) and 5 to be uncommon (at less than 30% of all sites). The accuracy of the ANN predictions was estimated as the percentage of correct predictions. The models have been developed using the Neuro Solutions 4 software. The Cross-validation technique has been used to determine the optimum architecture of the ANN and prevent overtraining. For the NSW data we used a combined set of 20 physical, chemical and biological predictor variables. The number of invertebrate families and native macrophyte species were used as outputs variables. The models were developed using Matlab 5.3 software. Because of the small size of the database we could not spare a subset for cross-validation purposes. Instead, models were trained using Bayesian regularisation (Foresee and Hagan, 1997). The accuracy of the ANN predictions was estimated as the correlation between actual and predicted output.

For both Victorian and NSW data, 30% of each dataset were chosen at random for the purpose of model validation. This data has not been used for training of the models, to avoid any possible confounding of the results by overtraining. Only the results for validation sets are demonstrated below.

4. RESULTS

4.1. Classification

In order to understand patterns of natural variability in Queensland streams we developed a SOM for a riffle habitat based on 16 environmental variables and 158 macroinvertebrate taxa. As a result we identified 4 distinctive clusters and 1 transitional cluster. These clusters are shown in Figure 2. The two partially overlapping clusters 2 and 4 cover South Eastern areas; cluster 1 is concentrated at the wet tropics area and large inland areas belong to cluster 3. Several samples overlap between clusters 1 and 4 which seems to indicate that the area of Central Queensland Coast should be considered as a separate spatial cluster.

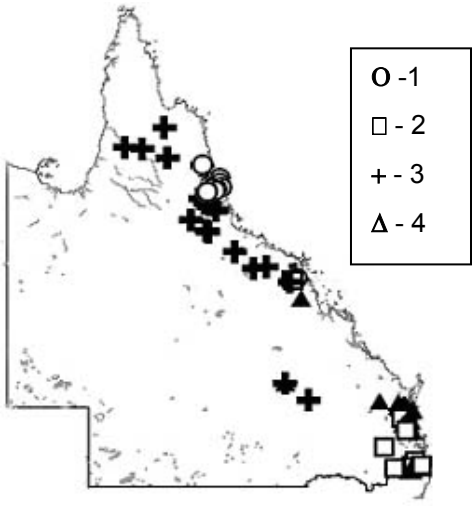

Figure 2. Application of SOM for Clustering of combined data for environmental variables and macroinvertebrate taxa.

Figure 3(a) shows the number of macroinvertebrate families found in riffle habitats of QLD streams. Despite the patchiness of the raw data it is clear that areas corresponding to different clusters show different patterns in diversity of macroinvertebrate fauna. Cluster 4 is characterised by the poorest taxonomical diversity (16 taxa on average). The rainforest areas behind clusters 1 and 2 are inhabited by the richest macroinvertebrate fauna (on average 24 to 23 taxa), with the inland cluster 3 being transitional (21 taxa on average). Even though the richness of aquatic macroinvertebrates is affected by various factors, rainfall data correspond the most with the previously found patterns. Fig. 3(b) shows the distribution of annual rainfall and Fig. 3(c) the ratio of monthly rainfall between the wet and the dry seasons. Cluster 1 corresponds with the highest annual rainfall (2551.808 mm in the average) and cluster 3 with the lowest (982.16mm). Cluster 5 is clearly different from its

surroundings, which supports the idea that it should be considered as a separate cluster. The pattern of seasonal rainfall (Figure 3(c)) explains well the difference between southern and northern parts of Queensland with clusters 1 and 3 having higher rainfall in the wet season in comparison with clusters 2 and 4 where the difference between seasons is not very significant.

4.2. Prediction

Results of the modelling data from Victoria are shown in Table 1.

Table 1. Percent of correct predictions of occurrence of macroinvertebrates in streams of Victoria (validation set).

	Taxa	% correct predictions
Very common	Oligochaeta	68.44
	Acarina	83.11
	Dytiscidae	74.22
	Elmidae	79.56
	Tipulidae	72.88
Common	Psephenidae	75.56
	Scirtidae sp	69.33
	Ceratopogonidae	67.55
	Coloburiscidae	84.44
	Physidae	82.67
Uncommon	Gordiidae	87.56
	Dugesiidae	78.22
	Ancylidae	74.22
	Ceinidae	92.00
	Gyrinidae	76.00

The average percent of correct predictions for all 15 taxa was 77.71%, for very common and common taxa 75.64 and 75.91 respectively, and slightly higher (81.6%) for uncommon taxa.

For the NSW data correlation between predicted and actual output on the validation set was 0.7 for the Number of Macroinvertebrate Families and 0.79 for the Number of Native Macrophyte Species (Figure 4).

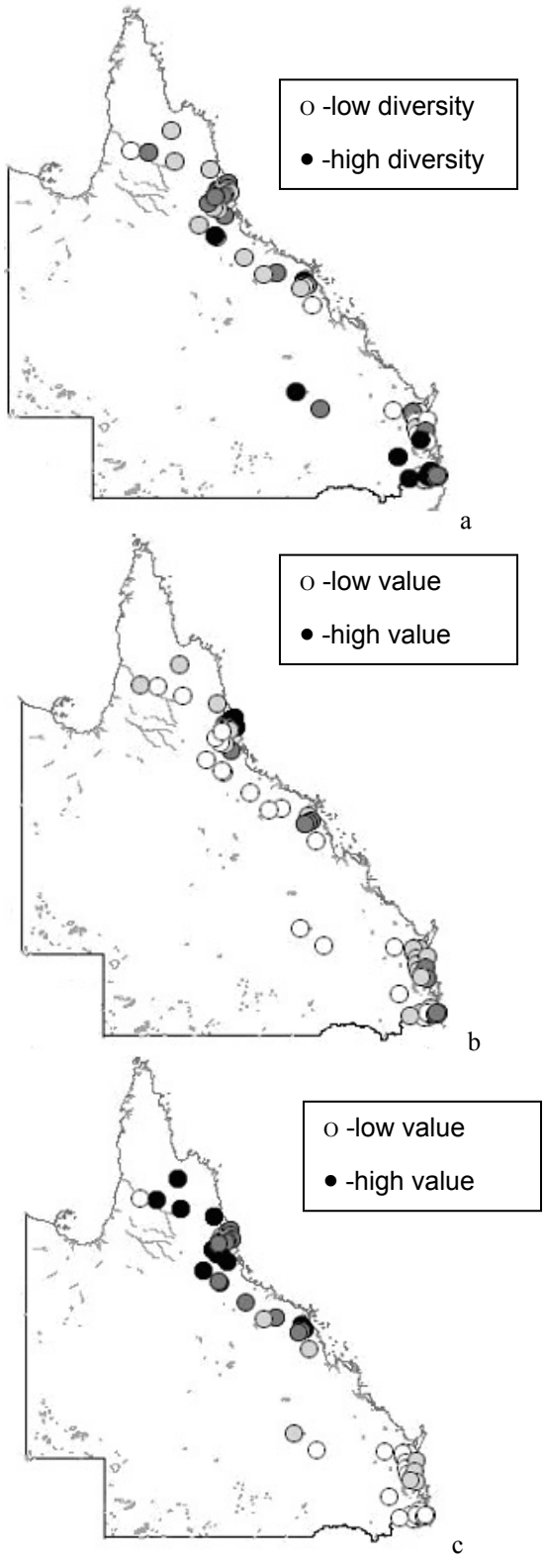

Figure 3. Distribution of a) macroinvertebrates diversity (number of macroinvertebrate families) (b) rainfall (c) ratio of the mean wet season monthly rainfall to the mean dry season monthly rainfall.

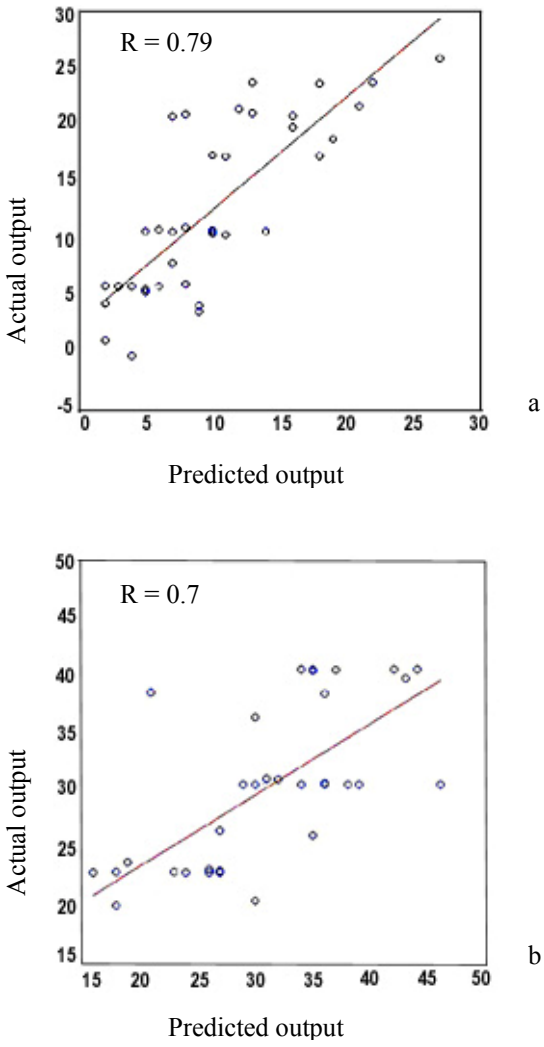

Figure 4. Predicted output versus actual output for the validation set (30% of the database not used for training) for a) number of native macrophytes species b) number of families of stream macroinvertebrates from NSW.

5. DISCUSSION AND CONCLUSION

The aim of this study was to demonstrate the application of the SDSS framework for the elucidation and prediction of Australian stream habitats. The proposed framework SDSS allows understanding of spatial patterns in stream data using unsupervised ANNs (SOMs). After Horrigan et al. (2003) have successfully applied SOMs for the classification of freshwater macroinvertebrate assemblages in Victoria we developed the method further to determine spatial patterns for natural variability of catchments using biological as well as physical variables and explaining those patterns by the corresponding component maps. This approach reveals unusual patterns and outliers that proves to be useful for environmental impact assessment. It can be developed further to a nationwide diagnostic system for catchment and stream systems.

We applied supervised ANN for the predictive modeling of the Victorian stream system based on the "reference approach". The correct predictions for stream macroinvertebrates in Victoria varied from 68.44 to 92% with an average of 77.71% which were slightly weaker than the results for the Queensland stream system obtained by Hoang et al. (2001) using the same "clean-water" or "reference" approach. This difference may result from the use of only 21 predictor variables for the Victorian stream compared to 39 for the Queensland streams.

Hoang et al. (2001) developed predictive models for the occurrence of 40 stream macroinvertebrate families using the "dirty-water" approach. Data used for this study were taken from both reference and degraded sites (2056 samples) containing physical and chemical variables. The average prediction accuracy of ANN models was 97%.

In this paper we also tested the applicability of ANN for modeling taxonomical diversity of macroinvertebrates and macrophytes in NSW streams by using the "dirty-water" approach. This was more challenging in comparison with the previous study because of: a) the limited size of the dataset, and b) variables being modeled were overall taxonomic reachness rather than single taxa occurrence.

Even though there were only 122 samples of the NSW streams system available, results of the predictive modeling of two biological variables for the "dirty-water" approach demonstrated that supervised ANNs can even cope with relatively small datasets from the diverse range of geographical locations and habitats.

The development of "dirty-water" models leads to "what if" or scenario analysis. It will allow not only to review known impacts of the past but also to predict potential impacts emerging from urban development and global changes on Australian stream ecosystems.

Extensive databases (like the database from QDNR) collected over many years over vast areas are most likely to contain the information on various conditions including extreme events like flood and drought. When ANN has learned the respective pattern from such data, it should be possible to model it in a range of different geographical locations and conditions. In a similar approach Dedecker et al. (2003) have assessed sensitivity and robustness of predictive neural network ecosystem models for the simulation of different management scenarios

using small dataset (120 samples). Three case studies have shown that ANN models are in general quite robust with a rather high predictive reliability.

The stream decision support system SDSS provides a flexible framework for further development. While new data and models can easily be integrated, the potential users have also easy access to archival and retrieval of data.

6. ACKNOWLEDGMENTS

This research is funded by the Queensland Department of Natural Resources and Mines. We would like to thank Satish Choy, John Marshall and the laboratory of aquatic ecology for providing data and valuable insight. We also thank Leon Metzeling from Victorian Environmental Protection Agency and Bruce Chessman from Centre for Natural Resources, NSW Department of Land and Water Conservation for providing additional data.

6. REFERENCES

Chon, T.S., Y.S. Park, K.H Moon and E.Y. Cha,. Patternizing communities by using an artificial neural network, *Ecological Modelling* 90, 69-78, 1996.

Dedecker, A. P., P. L. M. Goethalis, N. De Pauw, Sensitivity and robustness of predictive neural network ecosystem models for simulation of different management scenarios. *Ecological Modelling* (in press), 2003.

Foresee, F. D., and M. T. Hagan, Gauss-Newton approximation to Bayesian regularization, *Proceedings of the 1997 International Joint Conference on Neural Networks*, 1997.

Hoang, H., F. Recknagel, J. Marshall and S. Choy, Predictive modelling of macroinvertebrate assemblages for stream habitat assessments in Queensland (Australia), *Ecological Modelling* 146, 1-3, 195-206, 2001.

Hoang, H., F. Recknagel, J. Marshall, and S. Choy, Elucidation of hypothetical relationships between habitat conditions and macroinvertebrate assemblages in freshwater streams by artificial neural networks. In: Recknagel, F. (ed.), 2003. *Ecological Informatics. Understanding Ecology by Biologically-Inspired Computation*. Springer Verlag-Berlin, Heidelberg, New York, 179-190, 2003.

Horrigan, N., J., Bobbin, F. Recknagel, and L. Metzeling,. Patterning, Prediction and Explanation of Stream Macroinvertebrate Assemblages in Victoria (Australia) by Means of Artificial Neural Networks and Genetic Algorithms. *Ecological Modelling* (in press), 2003.

Kohonen, T., Self-organised formation of topologically correct feature maps, *Biological Cybernetics* 43, 59-69, 1982.

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

Schleiter, I. M., D. Borchardt, R. Wagner, T. Dapper, K. Schmidt, H. Schmidt and H. Werner, Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks, *Ecological Modelling* 120: 271-286, 1999.

Recknagel, F., M. French, P. Harkonen and K. Yabunaka. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling* 96, 1-3, 11-28,1997.

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

Walley, W. J. and V. N Fontama. Neural network predictors of average score per taxon and number of families at unpolluted river sites in the Great Britain. *Water Resources Research* 31(2): 201-210, 1998.