

**Neural Network Based Decision  
Support  
Framework for the Assessment and  
Management  
of Freshwater Stream Habitats**

A thesis submitted for the award of Doctor of Philosophy

Nelli Horrigan  
Discipline of Environmental Biology  
School of Earth and Environmental Sciences  
The University of Adelaide, Australia

February, 2005

II

Copyright © 2005 Nelli Horrigan

# Abstract

Modelling of stream macroinvertebrate communities has been widely accepted as an interesting and powerful tool to support water quality assessment and management. Stream Decision Support Framework (SDSF) offers an alternative approach to the current statistical models as Australian River Assessment Scheme (AusRivAs) for the derivation of scientific basis to support management applications regarding fresh water systems. Implementation of Artificial Neural Networks (ANNs) offers a possibility to overcome constraints of the statistical methods in dealing with high non-linearity of stream data.

This thesis includes several case studies illustrating application of Self Organising Map (SOM) and Multilayer Perceptron (MLP) neural networks to various tasks involving analysis, assessment and prediction of stream macroinvertebrates in three Australian states. The data for this study have been provided by the Queensland Department of Natural Resources (NR&M), EPA Victoria and the Department of Land and Water Conservation, New South Wales (NSW).

SDSF approach utilises predictive models for both 'referential' and 'dirty-water' approaches. Applicability and high accuracy of ANN models for the purpose of prediction both occurrence of individual taxa and taxonomic richness of stream macroinvertebrates have been demonstrated using data from Victoria and NSW. A comprehensive analysis of salinity sensitivity of stream macroinvertebrate has been demonstrated using both types of ANNs plus statistical methods, and pressure specific Salinity Index was suggested as a measurement of changes within macroinvertebrate communities in response to the secondary salinisation. Scenario analysis of the combined effect of increasing salinity and nutrient load demonstrated predictability and ecological meaningfulness of the Salinity Index.

Application of SOM has been demonstrated using the data from Queensland and Victoria in order to analyse natural variability of macroinvertebrate communities between reference sites. SOM component planes provided a valuable insight into the relationships between abiotic variables (as water quality and geoclimatic factors) and distribution of taxa and trophic structure of macroinvertebrate communities. Potential of SOM as data exploration tool has been also demonstrated for the analysis of the output of scenario simulation in order to understand the difference in response to salinisation in different sites.

Flexibility and potential of SDSF have been illustrated by using the combination of SOM and MLP, and combination of ANNs with statistical methods. Application of both SOM and Canonical Correspondence Analysis allowed the extraction of additional information and provided convenient visualisation of the relationships between water quality factors and the structure of macroinvertebrate communities.

In general, SDSF provides convenient, flexible and accurate approach for the analysis, assessment and prediction of stream biota. In addition to the freedom from the limitations inherent to the traditional statistical methods it allows many more options than currently used modelling frameworks, namely: highly accurate predictions using

## IV

both 'referential' and 'dirty-water' approaches, sensitivity analysis, scenario analysis and pattern exploration using SOM.

## **Statement of originality**

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Libraries, being available for photocopying and loan.

Nelli Horrigan



# Acknowledgements

I would like to thank my supervisors, Prof. Friedrich Rechnagel and Dr. Satish Choy for their advice and support.

This work has been made possible with the financial support of the Queensland Department of Natural Resources and Mines (NR&M) and I would like to express my gratitude to the staff of NR&M, for their ideas, feedback and hard work put into the collection and processing of the data. In particular I would like to thank Jon Marshall, Glenn MacGregor and Jason Dunlop.

I was fortunate enough to travel and meet many talented researchers during my PhD candidature, in particular I would like to thank Prof. Chon, Mi-Young Song and students and staff of the laboratory of Freshwater Ecology, Pusan National University for their hospitality and an opportunity to share the knowledge and ideas.

I would like to express my appreciation to my fellow students and office mates for their encouragement, sharing and support: Anita Talib, Lydia Cetin, Amber Whelk, Jason Bobbin, Hugh Willson, Tumi Bjornsson, Hongqing Cao and others who I may forgotten to mention.

Also I must thank my friends and family, in particular Greg Horrigan, Galina Putzka, Yulia Hudson, Galina Yastrebova and Lena Kupriyanova for their support and motivation.



# Publications and Scientific Communications During Candidature

## *Conference presentations*

Horrigan, N., Recknagel, F.A., Bobbin, J., Metzeling, L. Patterning, Prediction and Explanation of Stream Macroinvertebrate Assemblages in Victoria (Australia) by Means of Artificial Neural Networks and Genetic Algorithms. *3rd Conference of the International Society for Ecological Informatics, ISEI 2002, Rome, Italy.*

Horrigan, N. and Recknagel, F.A. Generic Artificial Neural Network Framework for Habitat Assessment and Prediction of Australian Stream Systems. *International Congress on Modelling and Simulation MODSIM 2003, 14-17 July 2003, Townsville, Australia.*

Horrigan, N., McGregor, G. and Dunlop, J. Salinity Sensitivity of Queensland stream macroinvertebrates: what field data has to say. *Australasian Society for Ecotoxicology, INTERACT 2004, Conrad Jupiters Gold Coast, QLD, Australia, 4-8 July 2004.*

Horrigan, N. and Choy, S. Understanding the effect of water quality on the trophic structure of stream macroinvertebrate communities using Self Organising Feature Map Neural Networks. *4<sup>th</sup> Conference of the International Society for Ecological Informatics, ISE4, BEXCO Busan, Korea, 24-28 October, 2004.*

Horrigan, N., Recknagel, F. and Choy, S. Predicting effect of dryland salinity outbreaks on stream macroinvertebrate communities in Central Queensland, Australia. *4<sup>th</sup> Conference of the International Society for Ecological Informatics, ISE4, BEXCO Busan, Korea, 24-28 October, 2004.*

## *Peer reviewed papers*

Horrigan, N. and Recknagel, F.A. (2003). Generic Artificial Neural Network Framework for Habitat Assessment and Prediction of Australian Stream Systems. *Proceedings of the International Congress on Modelling and Simulation MODSIM 2003, 14-17 July 2003, Townsville, Australia, Vol 2, 813-818.*

Horrigan, N., Recknagel, F.A., Bobbin, J., Metzeling, L. (2005). Patterning, Prediction and Explanation of Stream Macroinvertebrate Assemblages in Victoria (Australia) by Means of Artificial Neural Networks and Genetic Algorithms. In: 'Modelling Community Structure in Freshwater Ecosystems'. Lek, S.; Scardi, M.; Verdonschot, P.F.M.; Descy, J.-P.; Park, Y.-S. (Eds.), XII, 518 p.

Horrigan, N., Choy, S., Marshall, J. and Recknagel, F. (in press). Response of stream macroinvertebrates to changes in salinity and the development of a Salinity Index. *Marine and Freshwater Research.*

Horrigan, N., Recknagel, F. and Choy, S. (submitted). Predicting the effects of secondary salinisation on stream macroinvertebrate communities using artificial neural networks. *Ecological Modelling*.

Horrigan, N. and Choy, S. (submitted). Exploring the effect of water quality on trophic structure of stream macroinvertebrate communities using Self Organising Map Neural Networks. *Marine and Freshwater Research*.

# Contents

<b>1 General Introduction .....</b>	<b>1</b>
<b>2 Literature review .....</b>	<b>4</b>
2.1 Bioassessment of freshwater systems by aquatic macroinvertebrates .....	4
2.2 Facilitating the bioassessment of freshwater streams by computer modelling .....	5
2.2.1 Prediction of stream conditions .....	5
2.2.1.1 Statistical approach .....	5
2.2.1.2 Artificial neural networks with supervised learning .....	8
2.2.2 Ordination and clustering of stream conditions .....	18
2.2.2.1 Multivariate statistical approach .....	18
2.2.2.2 Artificial neural networks with unsupervised learning: Self organising maps .....	20
2.3 Comparison of ANN with other methods .....	26
2.4 Aims and hypotheses of the thesis .....	28
<b>3 Material and Methods .....</b>	<b>30</b>
3.1 Data .....	30
3.2 Data preprocessing and modelling .....	37
3.2.1 Unsupervised neural networks: Self Organising Maps .....	38
3.2.2 Supervised Neural Networks: Multilayer Perceptron and Feedforward networks .....	41
3.3 Structure and functioning of the Stream Decision Support Framework .....	43
<b>4 Ordination, clustering and correlation hunting using Self Organising Maps (SOM).....</b>	<b>45</b>
4.1 Exploring natural variability with SOM using referential datasets from Victoria and Queensland .....	45
4.1.1 Victoria dataset .....	46
4.1.2 Queensland dataset .....	52
4.2 Exploring relationships between environmental variables and diversity of stream biota in four NSW catchments .....	63
4.3 Exploring the effect of water quality on trophic structure of the macroinvertebrate communities using SOM in combination with Canonical Correspondence Analysis (CCA) .....	71
<b>5 Predicting macroinvertebrate taxa and macroinvertebrate communities in freshwater streams by MLP.....</b>	<b>95</b>
5.1 Using the clean-water (or referential) approach and Victorian dataset .....	95
5.2 Using the dirty-water approach and NSW dataset .....	97

5.3	Optimisation of the modelling design in respect to the cost efficiency of environmental monitoring.....	100
5.3.1	How many predictor variables is enough?.....	100
5.3.2	Generic models versus local models.....	103
5.3.3	Matter of time .....	105
5.3.4	Habitat issue.....	108
5.4	Prediction of SOM defined groups: case study for the comparison of the evolutionary algorithms and supervised neural networks.....	110
<b>6</b>	<b>Defining the relationships between water quality and macroinvertebrates using sensitivity analysis with MLP and SOM component planes.....</b>	<b>117</b>
6.1	Investigation into stability and quantitative applicability of the sensitivity analysis using supervised neural networks .....	116
6.2	Response of stream macroinvertebrates to changes in salinity and the development of a Salinity Index .....	122
<b>7</b>	<b>Scenario analysis based on the dirty-water approach.....</b>	<b>147</b>
7.1	Predicting the effect of secondary salinisation on stream macroinvertebrate communities in Central Queensland. ....	146
7.2	Using methods in combination: analysing results of the scenario analysis with Self Organising Maps (SOM). ....	155
<b>8</b>	<b>Discussion and the recommendations for further research .....</b>	<b>160</b>
	<b>Bibliography .....</b>	<b>174</b>

# List of Tables

Table 3.1. Variables contained in dataset from QLD with minimum, maximum and mean values.....	32
Table 3.2. List of variables in Victoria data set with minimum, maximum and mean values. ....	35
Table 3.3. List of biotic and abiotic variables available in NSW dataset with minimal, maximal and mean values.....	37
Table 4.1. Mean values of the continuous environmental variables in each of 6 SOM defined clusters, total abundance of macroinvertebrate and number of macroinvertebrate families are added for comparison. (L) – the variable was provided log-transformed.....	51
Table 4.2. Description of environmental variables and results of univariate analysis between 6 SOM defined clusters. ....	52
Table 4.3. Mahalanobis distances between 6 SOM defined clusters (environmental settings used as independent variables). ....	52
Table 4.4. Mean values of abiotic variables in each of 12 SOM defined clusters.....	56
Table 4.5. Univariate analysis of variance between 12 SOM defined clusters using environmental variables as predictors.....	57
Table 4.6. Mahalanobis distance between groups using set of environmental variables as predictors. ....	57
Table 4.7. Results of the univariate analysis of variables between SOM defined clusters, macroinvertebrate taxa (only first 20 are shown). ....	61
Table 4.8. Description of the variables from NSW dataset. ....	65
Table 4.9. Mean values of FFG and water quality variables for 7 SOM defined clusters, riffle habitat. ....	81
Table 4.10. Univariate analysis of variance between groups using FFG only. ....	83
Table 4.11. Mahalanobis distance between groups using FFG only. ....	83
Table 4.12. Mahalanobis distance between groups using combination of natural settings and water quality variables. ....	84
Table 4.13. Univariate analysis of variance between clusters using FFG only, edge..	85
Table 4.14. Mahalanobis distance between 12 SOM defined clusters using FFG, edge habitat.....	87
Table 4.15. Mahalanobis distance between 12 SOM defined clusters using physical settings and water quality variables, edge habitat.....	87
Table 4.16. Mean values of FFG and water quality variables for 12 SOM defined clusters, edge habitat. ....	88
Table 4.17. Multivariate effects (in order of model selection). ....	89
Table 4.18. Multivariate effects (in order of model selection), edge habitat.....	91

Table 5.1. Percent of correct predictions of occurrence of macroinvertebrates in streams of Victoria (testing set). .....	96
Table 5.2. The list of predictor variables used for the development of dirty-water models. ....	98
Table 5.3. Subsets of predictor variables used to investigate the relationship between the number of predictor variables and accuracy of the model. ....	102
Table 5.4. Comparative accuracy (% of correct predictions) of taxa specific models trained using different sets of predictor variables. ....	103
Table 5.5. Comparative accuracy of the generic models and models trained and tested on data subset from different geographical regions, expressed as percent of correct predictions, validation subset, equalized data. ....	105
Table 5.6. Comparative accuracy of the season specific and mixed-seasons models (% of correct predictions). ....	107
Table 5.7. Comparative accuracy (% of correct predictions) of the models trained and tested on the data from the same habitat versus models trained on one habitat and tested on another. ....	109
Table 5.8. Mean square error (MSE) and percentage of correct predictions (PCP) by applications of ANN and GA for each of 6 groups. ....	113
Table 5.9. Characterisation of the macroinvertebrate assemblage group 1 by means of environmental variables in term of descriptive statistics from SOM and rule set from GA. ....	116
Table 6.1. Estimation of mean accuracy and standard deviation of individual models. ....	119
Table 6.2. Variability in the estimation of the predictor importance. ....	121
Table 6.3. List of abiotic variables used for the study with Product Moment correlation with conductivity. ....	127
Table 6.4. Taxa specific conductivity ranges, trends shown by frequency plots and sensitivity analysis with ANN and Salinity Sensitivity Score, edge habitat. ....	143
Table 6.5. Taxa specific conductivity ranges, trends shown by frequency plots and sensitivity analysis with ANN and Salinity Sensitivity Score, riffle habitat. ....	145
Table 7.1. Input variables used for training of the predictive neural network. ....	150
Table 7.2. Product-Moment correlations between water quality variables in Central Queensland. ....	151
Table 7.3. Mean values of the water quality variables and actual PST in four SOM defined clusters. ....	158
Table 7.4. Results of the univariate analysis of variance between clusters 1 and 2. .	159

# List of Figures

Figure 2.1. Scheme of the biological neuron (Rojas, 1996). .....	8
Figure 2.2. Artificial neuron (Rojas, 1996).....	9
Figure 2.3. Multilayer perceptron model for the prediction of stream macroinvertebrates from environmental variables (Huong et al., 2001). .....	13
Figure 2.4. Structure of the Kohonen's network (Chon et al., 1996). .....	21
Figure 3.1. Map of the sampling sites in QLD dataset. ....	31
Figure 3.2. Locations of sampling sites in Victoria data set.....	34
Figure 3.3. Locations of the sampling sites in NSW data set. ....	36
Figure 3.4. Structure of the Stream Decision Support Framework (SDSF). .....	44
Figure 4.1. SOM outputs: a) U-matrix, b) Partitioning into 6 clusters by the K-means algorithm. ....	47
Figure 4.2. a) Distribution of macroinvertebrate groups resulting from SOM (sites belonging to the same group have the same marker), b) biological regions in Victoria based on benthic macroinvertebrates (Metzeling et al., 2001). .....	48
Figure 4.3. a) SOM hit diagram showing distribution of 6 groups (clusters) on SOM grid, b) SOM component planes for the environmental variables (see Table 3.2 for abbreviations), all data normalized between 0 and 1, darker shades correspond to higher values. ....	49
Figure 4.4. a)U-matrix and, b) k-means partitioning into 12 clusters, QLD reference sites, macroinvertebrates only.....	53
Figure 4.6. Bioregions of Queensland based on aquatic macroinvertebrates, defined by NR&M (in preparation). ....	58
Figure 4.5. Distribution of 12 SOM defined clusters (clusters are shown in groups of 4 for readability), reference sites, only macroinvertebrates.....	59
Figure 4.7. Distribution of rainfall pattern in QLD: a) mean annual rainfall, b) mean dry season rainfall, c) mean wet season rainfall. ....	60
Figure 4.8. SOM component planes of the first 10 abiotic and first 10 biotic variables best discriminating between SOM defined clusters (all variables normalized between 1 and 0). ....	62
Figure 4.9. Hit diagram, hits for four catchments shown in different color. ....	65
Figure 4.10. SOM component planes for the natural settings and risk factors in NSW dataset. ....	67
Figure 4.11. SOM component planes for the biotic variables in NSW dataset. ....	68
Figure 4.12. Scatterplots for the number of macroinvertebrate families versus water temperature and turbidity.....	71

Figure 4.13. Riffle habitat, box and whiskers plots for median, range (20-80%) and non-outlier minimum and maximum of trophic groups in different stream order categories a) reference sites, b) test sites. ....	76
Figure 4.14. Edge habitat, box and whiskers plots for median, range (20-80%) and non-outlier minimum and maximum of trophic groups in different stream order categories: a) reference sites, b) test sites. ....	77
Figure 4.15. SOM component planes for the proportional values of functional feeding groups and water quality variables: water temperature (WTEM), conductivity (COND), dissolved oxygen (DO), alkalinity (ALK), turbidity (TURB), total nitrogen (N) and total phosphorus (P), see Table 1 for units. ....	79
Figure 4.16. SOM component planes for proportional values of feeding functional groups and water quality variables, edge habitat. ....	80
Figure 4.17. Box and whiskers plots for median values of FFG and turbidity. Box – 20-80%, whiskers – minimum and maximum. ....	82
Figure 4.18. Spatial position of sites within clusters 1, 5, 6 and 7. ....	83
Figure 4.19. Box plots demonstrating changes in the percentage of collectors along stream order gradient in different turbidity conditions a) turbidity <5 NTU, b) turbidity >10NTU, c) turbidity >20 NTU. ....	84
Figure 4.20. Box plots for median values of proportion of collectors and predators FFG in each of 12 SOM defined clusters, edge habitat. ....	88
Figure 4.21. Bi-plot resulted from CCA with SOM defined clusters and eight environmental variables. Pie charts show mean percentages of each FFG within a cluster. ....	89
Figure 4.22. Bi-plot resulted from CCA with SOM defined clusters and environmental variables. ....	91
Figure 5.1. 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. ....	99
Figure 5.2. Accuracy of the year 1994 model when tested on data collected during other years. ....	108
Figure 5.3. Conceptual framework for the study involving prediction and explanation of SOM defined groups using MLP and GA. ....	112
Figure 5.4. Structure of an evolved rule tree. ....	113
Figure 6.1. Box and whisker plots showing variability between individual models used for the sensitivity analysis of the relationship between three taxa and two water quality variables. Center – median value, box - 20-80%, whiskers – minimum and maximum values. ....	120
Figure 6.2. Distribution of conductivity values in Queensland dataset (ranges of conductivity taken from Williams (1967)). ....	125
Figure 6.3. a) SOM component plane for conductivity ( $\mu\text{S}/\text{cm}$ ), b) SOM grid with hits divided according to the classification by Williams (1967). ....	129
Figure 6.4. Selected SOM component planes for: a) variables positively correlated to conductivity values, b) variables negatively correlated to the conductivity values	

(normalized data, darker shades indicate higher values, broken outline indicates high conductivity corresponding to subsaline and saline categories by Williams (1967)).	130
Figure 6.5. Scatterplot of the taxonomic richness versus conductivity for: a) edge and, b) riffle habitats.	131
Figure 6.6. Selected SOM component planes for macroinvertebrate taxa, a) collected in mostly low salinity conditions (sensitive (s)), b) macroinvertebrate taxa collected in high salinities conditions (very tolerant (vt), darker shades indicate more frequent presence, broken outline indicates high conductivity corresponding to subsaline and saline categories by Williams (1967)).	132
Figure 6.7. Plots of (a) ‘decreasing’ and (b) ‘increasing’ trends of selected stream macroinvertebrates along the conductivity gradient, edge habitat.	133
Figure 6.8. Typical sensitivity plots resulting from the sensitivity analysis of MLP for selected taxa, edge habitat. a) ‘Decreasing’ and, b) ‘increasing’ trends in the probability of occurrence of macroinvertebrate taxa along the conductivity gradient.	134
Figure 6.9. Percentage of sensitive and very tolerant taxa in 12 equal sized bins along the gradient of increasing conductivity, a) edge habitat, b) riffle habitat. Median values with boxes corresponding to 80 <sup>th</sup> and 20 <sup>th</sup> percentiles and horizontal bars to maximum and minimum.	135
Figure 6.10. Scatterplots of SI versus conductivity with fitted logarithmic trends, a) edge, $y = -0.29\ln(x) + 6.03$ , b) riffle, $y = -0.53\ln(x) + 8.36$ .	136
Figure 6.11. Salinity Index in 12 equal sized bins along increasing conductivity gradient for: a) edge and, b) riffle habitats, only sites with good water quality.	137
Figure 6.12. Scatterplots of SI versus flow (maximum water velocity), water temperature, habitat depth (HDepth), mean phi, mean annual rainfall, distance from source (DFS), altitude and longitude.	138
b)	141
Figure 6.13. CCA biplots showing effect of water quality variables after the effect of natural and temporal variability was partialled out a) edge, b) riffle. S - sensitive taxa, T - very tolerant taxa, * - generally tolerant taxa, DO - dissolved oxygen ( $\text{mg L}^{-1}$ ), Total P – total phosphorus ( $\text{mg L}^{-1}$ ).	141
Figure 7.1. Map of the Fitzroy and Burdekin catchments with sites marked accordingly to their location in different salinity hazard zones.	149
Figure 7.2. Scatterplot of alkalinity versus conductivity with fitted trendline.	152
Figure 7.3. Scatter plot of pH versus conductivity with fitted trendline.	152
Figure 7.4. Scatter plot of turbidity versus conductivity.	153
Figure 7.5. Scatter plot of total phosphorus versus total nitrogen with fitted trendline.	154
Figure 7.6. Actual versus predicted a) Salinity Index, and b) Percent of sensitive taxa, simulation dataset.	154
Figure 7.7. Box plots for simulation results for Scenario 1 and Scenario 2, median values, box 20-80%, whiskers minimum and maximum.	155

Figure 7.8. Box plot for the PST outputs for the Simulation 5 (+ 4000  $\mu\text{S cm}^{-1}$ ), for Scenario 1(Conductivity), Scenario 2 (Combined) and only increase in nutrients (+ 4mg  $\text{L}^{-1}$  of total nitrogen, total phosphorus = 0.1251x (total nitrogen)- 0.01). ..... 156

Figure 7.9. Box plot for the median values (box 80-20%, whiskers for maximum and minimum) of differences between simulation outputs in two SOM defined clusters. .... 158

Figure 8.1. PAEQANN interface. .... 167

Figure 8.2. Ordination (using SOM) screenshot. .... 167

Figure 8.3. Prediction (using MLP) and sensitivity analysis screenshot. .... 168