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

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

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 of 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. *4th 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. *4th 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

List of Tables

List of Figures

Chapter 1

General Introduction

Since European settlement, anthropogenic effects on Australian rivers have been considerable. Sewage, detergents and agricultural runoff have altered nutrient balances, mine wastes have caused heavy metal pollution, acidification and sedimentation, while land clearing and deforestation have caused sedimentation and secondary salinisation (Smith et al., 1999). To address concerns about declining conditions in Australian rivers and stream a number of different assessment schemes and protocols have been designed with Australian River Assessment Scheme (AusRivAs) being currently most widely used (Schofield and Davies, 1996). Even though the application of AusRivAs achieved some valuable success, some constraints appeared to have caused confounded assessment of biological impairment.

AusRivAs is based on a referential approach, which assesses habitat conditions in a river by predicting the macroinvertebrate families expected to occur in the absence of environmental stress, such as pollution or habitat degradation (Coysh et al., 2000). Predictions are derived from a set of environmental measurements (only variables not affected by human activity) used to characterise the sites. A predicted macroinvertebrate assemblage is compared with the actual assemblage, and the ratio of observed/expected (O/E) families is used as a measure of ecological habitat conditions (Parsons and Norris, 1996; Marchant et al., 1999; Smith et al., 1999).

Macroinvertebrate communities are influenced by a number of physical, chemical and biological factors. It is impossible to predict or analyse the effect of potential anthropogenic disturbances on the stream biota using only the referential approach as it utilises only predictor variables potentially unaffected by humans. ‘Dirty-water’ approach allows prediction of the possible consequences from various impacts by utilising distribution of macroinvertebrates and habitat characteristics from both reference and potentially impacted sites. ‘Dirty- water’ models utilise a wider range of input variables, including those that can be altered by anthropogenic impacts. This type of models can be used for two purposes: sensitivity analysis and scenario analysis. The principle behind scenario and sensitivity analysis is the same: trained models are simulated while variables in question are altered, and simulated output is analysed. However, the specific methodology and application of the sensitivity and scenario analyses can differ depending on the task.

Ecological data usually consist of many species and environmental variables, which vary and covary in nonlinear fashion (Lek and Guegan, 2000). Thus, nonlinear

modeling methods such as Artificial Neural Networks (ANNs) should be preferred for dealing with such data (Blayo and Demartines, 1991). Implementation of Artificial Neural Networks (ANN) instead of statistical methods offers a possibility to overcome constraints of the statistical methods in dealing with high non-linearity of stream data.

Two types of ANN have been widely applied for the ecological problems: Self Organising Map (SOM) for the classification and ordination and Multilayer Perceptron (MLP) for the predictions. A number of ecological case studies have shown that SOMs are an efficient classification tool (Chon et al.,1996; Park et al., 2001; Park et al., 2003; Brosse et al., 2001; Giraudel and Lek, 2001; Cereghino et al., 2000). MLPs were successfully applied to predict the occurrence of stream macroinvertebrates from the environmental variables (Walley and Fontama, 1998; Schleiter et al., 1999; Pudmenzky et al., 1998; Huong et al., 2001), fish distribution (Joy and Death, 2004) and species richness (Cereghino et al., 2003).

This study explores the potential of ANNs application to a number of ecological problems in an intergrated manner described by the Stream Decision Support Framework (SDSF). SDSF offers an alternative approach to that of AusRivAs and other methods to provide scientific understanding of freshwater streams to support the management decisions regarding the sustainable use of fresh water systems. Using stream datasets from three Australian states we demonstrated the usefulness of the SDSF for assessments of biological conditions using referential approach, prediction of taxonomic richness and scenario analysis, finding similar patterns in macroinvertebrate communities and relating them to the environmental variables, derivation of pressure-specific ecological index using results of ANNs based sensitivity analysis and addressed practical questions posed by aquatic ecologists from the Queensland Department of Natural Resource and Mines (NR&M).

Organisation of the Thesis

Chapter 2 introduces principles of ecological assessment using stream macroinvertebrates and principles of ANNs modeling and application to the ecological problems.

Chapter 3 describes three datasets available for the study. Practical implementation of SOM and MLP in the scope of this study is explained.

Chapter 4 demonstrated the applicability and potential practical use of SOM using several case studies: grouping macroinvertebrate assemblages into similar spatial clusters and explaining the patterns found by the environmental variables using the datasets from Queensland (QLD), Victoria and New South Wales (NSW) (4.1-4.2), and analyzing the changes within the trophic structure of macroinvertebrate communities in response to water quality parameters (4.3).

Chapter 5 presents the results of several case studies using predictive ANN models: implementation of the referential approach using dataset from Victoria (5.1), prediction of taxonomic richness using ‘dirty-water’ models (5.2) and prediction of

SOM defined clusters (5.4). It also contains answers to the questions formulated by the scientists of the Laboratory of Aquatic Ecosystem Health, NR&M, regarding the optimisation of the modeling design in respect to the cost efficiency of environmental monitoring. It discusses temporal and spatial implications for the building of accurate predictive models and compares variety of models using different number of predictive variables (5.3).

Chapter 6 explores sensitivity analysis with MLP. First part contains the results of the investigation into a stability and quantitative applicability of the sensitivity analysis. The second part contains a comprehensive study of salinity sensitivity of QLD stream macroinvertebrates using sensitivity analysis and variety of the other methods as SOM and Canonical Correspondence Analysis, and proposes Salinity Index as measurement of changes within communities in response to the changes in salinity.

Chapter 7 presents results of the scenario analysis using MLP. The changes in the Salinity Index and the percentage of sensitive taxa were simulated in response to the increase in conductivity and combined increase in conductivity and nutrients concentration using data from Central Queensland. Untraditional application of SOM for the analysis of simulated outputs is demonstrated.

Chapter 8 contains general discussion of the results and possible directions for the future research.

Chapter 2

Literature review

2.1 Bioassessment of freshwater systems by aquatic macroinvertebrates

Stream and river ecosystems experience great pressure by human activities such as population growth and economic development. Protection and maintenance of high quality stream water has become an increasingly important issue in recent years.

Water quality can be measured by different parameters such as biological oxygen demand, suspended sediments and bacterial counts. However, these parameters only reveal the quality of the water at the time of sampling, and their further relevance has to be inferred by extrapolation from limited data (Hellawell, 1986). These measurements may be efficient for regulating effluent discharges and protecting humans, they are not very useful for a large-scale management of catchments or for assessing the state of the river ecosystems (Norris et al., 1999). Biological monitoring, on the other hand, generally is considered to provide a more integrated appraisal of water and overall environmental quality (Hynes, 1960). In numerous cases the parameters like community structure and taxonomic richness have been proved to be the most sensitive indicators for quickly and adequately detecting alterations in aquatic ecosystem (Cairns and Pratt, 1993).

Biological assessments are less time consuming than other methods as a single series of samples represents the sum effects of the prevailing conditions. In addition, animal and plant communities are little affected by a temporary amelioration or a transient deterioration of the effluent (Mason, 1996). Bioassessment can reveal long-term effects on ecosystems after the cause of the impact has passed and is itself undetectable. Such assessment provides both numeric and narrative descriptions of resource condition (Karr, 1998). Cairns and Pratt (1993) considered the role of the bioassay as a diagnostic tool for the restoration of desirable ecosystem conditions and as a predictive tool for preventing environmental impact.

Much emphasis is being placed on rapid biological assessment, particularly using indices such as the Index of Biological Integrity (IBI, Karr, 1981), Ephemeroptera Plecoptera Trichoptera (EPT), Biological Monitoring Working Party (BMWP) (Hawkes, 1998), bentic IBI (Kerans and Karr, 1994), SIGNAL (Chessman, 2003),

AusRivAs (Simpson et al., 1997), and River Invertebrate Prediction And Classification Scheme (RIVPACS; Wright, 1995).

Amongst aquatic animals that can be used in bioassessment, macroinvertebrates proved to be a superior indicator for the quality of freshwater streams (Rosenberg and Resh, 1993). Macroinvertebrates in streams have relatively long life cycle, exposing them to pollutants over a long period of time and integrating the effect of short-term pollution episodes. From the practical point of view they can be relatively easy sampled and identified.

Freshwater macroinvertebrates are ubiquitous; even the most polluted or environmentally extreme stream habitats usually contain some representatives of this diverse and ecologically important group of organisms. Macroinvertebrates play important roles within the stream community as a fundamental link in the food web between organic matter resources (leaf litter, algae, detritus) and fish (Hynes, 1970; Allan, 1995).

The responses of aquatic macroinvertebrate communities to environmental disturbances have been incorporated into methods of bioassessment and biotic indices for the bioassessment of aquatic ecosystems. Commonly observed responses to antropogenic stress include increased abundance of certain species on the one hand but general loss of diversity, especially when affected by pesticide load or organic enrichment (Cranton et al., 1996) on the other hand. However, the intermediate disturbance hypothesis, as modified for streams, predicts that the biotic diversity will be highest in communities subjected to intermediate levels of disturbance. At low levels of disturbance, competitive interactions will result in lower diversity because of exclusion of species. High disturbance also will result in lower diversity because of exclusion of poor colonists or long –lived species (Ward and Stanford, 1983).

2.2 Facilitating the bioassessment of freshwater streams by computer modelling

2.2.1 Prediction of stream conditions

Stream modeling based on ecological knowledge and sufficient stream monitoring data can substantially facilitate and further improve assessment of stream habitats (Huong et al., 2001). This chapter provides an overview and comparison of different modeling methods and approaches available for the prediction of stream biota.

2.2.1.1 Statistical approach

Ecological research requires statistical analysis or even more complex numerical analysis to draw generalities and to detect and highlight patterns or trends in complex data set consisting of many variables. The most commonly used statistical methods are Analysis of Variances (ANOVA), multiple regression (MR), Discriminant Function Analysis (DFA) and time series analysis. They are all very powerful tools for developing predictive models and associating physical, chemical and biological

data together. However, these methods of statistical analysis often have stringent requirements of data, such as replicated collection of data, normal data distribution or high frequency of data collection.

Some requirements are difficult to meet, so that simplified assumptions must be used to apply these methods. These assumptions and data requirements usually restrict the capability of statistical methods to cope with the non-linearity and complexity of water ecosystems. Statistical methods tend to minimize non-linearity in the processes. They are simple to implement if the relationships with variables are linear. If they are non-linear, transformation into linear becomes a major limitation of statistical methods in working with non-linear relationship of variables in the aquatic system (Lek et al., 1996; Paruelo et al., 1997).

Modelling has been widely accepted as an interesting and powerful tool to support river quality assessment and management. The River Invertebrate Prediction and Classification System (RIVPACS) based on the statistical modelling was one of the first and the best known systems in this context. RIVPACS was developed to classify macroinvertebrate community types and to predict the fauna expected to occur in different types of watercourses, based on a small number of environmental variables. The statistical techniques used for RIVPACS are TWINSPAN classification of the reference sites based on their macroinvertebrate assemblages, followed by multiple discriminant analysis (MDA) of the resulting groups of sites using a limited number of environmental variables. Prediction of the fauna at a test site was achieved through MDA, leading to the calculation of probabilities of capture of individual taxa based on the prediction of group membership for the test site (Moss et al., 1987). The prediction is essentially a static'target' of the fauna to be expected at a site with well defined environmental features, in the absence of environmental stress.

In Australia, a similar predictive model called Australian River Assessment Scheme - AusRivAS was developed to use aquatic macroinvertebrates to assess the habitat condition of Australian rivers and streams (Schofield and Davies, 1996). AusRivAS models are based on RIVPACS, which also assess habitat conditions in a river by predicting the macroinvertebrate families expected to occur in the absence of environmental stress, such as pollution or habitat degradation (Coysh et al., 2000). Predictions are derived from a set of environmental measurements used to characterise the sites. A predicted macroinvertebrate assemblage is compared with the actual assemblage, and the ratio of observed/expected (O/E) families is used as a measure of ecological habitat conditions (Parsons and Norris, 1996; Marchant et al., 1999; Smith et al., 1999). There are two major differences between AusRivAS and RIVPACS. Firstly, macroinvertebrates are only identified to the family level in AusRivAS. Secondly, major aquatic habitats (channel, riffle etc) are sampled and processed separately in AusRivAS (Smith et al., 1999). The rationale behind habitat – specific sampling is that each habitat has a distinct macroinvertebrate community and within a given region, differences among habitats are greater than differences between sites. Unless comparisons between sites are based on the same habitats, they may be confounded by the occurrence of different habitats at each site (Parson and Norris, 1996).

The modelling approach for AusRivAS was similar to that of RIVPACS. Model building occurred in five steps. First, reference sites were classified into groups with similar macroinvertebrate communities using an agglomerative hierarchical fusion

technique, Unweighted Pair-Group arithMetic Averaging (UPGMA). Second, once the optimal classification was chosen, stepwise discriminant function analysis (DFA) was used to identify which environmental variables discriminated best between groups in the classification. Third, the DISCRIM procedure in the SAS statistical package was used to incorporate predictor variables into a discriminant function and assign sites to groups identified in the classification. Fourth, the probability of each family occurring at each site was calculated by multiplying the probability of a site belonging to a classification group by the probability of the occurrence of family in that group and then summing the products to give the number of families expected (E). Fifth, using a preliminary model, O/E ratios of reference sites were calculated. The O/E score itself was used as a measure of impact at disturbed sites, with lower scores indicating greater disturbance (Simpson et al., 1997; Smith et al., 1999; Coysh et al., 2000).

The AusRivAS model had been applied to study the effect of habitat- specific sampling on the biological stream assessment for the Australian Capital Territory (Parson and Norris, 1996), to classify macroinvertebrate communities across drainage basins in Victoria (Marchant et al., 1999), and to assess ecological conditions of rivers in Western Australia (Smith et al., 1999). Even though the applications achieved some valuable success, some constraints appeared to have caused confounded assessment of biological impairment. Although statistics can be used to validate metric choices and predictions while building multimetric indices, excessive dependence on the outcome of statistical tests can obscure meaningful biological patterns. A narrow focus on probability values (P-value) rather than on biological consequences limits the value of biological assessment. Dependence on narrow statistical approaches overlooks the fact that a statistically significant result (small P-value) may not indicate a large important effect, as researchers often assume; similarly, a statistically insignificant effect (large P- value) may well be of biological significance (Karr, 1999).

An investigation of the RIVPACS classification based on statistical methods revealed that the composition of a few of the classification groups was less than optimal and could adversely affect the performance of parts of the prediction system (Wright et al, 1991). Moreover, the RIVPACS and AusRivAS statistical approach may be more difficult to apply to sites where environmental conditions are extreme or highly unpredictable. As a consequence it may be difficult for statistical methods to cope with substantial year by year variations of the biota (Wright, 1995).

These constraints are caused by the assumptions and limited implementation of statistical methods in dealing with the high non-linearity of stream data. As new computational techniques are becoming widely available, a number of alternative ordination and classification procedures are now being examined to determine whether a new procedure can deliver more reliable predictions (Wright, 1995).

Chessman (1999) pointed out that classification in RIVPACS and AusRivAs is an unnecessary step that could be avoided. Where the invertebrate fauna shows a continuum of variation rather than a discrete occurrence of community types, the use of cluster analysis seems particularly dubious since the groups defined after cluster analysis will be artificial and rather arbitrary. Chessman (1999) also suggested that the use of abundance instead of taxa absence-presence might prove more useful where pollution results in reduced abundances rather than the total elimination of sensitive taxa. He proposed an alternative to the RIVPACS method that predicts abundance and

does not require site classificaiton. Macroinvertebrate indices generated by the new method showed a greater distinction between human-disturbed and undisturbed test sites, and a similar or higher degree of correlation with physical and chemical indicators of human disturbance.

2.2.1.2 Artificial neural networks with supervised learning

Modelling freshwater quality is extremely difficult, as the interrelations between various influences are not well known. Highly complex and heterogeneous nature of ecological data requires methods, which could overcome the limitations and inflexibility of statistical methods.

ANNs are non-linear mapping structures based on the function of the human brain. They are considered universal and highly flexible approximators for any data and are powerful tools for ecological modelling, especially with high non-linearity occasions when the data relationships are unknown (Lek and Guegan, 2000). They do not require assumptions about mathematical relationships between state variables and the nature of the distribution of data. All neural networks have in common the ability to learn from data. ANNs can identify and learn correlations between input data and corresponding target values. After training, ANNs are able to predict the output of new independent input data.

The natural neuron has four main structural elements: dendrites, synapses, cell body and axon (Figure 2.1). Dendrites receive the signals at the contact regions with other cells called synapses. Organelles in the body of the cell produce all the necessary chemicals for the continuous working of the neuron. The output signals are transmitted by the axon, of which each cell has at most one (Rojas, 1996).

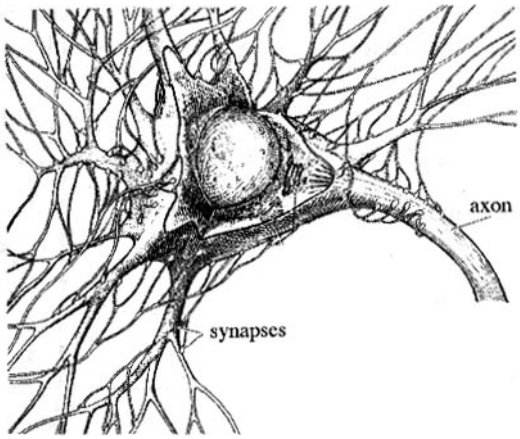


Figure 2.1. Scheme of the biological neuron (Rojas, 1996).

An artificial neuron for computing will have input channels, a cell body and an output channel. In ANNs, the computational or processing element is called a neuron (node

or unit). Like a real neuron, the neuron in ANN has many inputs, but only a single output which can pass an information to other neurons in the network. Figure 2.2 shows the structure of an abstract neuron with *n* inputs. Each input channel *i* can transmit a real value *x*. The function *f* computed in the body of the abstract neuron can be selected arbitrarily. Usually, the input channels have an associated weight, which means that the incoming information x_i is multiplied by the corresponding weight w_i. The transmitted information is integrated at the neuron (usually just by adding the different signals) and the function is then evaluated.

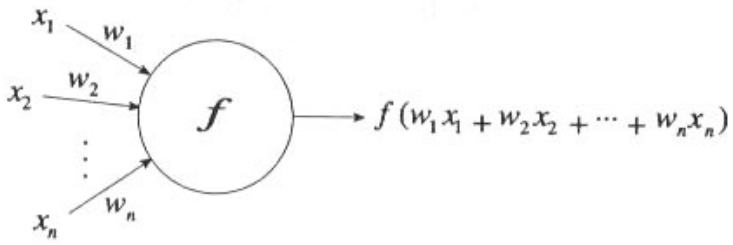


Figure 2.2. Artificial neuron (Rojas, 1996).

There are three basic attributes that characterize the models of Artificial Neural Networks: models of the processing elements (neurons), models of synaptic interconnections, and the training or learning rules for updating the connecting weights.

Processing Elements

The information processing of a PE consists of two parts: input and output. Associated with the input of a PE is an integration function *f*. The function combines information, activation, or evidence from an external source or other PEs into a *net input* to the PE. The simplest case is a linear function of the input x_j to the PE:

$$f_i = net_i = \sum_{j=1}^{m} w_{ij} x_j - \theta_i ,$$

Where θ_i is a certain threshold, w_{ij} represents the strength of the synapse connecting neuron *j* (source) to neuron *i* (destination).

One of the most commonly used functions is hyperbolic tangent (tanh) function:

$\tanh z = e^z - e^{-z}/e^z + e^{-z}$,

where e is the base of the natural algorithm.

Connections

Architecture defines the network's structure, that is not only the number of processing elements but also their interconnectivity. Each PE is connected to other PEs or to itself; both delay and lag-free connections are allowed (Lin and Lee, 1996). There are five basic types of connection geometries.

In the single-layer feedforward network, a PE is combined with other PEs to make a layer of these nodes. Inputs can be connected to these nodes with various weights, resulting in a series of outputs, one per node. Several layers can be interconnected to form multilayer feedforward network. Input layer receives inputs and typically performs no function other than buffering of the input signals. The outputs of the network are generated from the output layer. Any layer between the input and output layers is called a hidden layer because it is internal to the network and has no direct contact with the external environment. There may be no or several hidden layers in an ANN. The two mentioned types are feedforward networks because no PE output is an input to a node in the same layer or in a preceding layer.

The outputs can be directed back as inputs to same- or preceding-layer nodes, in this case, the network is a feedback network. If PE output is directed back as input to PEs in the same layer, the network is lateral feedback. Feedback networks that have closed loops are called recurrent network. A single node with feedback to itself is the simplest recurrent neural network.

In a single-layer network with a feedback connection PE output can be directed back to the PE itself, to other PEs, or to both. In a multilayer recurrent network, a PE output can be directed back to the nodes in the preceding layer. A PE output can be also directed back to the PE itself and to the other PEs in the same layer.

Learning Rules

ANNs may be broadly classified according to whether they learn in a supervised or unsupervised way (Bishop, 1995).

Supervised learning denotes a method in which some input vectors are collected and presented to the network. The output computed by the network is observed as the deviation from the expected answer that is measured. The weights are corrected according to the magnitude of the error in the way defined by the learning algorithm. This kind of learning is also called learning with a teacher, since a control process knows the correct answer for the set of selected input vectors (Rojas, 1996).

Unsupervised learning is used when the exact numerical output a network should produce is unknown. It is mostly used for classification or clustering problems.

Multilayer Perceptron

The most common type of supervised learning is the back propagation algorithm (Rumelhart et al., 1986), mostly executed with multilayer feed-forward neuronal networks or multilayer perceptron (MLP).

Backpropagation (BP) algorithm is preferred in ecological modelling, especially in water quality modelling. The architecture of the BP network is a layered feed forward neural network, in which the non-linear elements (neurons) are located in the hidden layer. The neurons feed a non-linear function by the sum of their inputs coming either from input nodes by feed forward or from output nodes by feedback. Neural networks determine the weighted connectance between the input and output nodes by these neurons (Recknagel et al., 1997; Recknagel et al., 1998; Lek et al., 1999).

Backpropagation is an algorithm based on a relatively simple concept: if the network gives the wrong answer, then the weights are corrected so that the error lessens, so future reponses of the network are more likely to be correct (Lek et al, 2000). The neurons in a backpropagation network are connected in layers, with units in layer k passing their activations only to neurons in the layer k+1. In solving a problem, activation passes from the input units, through one or more internal layers of neurons (hidden layer) and ultimately passes to the output layer and the environment. Figure 2.3 shows the scheme of the multilayer perceptron used for prediction of stream macroinvertebrate taxa from environmental variables (Huong et al., 2001).

A brief algorithm of backpropagation in neural networks is following (from Lek and Guegan, 1999):

(1) Initialize the number of hidden nodes

(2) Initialize the maximum number of iterations and the learning rate (η). Set all weights and thresholds to small random numbers. Thresholds are weights with corresponding inputs always equal to 1.

(3) For each training vector (input $X_p = (x_1, x_2, \ldots, x_n)$, output Y) repeat steps 4–7.

(4) Present the input vector to the input nodes and the output to the output node;

(5) Calculate the input to the hidden nodes: $a_j^h = \Sigma_{i=1}^{n} W_{ij}^h x_i$. Calculate the output from the hidden nodes: $x_j^h = f(a_j^h) = \dfrac{1}{1+e^{-a_j^h}}$.

Calculate the inputs to the output nodes: $a_k = \Sigma_{j=1}^{L} W_{jk} x_j^h$ and the corresponding outputs: $\hat{Y}_k = f(a_k) = \dfrac{1}{1+e^{-a_k}}$.

Notice that $k = 1$ and $\hat{Y}_k = \hat{Y}$. L is the number of hidden nodes.

(6) Calculate the error term for the output node: $\delta_k = (Y - \hat{Y}) f'(a_k)$ and for the hidden nodes: $\delta_j^h = f'(a_j^h) \Sigma_k \delta_k W_{jk}$

(7) Update weights on the output layer: $W_{jk}(t+1) = W_{jk}(t) + \eta \delta_k x_j^h$ and on the hidden layer: $W_{ij}(t+1) = W_{ij}(t) + \eta \delta_j^h x_i$

As long as the network errors are larger than a predefined threshold or the number of iterations is smaller than the maximum number of iterations envisaged, repeat steps 4–7.

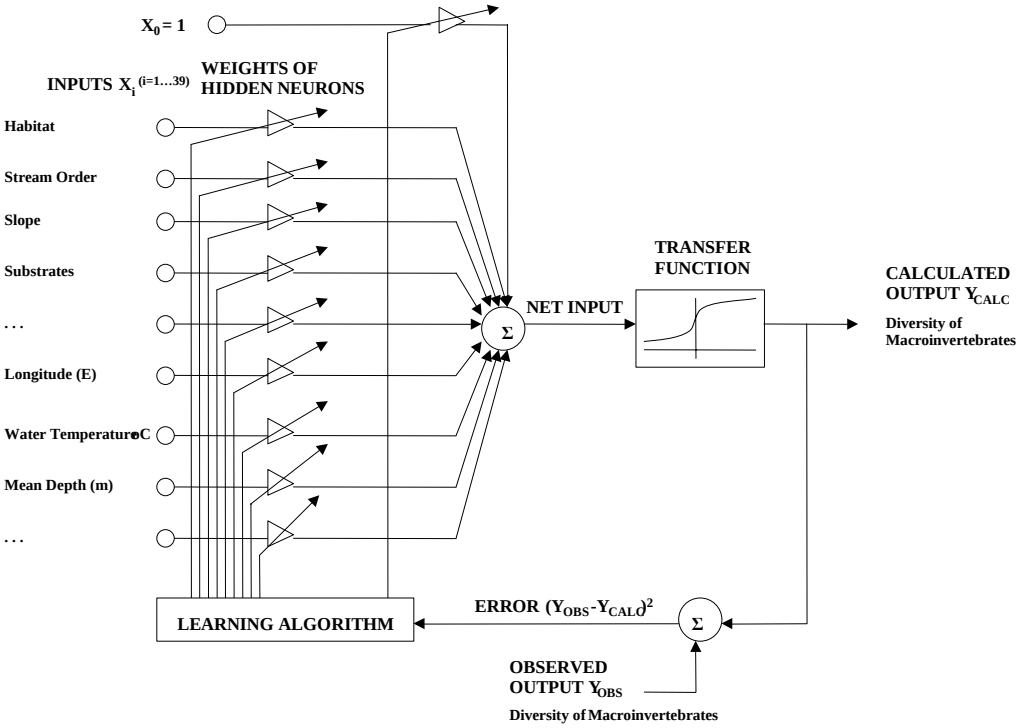


Figure 2.3. Multilayer perceptron model for the prediction of stream macroinvertebrates from environmental variables (Huong et al., 2001).

The error for a neuron in the layer directly below the output layer is a function of the errors on all the units that use its output. In general, the error for a neuron at layer n is a function of the errors of all neurons at layer *n+1* that use its outputs. In a BP network, activation moves backward in a similar fashion (Luger and Stubblefield, 1992). Once BP has computed the error for each neuron in the network, the individual units may learn by applying the delta rule, the amount of learning is represented as the difference (delta) between the desired and computed outputs.

Application of predictive ANN in freshwater ecology

In a review of computer-aided research in biodiversity, Edwards and Morse (1995) underlined that ANNs have an important potential. There have been a number of studies (Walley and Fontama, 1998; Schleiter et al., 1999; Chon et al., 2000) that showed that ANN performed better than more classical modelling methods.

Walley and Fontama (1998) firstly reported a successful application of ANN in prediction of macroinvertebrate taxa in unpolluted river sites and compared with the performance of RIVPACS. The objectives of predictions were average score per taxon (ASPT) and number of families presents (NFAM). Models were based on the standard backpropagation networks. The results showed that the ASPT model achieved a significantly higher level of performance in independent test data than the NFAM model. Results of their study demonstrated the ability of ANN in training with values of biological indices and understanding the relationship between environmental variables and biotic indices that is often a very complicated and non-linear problem. It

was concluded from study that the neural networks performed marginally better than RIVPACS. They also discussed further improvement to the performance of neural network by extending the environmental data to include relevant catchment characteristics.

Schleiter et al. (1999) went one step further to model the population dynamics of macroinvertebrates in German streams using ANN. They tested the suitability of ANN for system analysis and impact assessment: (1) in temporal dynamics of water quality; (2) in bioindication of chemical and hydromorphological properties using benthic macroinvertebrates; (3) and long-term population dynamics of aquatic insects. The satisfactory results of the study showed that ANN can meaningfully be used in the analysis of effect-relation of species, including the identification and assessment of complex impact factors, and also for forecasting system behaviour which have specific, very complex and non-linear features. However, they admitted that as ANNs learn from examples, their quality depends heavily on the representativeness and compatibility of the database.

Chon et al. (2000a) applied Artificial Neural network to classify and predict multivariate stream data even in a short period using benthic communities. This study demonstrated that temporal ANNs could be utilized to forecast and analyze short-period changes in multivariate data sets. The recurrent neural network appeared to be effective in patterning development of benthic communities in streams responding in a diverse manner to a wide range of pollution. The study also showed the advantage of specific forecasting for an individual taxon is that it could assist to characterise community changes.

Pudmenzky et al. (1998) developed preliminary ANN models for predicting the distribution of macroinvertebrates in the Queensland stream system based on environmental variables. The network was trained with both categorical and continuous attribute input data. The ANN proved promising in predicting the taxa, which had the most even equal distribution of presence/absence (probability of occurrence around 0.5). As work had been done with a shareware version of the software package, only a subset of the data could be investigated. However, this is the first work done in applying ANN to biological assessment of habitat condition in Australia. Further research is highly recommended to investigate the possibility of ANN as computational alternative to AusRivAS in supporting bioassessment of habitat condition.

Maier and Dandy (1996) used ANN as a viable means of forecasting salinity in the River Murray (South Australia) 14 days in advance. The results obtained had less than 7% average absolute percentage error. It was concluded that, ANN models appear to be useful tool for forecasting salinity in rivers.

Recknagel et al. (1997) applied ANNs to the task of modelling and prediction of algal blooms and to identification of the variables that play a major role in algal growth. In their study, major ecological factors of all chemical physical and biological categories, which could clearly define the environmental conditions of the aquatic system, were included as input variables and five dominating phytoplankton species were used as output variables. The resulting predictions on succession indicate the ability of ANNs to fit the complexity and non-linearity of complicated ecological phenomena. If an expanded database is available, not only a specific aim can be

investigated but also cost-benefit strategies for management can be addressed applying ANN to scenario and sensitivity analysis.

ANN had been applied very successfully to eutrophication processes. Research has been done in Italy (Scardi, 1996), Japan (Yabukana et al., 1997), and Turkey (Karul et al., 2000). Models used physical and chemical parameters and also biological variables as inputs to predict the behaviour of chlorophyll – a and other typical eutrophication indicator. The studies showed that nonlinear relationships in the eutrophication phenomenon could be modeled reasonably well. The ANN model can also estimate an extreme value that lies outside the boundaries of the training set. Conclusions were made that ANN models can be used to estimate the densities of certain species as functions of environmental parameters.

Wen and Lee (1998) applied ANN to the problem of optimising water quality management in a river basin. Their study focused on the objectives of environmental quality, treatment cost of wastewater and the assimilative capacity of a river to provide a solution to water quality management problems. The results of their work show that using the backpropagation algorithm and feed forward neural network, a multi objective programming model can simulate the decision makers' preferences and successfully overcome the disadvantages of unknown preferences of decision makers.

Recknagel and Wilson (2000) discussed the potential of ANN models in working with aquatic ecosystems. They compared presentations of 6 prototypes of inductive and deductive models for phytoplankton including a regression model; time series model; deterministic models for functional algal group succession and algal population; heuristic model; and ANN. The result of comparisons showed that only ANN provides an ability to predict both timing and magnitudes of species dynamics and species succession in the lake. ANN models can support both prediction and elucidation of ecosystem behavior with the potential to provide new insight into mechanisms of systems from the results.

Maier et al. (1998) used ANNs for modelling the incidence of cyanobacteria in rivers by forecasting the occurrence of a species group of Anabaena in the River Murray, Australia. ANNs provided a good forecast of both the incidence and magnitude of a growth peak of cyanobacteria within the limit required for water quality monitoring. The models also defined predominant variables in determining the onset and duration of cyanobacteria growth.

Lek-Ang et al. (1999) developed predictive modelling of Collembolan diversity and abundance on a riparian microhabitat scale. Biological variables that were retained to describe its structure in this model included abundance of dominant species, species richness and biological indices. In the input layer, the main environmental variables were considered. 80% samples were chosen randomly for the training process and the remaining 20% were used for model validation. The resulting habitat profiles illustrated the complex influence of each variable on the biological parameters of the assemblage and also the non-linear relationship between dependent and independent variables. The study gave satisfactory results over practically the whole range of values of dependent variables, which showed ANNs potential to predict biodiversity and structural characteristics of species assemblages.

Gozlan et al (1999) applied ANN with the aim to predict the abundance of six fish species in the river Garonne with back propagation as learning algorithm. The ANN was successful in predicting the abundance of 0+ fishes on a microhabitat scale, indicating that technique merits more frequent use in ecology and biodiversity studies. The explanatory part of the analysis, coupled with the predictive power of ANN, should facilitate the ecologically oriented management of aquatic ecosystems, providing that the duration of the study is extended.

Interpreting variable importance and Sensitivity analysis

Although a number of studies shown ANNs having higher predictive power compared to traditional statistical approaches, neural networks often called 'black boxes' because it is difficult to extract explanation of the relative influence of the independent variables in the prediction process. Fortunately, recent studies have provided a variety of methods for quantifying and interpreting the contributions of the variables in the neural networks (Olden and Jackson, 2002; Gevrey et al., 2003; Olden et al., 2004).

A variety of methods are available for the estimation of the contribution of predictor variables in relationship to the output. For example, Partial Derivative method (PaD) provides a profile of the output variations for small changes of each input variable and classification of the relative contributions of each variable to the network output. 'Stepwise' method is based on the classical stepwise approach that consists of adding or rejecting step by step one input variable and noting the effect on the output results (Gevrey et al., 2003). 'Profile' method proposed by Lek (Lek et al., 1995) studies each input variable successively when the others are blocked at fixed values. In the neural network, the connection weights between neurons are the linkages between the input and the output of the network, and therefore are the link between the problem and the solution (Olden and Jackson, 2002). Garson algorithm or 'Weights' method includes partitioning the connection weights to determine the relative importance of the various inputs.

Even though a variety of methods are available in order to make 'black-box' ANN more transparent, the 'Profile' method is the most relevant for the ecological applications as it is the only technique that provides two elements of information on the contribution of the variables (Dedecker et al. in press). On the other hand this method presented the order of contribution of the different environmental variables, on the other hand gave direct interpretation of the effect of river characteristics on the abundance or occurrence of particular taxa by plotting the simulated output against changes in the particular variable. The other methods are merely able to classify the variables by order of their importance, in other words, to reveal their contribution to the output. In spite of their different ways of computation, difference in sensitivity and stability of the methods were rather small (Gevrey et al., 2003; Dedecker et al., in press).

Scenario analysis

Scenario planning is a systematic method for thinking creatively about possible complex and uncertain futures (Peterson et al., 2003). The central idea of scenario planning is to consider a variety of possible futures that include many of the important uncertainties in the system rather than to focus on the accurate prediction of single outcome. In many ecological situations uncertainty is substantial and irreducible and arises from the problems of ecological predictions. Scenarios describe futures that could be rather than futures that will be. A set of scenarios should usefully expand and challenge current thinking about the system (Peterson et al., 2003).

'Dirty-water' models, which include variables potentially affected by the human activity can be used for the scenario analysis. Accuracy and flexibility of the neural networks makes them extremely attractive tool for the simulation of the potential futures. I am aware of only few studies involving scenario analysis with ANNs in the relationship to freshwater ecology.

Poff et al. (1996) used an artificial neural network to evaluate the hydrological responses of two streams in the northeastern US having different hydroclimatologies to hypothetical changes in precipitation and thermal regimes associated with climate change. Four scenarios of climate change were used to evaluate stream response to climate change: +25% precipitation, -25% precipitation, 2x the coefficient of variation in precipitation regime and +3C temperature. Responses were expressed in hydrological terms of ecological relevance, including flow variability, baseflow conditions, and frequency and predictability of floods. ANN were used to generate synthetic daily hydrograph with high goodness of fit (r^2 >0.8). Increased average precipitation induced elevated runoff and more frequent high flow events, while decreasing precipitation had the opposite effect. Elevated temperatures reduced average runoff. In general, the rainfall-dominated stream exhibited greater relative response to climate change scenarios than did the snowmelt stream. The fact that ANN does not rely on mechanistic response can be however viewed as advantage because it does not require that assumptions be made about specific indirect effects that may result from scenarios of climate change. Because this technique is not mechanistic it does not require extensive parameter estimation – a modeling process that may result in large propagation of errors.

The limitation of the ANN models (in comparison with process-based or deterministic models) is that their expertise is limited by the data available for training, thus simulation of scenarios using data range beyond the models' range as extreme events can be difficult. However, it is possible to add virtual datasets based on the expert opinion in order to be able to simulate response of the biota to 'extreme' events. For example, Dedecker et al. (in press) built and trained MLP model using presence absence data, collected over 2 years in Zwalm river basin in Flanders, Belgium. Fifteen structural, physical and chemical environmental variables were used as predictors. Model training methodologies were elaborated to generate appropriate models to simulate 'extreme' scenarios concerning flow control and water quality management. A virtual data set based on ecological expert knowledge was created to introduce 'extreme' value to the model. The obtained results indicated that the presence/absence of Asellidae in the 'extreme' validation set was predicted significantly better when the number of extreme examples in the training set increased. However, the overall predictive power of the ANN models decreased when

a relatively large virtual data set were applied. Three case studies have shown that ANN models are in general quite robust with a rather high predictive reliability. The reliability of the models has to be assessed via simulations made by ecological experts who can deliver knowledge that is often not included in the database used for the model induction.

2.2.2 Ordination and clustering of stream conditions

People wish to know how human activity influences the fascinating diversity of biological communities. Yet this very diversity creates problems for the statistical analysis of ecological observations: it implies a large number of species and a large inherent variability. A set of community samples and associated environmental measurements typically yields an enormous amount of noisy data, which is difficult to interpret (terBraak and Verdonschot, 1995). In order to find a solutions when dealing with this kind of data, ecologists have employed a number of methods of multivariate analysis including, clustering and ordination (Gauch, 1989) and newer techniques such as artificial neural networks (Legendre and Legendre, 1998).

Understanding patterns in communities within ecosystem is a first necessary step towards effective management of the ecosystem. Development of methods for patterning spatial and temporal changes in biological communities has been an important issue in ecosystem management. Traditionally, a variety of multivariate statistical methods have been used for patterning biological communities. In the context of this chapter we will consider the most commonly used statistical methods such as Principle Component Analysis and Canonical Correspondence Analysis and neural based Self Organising Maps.

2.2.2.1 Multivariate statistical approach

Data for the complete assemblages involving the abundances of all component species can contain as many dimensions to these data as there are species of sites (Gauch, 1989). Such a large number of dimensions are almost impossible for the human mind to comprehend. A reduction of dimensionality sometimes is necessary in order to conduct further data analysis and modeling. Unconstrained ordination is a tool commonly employed to examine data structure and to reduce the dimensionality. Principle Component Analysis (PCA) and Correspondence Analysis (CA) are both commonly used eigen based ordination methods.

> *A non-zero vector C is called an eigenvector of a square matrix A if and only if there exists a number (real or complex) λ such that $AC = \lambda C$. If such a number λ exists, it is called an eigenvalue of A (http://www.sosmath.com).*

Both PCA and CA perform an eigen analysis on a matrix of distances, with distance being the spatial distance between the objects (sites) in ordination space. PCA the oldest ordination method, uses Euclidean distance in the analysis, while CA uses Chi-

Squared distance (Legendre and Legendre, 1998). As a result the biplot of an ordination of either PCA or CA preserves the respective distance among the sites or species.

The arch and horseshoe effects relate to a commonly seen property in biplots created from PCA and CA. They refer to the representation of sites along an arch or horseshoe shape, which misconstrues the distance among sites. These effects result because each axis in the ordination may not be independent from that preceding it. That is, the axes are correlated in some way (Legendre and Legendre, 1998). This can lead to incorrect interpretation of the biplot and should be kept in mind when using these techniques.

An alternative to eigen analysis based approaches is Non-Metric Multidimensional Scaling (NMDS). Where PCA and CA perform the eigen analyses on a matrix of distance in ordination space, NMDS performs the analyses directly on a matrix of ecological dissimilarity (Gauch, 1989). Considering the rank order of dissimilarities among variables (species) avoids the assumption of linearity and replaces it with the less constraining assumption of monotonicity, where the relationship is assumed to either increase or decrease with no consideration of the nature of this relationship (i.e. there is no constraint on whether this relationship is linear, logarithmic or exponential)(Digby and Kemptop, 1987; Gauch, 1989).

In the methods considered so far, the distribution of biota is not directly related to the environmental conditions, rather, the gradients in distribution of biota can be later related to environmental variables (Okland, 1996). Recently, there has been an increasing trend to move from using multivariate techniques only for pattern recognition to identifying relationships between the assemblage and its environment (De'ath, 2002).

Canonical Correspondence Analysis (CCA) is a constrained ordination method that combines the multivariate setting of CA and multiple regression. CCA maximises the fit of sites and /or species to a set of environmental (explanatory) variables.

Each species occurrence is confined to a limited range, its niche. Species tend to separate their niches, partly so to minimise the competition. If the separation is strong, successive species replacements occur along the environmental gradient. The composition of biotic communities thus changes along the environmental gradient according to unimodal function. Some species may prefer extreme environmental conditions or their optima may fall outside the environmental region actually sampled in a particular study so that their observed response function is not unimodal but monotonic decreasing or increasing. CCA extracts the "best" synthetic gradients from field data on biological communities and environmental features: it forms a linear combination of environmental variables that maximally separates the niches of the species. Niche separation is expressed as weighted variance of species centroids on a standardized gradient, the species centroids being the weighted average of gradient values of the sites at which the species occurs. The first synthetic gradient is termed the first ordination axis. The achieved maximum amount of niche separation is given by the eigenvalue of the ordination axis. Subsequent ordination axes are also linear combinations of the environmental variables that maximally separate the niches, but subject to the constraint that there are uncorrelated with the axis or axes extracted previously (terBraak and Verdonshot, 1995).

The primary result of a CCA is an ordination diagram, i.e. a graph with a coordinate system formed by ordination axes. The coordinates of the site points are the values (termed scores) of the sites on the two best synthetic gradients. Each species point in the diagram is at the centroid (weighted average) of the site points in which it occurs. The species points thus indicate the relative location of the two-dimensional niches of the species in the ordination diagram. The arrow for a quantitative variable runs from the origin (center) of the diagram to an arrowhead, the coordinates of which are the correlations of the variable with the axes. The arrow points in the direction of maximum change in variable and the arrow length is proportional to the maximum rate of change. In the perpendicular direction the variable does not change in value. Informally, the length of an environmental arrow indicates the importance of the variable (terBraak and Verdonshot, 1995).

Each eigenvalue of CCA can be converted to a percentage variance accounted for by dividing the eigenvalue (x100) by the total inertia of the abundance data, inertial being a measure of weighted variance that is closely related to the chi-square statistic. For ecological data the percentage explained inertia is typically low (<10%). This is nothing to worry about, it is an inherent feature of data with a strong presence-absence aspect (terBraak and Verdonshot, 1995).

Decision criteria include the magnitude of the eigenvalues themselves (as a rule of thumb, eigenvalues >0.3 indicate strong gradients), the statistical significance as judged by Monte Carlo permutation tests and, even more importantly the ecological interpretability (terBraak and Verdonshot, 1995).

Partial CCA

Often it may be known that certain environmental factors dominate the composition of a species assemblage. In this case it may be of interest to consider the effect of other environmental variables having removed the effect of this overarching environmental factor or factors (terBraak and Verdonshot, 1995).This is done through partial CCA. Partial CCA removes the effect of one set of explanatory variables, covariables, before conducting standard CCA on the residual variation using another set of explanatory variables (Legendre and Legendre, 1998). It is an extremely useful tool, which allows the user to identify important environmental variables that may be hidden by dominant environmental gradients. Further, it has been adapted by ecologists to identify spatial and temporal dependencies that exist in a multivariate dataset (Borcard et al., 1992).

Partial CCA has been widely used throughout the ecological literature since its introduction across a wide range of taxa including benthic macroinvertebrates in the littoral zone of lakes (Johnson and Goedkoop, 2002), the macroibenthos in estuaries (Ysebaert et al., 2003), as well as freshwater fish (Godinho et al., 2000).

2.2.2.2 Artificial neural networks with unsupervised learning: Self Organising Maps

Even though, very useful and widely applied for a number of ecological studies, statistical methods as PCA and CCA still capable of describing linear relationships

only and require the data meeting the assumption of the normal distribution. The models based on unsupervised learning methodology, particularly Kohonen Self-Organising Maps (Kohonen, 1995) recently have become widely used tool for clustering multivariate biological data as they allow to overcome limitations of statistical methods and allow the analysis of data containing complex non-linear relationships (Lek et al., 2000).

SOM identifies patterns in data, clusters them into a predefined number of classes, and orders the classes in a two-dimensional output space such that near neighbors in data space are near neighbors in output space. Clustering and ordering are integrated into one process using a similarity metric based on Euclidean distances and a neighborhood function which ensures that near neighbors in the output space represent similar patterns (Walley and O'Connor, 2001).

This method allows to group objects together on the basis of their perceived closeness in n-dimensional hyperspace. The Kohonen network includes two types of units: an input layer and an output layer (Figure 2.4). The input unit is simply a flow-through layer for the input vectors (Lek et al., 2000). In the output layer the model typically consists of a two-dimensional network of neurons arranged on a grid laid out in a lattice. Lattice could be

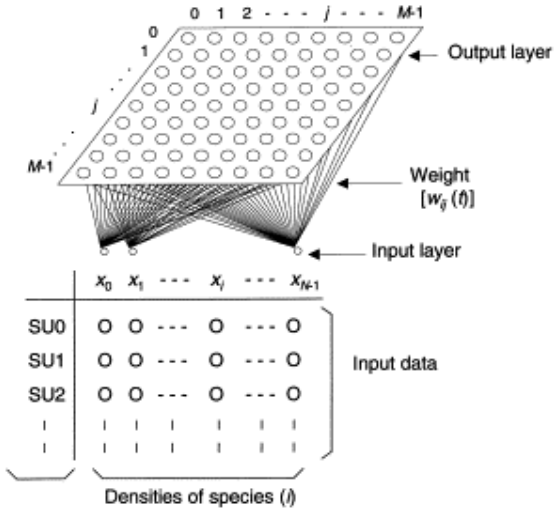


Figure 2.4. Structure of the Kohonen's network (Chon et al., 1996).

different geometrical form but hexagonal is usually preferred as it does not favor horizontal or vertical directions. Each neuron is connected to its nearest neighbors on the grid. The neurons store sets of weights. For an input *x*, each neuron *j* (weight: w_j) calculates its activation level, defined as:

$$||w_j - x|| = \sqrt{(w_{ij}-x_i)^2}$$

which is the Euclidean distance between the points represented by the weight vector and the input in n-dimensional space. A node whose weight vector closely matches the input vector will have a small activation level, and a node whose weight vector is

very different from the input vector will have a large activation level. The node in the network with the smallest activation level is deemed to be the winner for the current input vector.

The winning node and some the node around it then allowed adjusting their weight vectors to match the current input vector more closely. The units allowed to adjust their weights, are called neighborhood of the winner. The size of the neighborhood is decreased linearly after each training epoch, until it includes only the winner itself. Input patterns, which allow the same node to win are then judged to be in the same cell on the final map output (Lek et al., 2000).

A brief algorithm of SOM is following (from Lek and Guegan, 1999):

Initialise the time parameter t: $t = 0$.

(1) Initialise weights W_{ij} of each neuron j in the Kohonen map to random values (for example, random observations).

(2) Present a training sample $x(t) = [x_1(t),\ldots,x_n(t)]$ randomly selected from the observations.

(3) Compute the distances d_j between x and all mapping array neurons j according to: $d_j(t) = \Sigma_{i=1}^{n}[x_i(t) - W_{ij}(t)]^2$ where $x_i(t)$ is the i^{th} component of the N dimensional input vector and $W_{ij}(t)$ is the connection strength between input neuron i and map array neuron j at time t expressed as a Euclidean distance.

(4) Choose the mapping array neuron j^* with minimal distance d_{j^*}: $d_{j^*}(t) = \min[d_j(t)]$.

(5) Update all weights, restricted to the actual topological neighbourhood $NE_{j^*}(t)$: $W_{ij}(t+1) = W_{ij}(t) + \eta(t)(x_i(t) - W_{ij}(t))$ for $j \in NE_{j^*}(t)$ and $1 \leq i \leq n$. Here $NE_{j^*}(t)$ is a decreasing function of time, as is the gain parameter $\eta(t)$.

(6) Increase the time parameter t

(7) If $t < t_{\max}$ return to step 2

Aside from convenient visualisation and ability to deal with non-linearities SOM is attractive for biologists by its way of working with missing data and outliers. Statistical methods like PCA are particularly sensitive to these two problems and input data often has to be pruned before processing. SOM discards data with too many missing components during the training process and then maps it on the finished map. When outliers are present, SOM positions each in its place in one unit of the map, and only the weights of that neuron and its nearest neighbors are affected. There is no effect on the other neurons and outliers are easily detected in by observation of scattered data in an area of the map (Lek et al., 2000).

The analysis using visualization of component planes is comparable to PCA, but more directly describes the discriminatory power of the input variables in the mapping procedure (Kohonen, 2001). A clear distribution gradient of a variable represents a high contribution to the classification of input vectors. When there are strong relationships between input and output variables, the component planes show clear gradients and similar patterns of their distribution on the trained SOM map. According to the distribution gradients of the environmental variables on the SOM map, influence of environmental variables on the classification of the sampling sites as well on biotic variables as diversity indices could be assessed effectively (Park et al., 2003).

Applications of SOM in freshwater ecology

In the last couple of years SOMs were extensively used for finding patterns and classification of ecological communities. SOMs were successfully applied to the genetic analysis of French trout population (Giraudel et al., 2000) in order to separate wild trout from domestic (born in hatcheries). Kohonen's map with Fuzzy Clustering Algorithm showed superior results in comparison with ANN based on supervised learning.

Walley and O'Connor (2001) have built a river pollution diagnostic system based on non-neural algorithm inspired by SOM, which is presently being tested by Environmental Protection Agency. It based on existing information on occurrence and abundance of macroinvertebrate taxa in England and Wales organised in 250 color coded clusters. The clusters have been arranged to form a hexagonal output 'map' such that clusters that are close together represent similar sets of river samples.

The user could see all the available information on cluster by clicking on it. The report provides information on classes, stress types, site characteristics, water chemistry averages, etc for the samples in the cluster. A list of reference numbers for the samples within the cluster is given at the bottom of the screen. Information on new sites could be loaded from file or manually, each new site will be allocated to the existing cluster. The system is also linked to GIS, which allows users easily to see the location of the site considered.

Spatial analysis of stream invertebrate distribution in the drainage basin had been studied (Cereghino et al., 2000). The study provided a stream classification based on characteristic EPTC (Ephemeroptera, Plecoptera, Trichoptera, Coleoptera*)* insect assemblages at species level. The main interest of their results is that the stability of these theoretical assemblages may be used to refine representative and/or reference sites for biological surveillance, as a change in species composition within a given region can be considered as a biological indicator of environmental changes.

Obach et al. (2001) used SOM for visualisation of data, outliers detection, hypothesis generation and detection basic pattern in annual abundance of aquatic insects in small stream in central Germany. Species groups with similar ecological requirements were distinguished. Furthermore, they applied radial basic function neural networks combined with a SOM (RBFSOM) and successfully predicted the annual abundance of selected species from environmental variables.

Park et al. (2001) applied SOM for patterning and prediction of exergy (single measurement to express the information level of communities) of benthic macroinvertebrate communities in streams. The trained mapping was able to characterize the development trend of exergy at different groups of sample sites in different time period. Park et al (2003) implemented SOM as a part of counterpropagation neural network (CPN) in order to pattern aquatic macroinvertebrate communities. Sampling sites were classified into 5 groups relatively the pollution status and habitat type using 34 environmental variables, Specie Richness of macroinvertebrates and Shannon diversity. They also used component planes to visualize environmental variables and diversity indices and evaluate the relationships between variables.

Chon et al. (2001) used SOM in combination with Adaptive Resonance Theory network (ART) for analysis of patterns of temporal variation in community dynamics of benthic macroinvertebrates collected in the Suyong River in Korea. The sampled data for each month was initially trained by ART, the weights of which preserved conformational characteristics among communities during the process of the training. Subsequently these weights were rearranged sequentially from 2 to 5 months, and were provided as input to the Kohonen network to reveal temporal variations in communities. The network was then able to extract the features of community dynamics in a reduced dimension covering the specified input period. Authors concluded that neural networks can be successfully used for comprehensive understanding of data features in community dynamics in the spatio-temporal domain.

Some other machine learning methods

Genetic algorithm (GA)

Genetic algorithms are a part of evolutionary computing, which is a rapidly growing area of artificial intelligence. The Genetic Algorithm is a model of machine learning, which derives its behavior from a metaphor of the processes of evolution in nature. This is done by the creation within a machine algorithm of a population of individuals represented by chromosomes, in essence a set of character strings that are analogous to the base-4 chromosomes that we see in our own DNA. The individuals in the population then go through a process of evolution.

First, pairs of individuals of the current population are selected to mate with each other to form an offspring, which then form the next generation. Selection is based on the survival-of-the-fittest strategy, but the key idea is to select the better individuals of the populations, as in tournament selection where the participants compete with each other to remain in the population (Jain and Martin, 1999). The most commonly used strategy to select pairs of individuals is the method of roulette-wheel selection, in which every string is assigned a slot in simulated wheel sized in proportion to the string’s relative fitness. This ensures that highly fit strings have a greater probability to be selected to form the next generation through crossover and mutation. After selection of the pairs of parent strings, the crossover operator which involves the swapping of genetic material between the pairs is applied to each of these pairs. The two individuals (children) resulting from each crossover will be subjected to the mutation operator in the final step to forming the new generation. The mutation operator alters one or more bit values at randomly selected locations in randomly selected strings. Mutation takes place with a certain probability, which, in accordance

with its biological equivalent, typically occurs with a very low probability. The mutation operator enhances the ability of GA to find a near optimal solution to a given problem by maintaining a sufficient level of genetic variety in the population, which is needed to make sure that the entire solution space is used in the search for the best solution (Jain and Martin, 1999).

Idea of evolutionary computing was introduced in the 1960s by I. Rechenberg (cited by Jain and Martin, 1999) in his work "Evolution strategies" (Evolutionsstrategie in original). His idea was then developed by other researchers. Genetic Algorithms (GAs) were invented by John Holland and developed by him and his students and colleagues. This lead to Holland's book "Adaptation in Natural and Artificial Systems" published in 1975 (Jain and Martin, 1999).

In 1992 John Koza has used genetic algorithm to evolve programs to perform certain tasks. He called his method "genetic programming" (GP). LISP programs were used, because programs in this language can be expressed in the form of a "parse tree", which is the object the GA works on (Jain and Martin, 1999).

Genetic algorithms are used for a number of different application areas. They are most appropriate for optimization type problems and have been applied successfully in a number of automation applications including job shop scheduling, proportional integral derivative (PID) control loops, and the automated design of fuzzy logic controllers and ANNs.

As a relatively new application GA was used for ecological research in very few cases. Bobbin and Recknagel (2001) applied it to the construction of rule-based model for the prediction and explanation of algal blooms in the Japanese Lake Kasumigaura. Different models have been evolved for two common groups of blue-green algae. The models show that there is a difference in the environmental preferences of the two groups, and that this difference could be learned. Learned patterns are represented explicitly as classification rules, which allow their underlying hypothesis to be examined.

Whigham (2000) has applied genetic programming for induction of spatial models for the prediction of habitat types and population distribution of Australian greater glider.

d'Heygere et al. (2001) applied evolutionary algorithms to select input variable combinations of classification tree models predicting benthic macroinvertebrate communities in watercourses of Flanders (Belgium). Different sets of input variables result in different CCIs and a manual selection of the most convenient model based on trial and error is very labour intensive. Selection of variables by genetic algorithm eased the choice of input variable combinations of the classification tree models drastically.

Both ANNs and evolutionary algorithm are novel approaches and there has not been many comparisons of their performance and applicability to particular ecological problems. Recknagel et al. (2002) compared potentials and achievements or artificial neural networks and genetic algorithms in terms of forecasting and understanding of algal blooms in Lake Kasumigaura (Japan). Examples presented in this paper showed that models explicitly synthesized by genetic algorithms not only perform better in

seven-day-ahead prediction of algal blooms than artificial neural networks models, but provide more transparency for explanation as well.

2.3 Comparison of ANN with other methods

Maier and Dandy (1996) compared ANNs to statistical ARMA (Auto Regressive Moving Average) class of models widely used for modeling water resources time series in terms of advantages and disadvantages. They found that ANNs are more flexible in working with complex non-linear system and in providing long term forecasting. Similar comparisons between ANNs and other classes of statistical modeling provided by Lek et al. (1996) and Paruelo and Tomasel (1997) also emphasized the flexibility of ANNs.

Ball et al. (2000) compared performance of ANNs with a range of statistical methods like Linear Regression, Multiple Regression, Principle Component Analysis and Combination of PCA with Least Squares Regression (LSR). It was shown that ANN models produced the best performance. The conventional statistical techniques produced a poor performance when modeling the data and were unable to produce accurate prediction on unseen data. The best non-ANN method of modeling was achieved by combining PCA with LSR. Jeong and Joo (submitted) have also compared Multiple Linear Regression and MLP for the prediction of phytoplankton dynamic in a regulated Nakdong River. MLP model has shown higher time-series predictability than Multiple Linear Regression model.

Brosse et al. (2001) made a comparison between PCA (Principle Component Analysis) and Kohonen networks capabilities to analyze the spatial occupancy of several European freshwater fish species in the littoral zone of a large French lake. Both methods provided insight on the major trends in fish spatial occupancy. However, a more detailed analysis showed that only SOM was able to reliably visualize the entire fish assemblage in two-dimensional space, when PCA provided irrelevant ecological information for some species. These drawbacks were afforded to data heterogeneity, scarce species being poorly represented on the PCA plane. The author concluded that SOM constitute a more reliable data representation method than PCA when complex ecological data sets are used.

Giraudel and Lek (2001) compared performance of SOM with that of some statistical methods like Polar Ordination, Principle Component Analysis, Correspondence analysis and Non-metric multidimensional scaling. It has been pointed that comparison of SOM and statistical method is not easy when non-linear algebra is involved in the computation, formal proofs are almost impossible and an experimental approach is the only way. SOM doesn't have problems like horseshoe effect (PCA) or arc effect (CoA), although it is not possible to control the direction of the gradients with SOM. As advantage, new samples can be added on the Kohonen's map without affecting the ordination. For each statistical ordination method, the distance or the similarity distance have to be chosen, unlike almost all conventional methods, SOM allows a large choice, requiring an adaptation of the learning equation. Authors concluded that SOM seems fully usable in ecology, it can perfectly complete classical techniques for exploring data and for achieving community ordination. SOMs provide

a visual way to find structures in ecological communities and can be recommended to be used in an exploratory approach in which unexpected structures might be found.

Chon et al. (1996) applied SOM to clustering and patternizing benthic macroinvertebrates collected in the Korean river Suyong. The grouping resulting from learning by the Kohonen network was comparable to the classification by conventional clustering methods. Through patternizing, the network showed a possibility of producing easily comprehensible low dimensional maps under the total configuration of community groups in a target ecosystem. Changes in spatio-temporal community patterns may also be traced through the recognition process.

Paruelo and Tomasel (1997) compared results of prediction of functional characteristics of ecosystems by ANN and regression models. They tested the predictive power of ANNs and RMs using simulated data for six functional traits derived from the seasonal course of the normalized difference vegetation index (NDVI). For the six traits analyzed, the ANNs were able to make better predictions than RMs. The correlation between observed and predicted values of each of the six traits considered was higher for ANNs than for PMs. ANNs showed clear advantages to capture inertial effects. The ANN used was able to use previous year information on climate to estimate current year NDVI much better than the RM that used the same input information.

Huong et al. (2001) compared performance of multi-layered perceptron models for 37 macroinvertebrate taxa based on 896 stream data sets from the Queensland stream system with that of AusRivAs. The ANN model validation by means of 167 independent data sets revealed 73% as lowest rate and 82% as average rate of correct ANN predictions of stream site habitats. The increase of correct predictions was 30%, if ANNs and the statistical stream model AusRivAS were compared based on the same data sets.

Olden and Jackson (2002) provided a comprehensive comparison of traditional and alternative techniques for predicting species distributions using logistic regression analysis, linear discriminant analysis, classification trees and artificial neural networks to model: 1) the presence/absence of 27 fish species as a function of habitat conditions in 286 temperate lakes located in south-central Ontario, Canada and 2) simulated data sets exhibiting deterministic, linear and non-linear species response curves. On average, neural networks outperformed the other modelling approaches, although all approaches predicted species presence/absence with moderate to excellent success. When simulated non-linear data was used classification trees and neural networks greatly outperformed traditional approaches, whereas all approaches exhibited similar correct classification rates when modelling simulated linear data.

The only case where ANN did not outperformed statistical models was reported by Manel et al. (1999). The authors assessed the occurrence of a common river bird, the Plumbeous Redstart *Rhyacornis fuliginosus*, along 180 independent streams in the Indian and Nepali Himalaya. They compared the performance of multiple discriminant analysis (MDA), logistic regression (LR) and artificial neural network (ANN) in predicting this species presence or absence from 32 environmental variables. Model performance was assessed from two methods of data partitioning. In 'leave-one-out' approach, LR correctly predicted more cases (82%) than MDA (73%)

or ANN (69%). It was concluded that ANN does not yet have major advantages over conventional multivariate methods for assessing bird distribution.

The number of studies which reported better performance of ANNs in comparison with traditional statistical methods is surprising taking into consideration that performance of neural network could be confounded by a number of factors. ANN models are highly dependent on quality of data used for learning. Choice of the model configuration, transfer function, number of processing element, connections and learning rules can also significantly affect its performance. ANNs have shown a great potential to become a highly reliable modelling tool for the ecological and environmental research, however, in order to be able to use ANNs' full capability ecologists need to understand the main principles of their organisation and performance.

2.4 Aims and hypotheses of the thesis

This study was aimed to address the main questions faced by practicing environmental biologists when attempting to analyze, assess and predict conditions of freshwater biota in Australia. This thesis does not go deep into many technical issues relating to the architecture and optimization of ANNs, but instead addresses the potential and applicability of various ANNs to the practical ecological problems as natural variability versus changes due to the anthropogenic impact, assessment and prediction of the changes in stream biota in response to changes in environmental variables, trade-off between complexity and accuracy of the predictive model, etc.

I intentionally utilized the most commonly used and easily understandable types of the neural networks which performances have widely been estimated, namely Multy-Layered Perceptron and Self Organising Maps in order for my research to be accessible and easy understandable for not only ecological modelers but for as many practicing ecologist as possible. When designing my research I aimed for the ecological relevance and practicality first, and then for the technical accuracy and robustness. My primary goal was to illustrate the applicability and flexibility of ANN using the widest variety of ecological problems practically possible for me to consider in the duration of my candidature. In particular I addressed the following questions:

- Huong et al. (2002) studied the applicability of ANN models for Queensland Streams. Are these models based on the referential approach applicable to the other geographical regions of Australia? Does the accuracy of the predictions depend on the rarity of taxa?

- Is it possible to directly predict taxonomic richness of freshwater biota using ‘dirty-water’ models?

- What is the potential of SOM to provide an explanation into the natural variability in macroinvertebrate communities?

- What is the potential of SOM for the understanding of the relationship between abiotic and biotic variables?

- What is the potential of combining two different types of ANNs as MLP and SOM: is it possible to predict SOM defined clusters using MLP?

- What is the potential of combining SOM with statistical techniques as CCA: is it possible to provide an additional insight and more convenient visualization of the relationships between trophic structure of macroinvertebrate communities and water quality variables using combination of these methods?

- What is the potential of application of SOM to the outputs of predictive models in order to provide clearer understanding of resulting patterns?

- How predictable and stable the results of sensitivity analysis when applied to the real ecological data?

- What is the potential of sensitivity analysis for the determination of sensitivity of macroinvertebrate taxa to the salinity of water?

- Can predictive ANNs be used for the simulation of the effect of secondary salinisation on macroinvertebrate communities using the real data?

- How does the accuracy of the model relate to the number of predictor variables?

- Which predictive models are more accurate: generic or local?

- To what extent does the temporal variability affect the accuracy of predictive models?

- To what extent the variability between habitats affects the accuracy of predictive models?

Chapter 3

Material and Methods

This chapter provides a general overview of the data and methods, however, specific data sets and modelling approach used for each case study are described in each relevant chapter.

3.1 Data

Queensland (QLD) data

The main dataset used for the current study was provided by the Queensland Department of Natural Resources and Mines (NR&M). The dataset contains information about habitat characteristics of 896 reference and 1159 test sites, 5416 samples in total (Figure 3.1). Reference sites are those considered to be in a near pristine conditions (see Conrick and Cockayne, 2000). The data was collected in spring and autumn from 1994 to 2002 at 5 different habitats: riffle, edge, rocky pools, sandy bottom and macrophytes. I used riffle and edge habitats subsets for most of the case studies. Riffle subset contains 1333 samples and edge subset 2442 samples. Different sets of variables describing geoclimatic features and water quality were used for different case studies with maximal number of 50 variables (Table 3.1).

The sites were sampled according to standard AusRivAs rapid assessment protocols (Conrick and Cockayne, 2000) and processed by experienced NR&M staff. This involves sampling 10 m of habitat using a standard 250 μm dip net with live picking in the field and identification of samples in the laboratory. The dataset contains occurrence patterns (presence-absence) of 168 macroinvertebrate taxa mostly identified to family level.

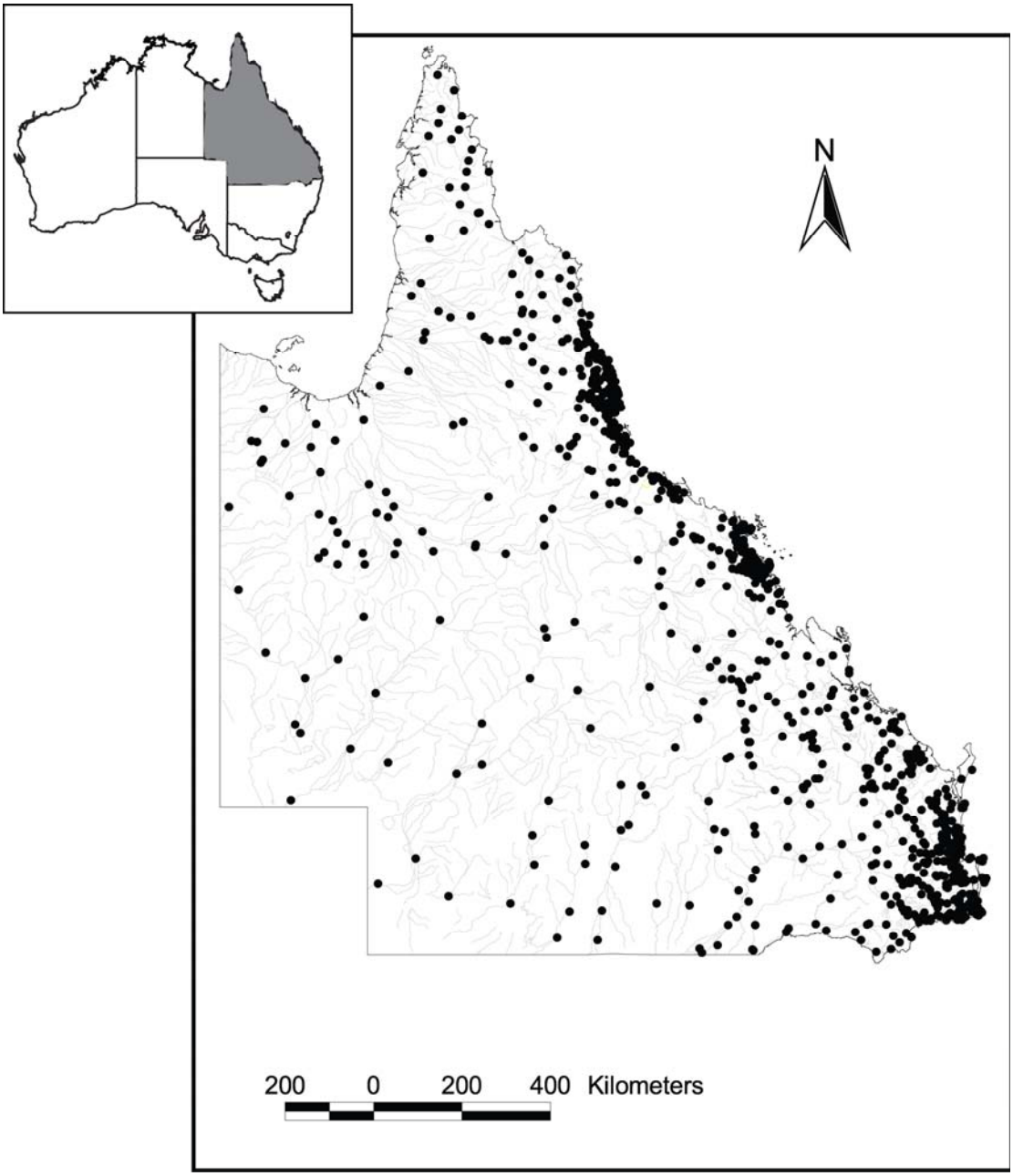


Figure 3.1. Map of the sampling sites in QLD dataset.

Table 3.1. Variables contained in dataset from QLD with minimum, maximum and mean values.

Variable	Abbreviation	Units	Type	Mean	Min	Max
Habitat	Habitat	Edge or Riffle	categorical			
Season	Season	Autumn or spring	categorical			
Position of site in the catchment in relation to watershed.	0-2 Reach		categorical		0	2
Latitude	Latitude	Decimal	continuous	-21.96	-29.01	-11.15
Longitude	Longitude	Decimal	continuous	148.25	138.11	153.51
Altitude	Altitude	m	continuous	168.01	1.00	950.00
Distance From Source	DFS	km	continuous	104.74	0.56	1198.80
Stream order	Stream Order		categorical		1	9
Depth at habitat	Depth	m	continuous	0.29	0.00	2.00
Width at habitat	Width	m	continuous	16.36	0.30	2000.00
Mean Depth over the sampled area	Mean Depth	m	continuous	0.46	0.05	16.00
Mean Channel Width	Mean Channel Width	m	continuous	70.68	2.00	2500.00
Instantaneous Discharge	Instantaneous Discharge	cumec	continuous	2.33	0.00	1850.00
Slope	Slope	km/m	continuous	0.00	0.00	0.10
Soil type number	Soil Type Number		categorical		2	38
Vegetation type number	Vegetation Type Number		categorical		2	22
Mean phi	Mean phi		continuous	-1.35	-13.00	7.00
Assessment of the site from 0 being poor to 4 indicating excellent habitat conditions. See Conrick & Cockayne (2000).	0-4 . Habitats		categorical		1.00	6.00
Number of substrate categories at the site	0-8.substrate categories		categorical		1.00	7.00
Percentage of bedrock in the substrate	Bedrock	%	continuous	4.87	0.00	100.00
Percentage of boulder in the substrate	Boulder	%	continuous	4.28	0.00	95.00
Percentage of cobble in substrate	Cobble	%	continuous	14.10	0.00	100.00
Percentage of gravel in the substrate	Gravel	%	continuous	10.78	0.00	95.00
Percentage of pebble in substrate	Pebble	%	continuous	9.99	0.00	100.00
Percentage of sand in the substrate	Sand	%	continuous	30.03	0.00	100.00
Percentage of Silt/Clay in substrate	Silt/Clay	%	continuous	25.96	0.00	100.00
Detrital cover	Detrital cover	%	continuous	20.80	0.00	100.00
Mean annual rainfall	Mean annual rainfall	mm	continuous	1361.13	165.00	4500.00
Percentage rainfall in wet season	Percentage rainfall in wet season	%	continuous	76.11	51.90	110.14

Continuation of Table 3.1.

Variable	Abbreviation	Units	Type	Mean	Min	Max
Range in dry season monthly means	Range in dry season monthly means		continuous	57.16	3.00	304.00
Range in wet season monthly means	Range in wet season monthly means		continuous	189.11	17.00	1384.00
Ratio of mean dry season rainfall to mean wet season monthly rainfall	Ratio of mean dry season rainfall to mean wet season monthly rainfall		continuous	4.21	1.24	26.40
Maximum velocity at sampled area	Velocity - max	m/s	continuous	0.29	0.00	4.00
Mean daily maximum temperature	Mean daily max temp	°C	continuous	27.74	20.30	33.90
Mean daily minimum temperature	Mean daily min temp	°C	continuous	16.07	8.20	24.20
Electrical Conductivity adjusted for temperature	Conductivity	µS/cm	continuous	334.98	6.00	12000.00
Alkalinity	Alkalinity	mg/L CaCO3	continuous	85.32	0.00	999.00
pH	pH		continuous	7.46	4.14	10.00
Total hardness	Total Hardness	mg/L CaCO3	continuous	90.89	1.90	3750.00
Turbidity (NTU)	Turbidity (NTU)		continuous	31.18	0.00	1922.00
Total nitrogen	Total N	mg/L as N	continuous	0.57	0.04	35.02
Total phosphorus	Total P	mg/L as P	continuous	0.06	0.00	4.50
Water temperature	Water Temp	°C	continuous	22.19	5.10	38.40
Concentration of bicarbonate	HCO3-	mg/L	continuous	96.31	0.00	1158.00
Concentration of potassium	K+	mg/L	continuous	2.70	0.08	76.00
Concentration of carbonate	CO3--	mg/L	continuous	0.76	0.00	35.50
Concentration of calcium	Ca++	mg/L	continuous	17.01	0.10	686.40
Magnesium concentration	Mg++	mg/L	continuous	12.45	0.10	1705.00
Concentration of sulfate	SO4--	mg/L	continuous	9.69	0.00	3568.00

Victoria data

The dataset from the state of Victoria has been provided by Leon Metzeling, Victorian EPA. It contains abundances of 128 macroinvertebrate families sampled at 407 stream sites (Figure 3.2) between March 1990 and November 1998. The sampling sites were chosen in order to represent the main types of rivers in each of the 25 drainage basins defined by the Australian Water Resources Council (AWRC). Most sites were sampled on four occasions in spring and autumn over the two consecutive years and seasonal habitat data for single sites were combined (Marchant et al., 1999).

At each site, two habitats were sampled separately: the main-channel (often a riffle) and the bank or edge of the channel. In order to simplify clustering, only the database including edge habitats was used for this study. A sample consisted of a macroinvertebrate collection over a 10m transect for each habitat using a D-frame hand net (0.25 mm mesh), followed by 30 minutes picking of live specimens. Macroinvertebrates were preserved in 70% ethanol and identified to family level. Specimens of Oligochaeta, Hydracarina, and Nematoda were not identified further (Marchant et al., 1999).

Only environmental variables presumably not affected by human activity (natural variables) were used for this study. The variables Distance from source, Slope, Altitude, Catchment area, Width, Alkalinity, Macrophyte taxa and Macrophyte abundance category were log-transformed (Table 3.2).

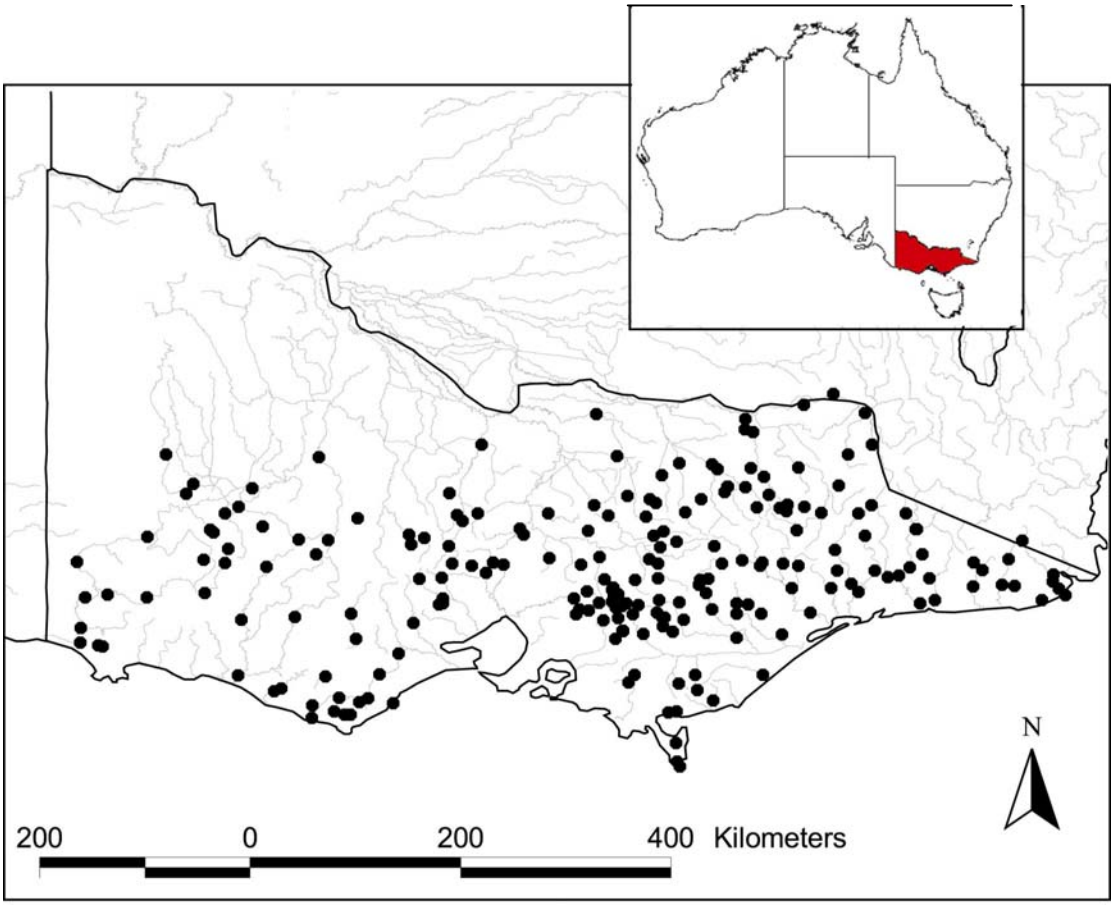


Figure 3.2. Locations of sampling sites in Victoria data set.

Table 3.2. List of variables in Victoria data set with minimum, maximum and mean values.

Variable	Abbreviation	Units	Type	Mean	Min	Max
Alkalinity (L)	LALK	mg/L	continuous	1.44	0.39	2.69
Catchment area (L)	LAREA	km	continuous	2.35	-0.52	4.13
Vegetation category	VEGCAT		categorical		1	4
Number of macrophyte taxa (L)	LMACTAXA		continuous	0.51	0	1.14
Slope(L)	LSLOPE	km/m	continuous	0.7	-0.79	2.51
Distance from source (L)	LDFS		continuous	1.41	-1	2.73
Longitude	LONG	Decimal	continuous		141.2412	149.6898
Latitude	LAT	Decimal	continuous		-39.1167	-35.9295
Macrophyte category (L)	LMACCAT		categorical		0	0.69
Shade	SHADE		categorical		0	5
Reach phi	REACHPHI		continuous	-1.99	-8.41	8.75
Altitude(L)	LALTITUDE	m	continuous	2.28	1	3.2
Width (L)	LWIDTH	m	continuous	0.88	-0.52	2
Substrate heregeneity	SUBHETERO			2.65	0	5
Percentage of Pebble	PEBBLE	%	continuous	11.77	0	80
Percentage of Cobble	COBBLE	%	continuous	20.3	0	80
Percentage of Boulder	BOULDER	%	continuous	12.28	0	80
Percentage of Bedrock	BEDROCK	%	continuous	6.61	0	99
Percentage of Gravel	GRAVEL	%	continuous	9.71	0	70

New South Wales (NSW) data

The database from NSW has been provided by Bruce Chessman from NSW Department of Land and Water Conservation and was the result of a Multi-Attribute River Assessment (MARA) survey of 122 sites on unregulated streams in 12 sub-catchments within four catchments (Figure 3.3). It contains 22 variables (Table 3.3) describing:

- Physical settings and diversity (geographical position, altitude, slope, distance from source, rainfall, substrate heterogeneity, etc.)
- Biological variables (diversity of macroinvertebrates, diversity of macrophytes, diversity of native fishes)
- Risk factors (flow, water temperature, phosphorus, nitrogen, oxygen, organic matter, etc.)

Sampling of macroinvertebrates has been conducted using standard methods (see description of Queensland data). More detailed description of NSW data can be found in Chessman (2002).

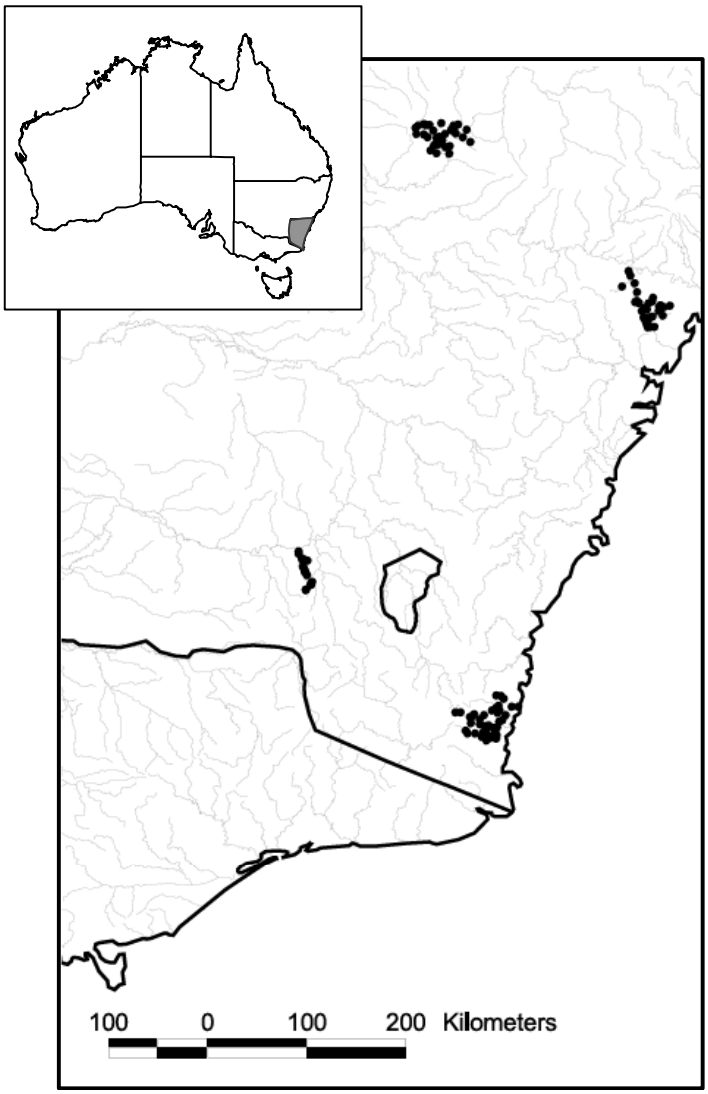


Figure 3.3. Locations of the sampling sites in NSW data set.

Table 3.3. List of biotic and abiotic variables available in NSW dataset with minimal, maximal and mean values.

Variable	Abbreviation	Units	Type	MEAN	MIN	MAX
Site elevation	Elev	m	continuous	268.16	5.00	780.00
Site slope	Slope	m/km	continuous	8.52	0.10	88.89
Site discharge	Flow	m^3/s	continuous	0.28	0.00	6.13
Average of maximum and minimum stream width per quadrat	Width	m	continuous	7.27	0.22	44.38
Average of maximum stream depth per quadrat	Depth	m	continuous	0.72	0.01	3.19
Number of diatom species per site	DiatomSp		continuous	38.31	6.00	78.00
Number of native aquatic macrophyte species per site	MacrNaSp		continuous	11.17	2.00	29.00
Number of native macroinvertebrate families per site	InveNaFa		continuous	31.12	13.00	59.00
Number of native fish species per site	FishNaSp		continuous	2.17	0.00	9.00
Macroinvertebrate family biotic index (SIGNAL 1995 version) (range 1-10)	SIGNAL95		continuous	5.42	4.35	7.02
Number of native fish individuals per hour of electrofishing	FishNNPH		continuous	201.16	0.00	1630.00
Water temperature at 0.2 m	TempSur	^oC	continuous	20.14	6.40	38.00
Turbidity at 0.2 m	Turb	NTU	continuous	12.30	0.40	64.70
Electrical conductivity at 0.2 m	EC	uS/cm	continuous	370.45	33.00	2330.00
pH at 0.2 m	pH		continuous	7.42	4.42	8.70
Ammoniacal nitrogen at 0.2 m	NH3	mg/L	continuous	0.06	0.01	1.60
Oxidised (nitrate plus nitrite) nitrogen at 0.2 m	NOx	mg/L	continuous	0.05	0.01	1.00
Filterable phosphorus at 0.2 m	FiltP	mg/L	continuous	0.03	0.00	0.85
Bank erosion score (range 0-100)	Erosion		categorical	8.37	0.00	96.43
Number of alien fish individuals per hour of electrofishing	FishANPH		continuous	273.12	0.00	4736.67
Stock damage score (range 0-100)	Stock		categorical	13.64	0.00	78.13
Catchment area above site	CatArea	km	continuous	231.75	1.00	1815.75

3.2 Data preprocessing and modelling

All data were preprocessed before modeling. Missing data were substituted for the average of variable in the relevant subset. Data manipulation and detailed analysis for each case study described in each relevant chapter. I used the following software for the supervised models: Matlab 5.3 with Neural Networks toolbox and Neural Solutions 4.0. Self Organising Maps were built and visualised using the SOM Toolbox for Matlab 5.3, developed at the Laboratory of Computer and Information Science (CIS) at Helsinki University of Technology and MOPED 1.11 (Modelling Patterns in Environmental Data, by NIWA). Supporting statistical analysis was conducted using STATISTICA 6.0 and various functions in Matlab 5.3. Canonical Correspondence Analysis was conducted using CANOCO 4.5 software package.

3.2.1 Unsupervised neural networks: Self Organising Maps

SOM is an excellent tool in the visualisation of high dimensional data. As such it is most suitable for data understanding phase of the knowledge discovery process, although it can be also used for data preparation, modelling and classification as well (Vesanto et al., 2000). A number of techniques are available for the visualisation of SOM. Based on the purpose, they can be divided into three groups: visualisation of components, visualization of cluster structure and shape and visualisation of data on the map.

Component planes

Each component plane shows the values of one variable in each map unit. The component plane can be thought of as a slice of the map: it consists of the values of a single vector component in all map units. Coupled with the clustering information, the component planes show the values of the variables in each cluster. By comparing component planes with each other, correlations are revealed as similar patterns in identical positions of the component planes: whenever the values of one variable change, the other variables change too. Component planes are very convenient when one has to visualize a lot of information at once. Based on overall view it is easy to select interesting component combinations for further investigation (Vesanto et al., 2000).

Clustering and visualization of cluster structure

The first and most well known type of SOM visualisation is a Unified distance matrix (U-matrix) (Vesanto et al., 2000). U-matrix visualizes distances between neighbouring map units, and helps to see the cluster structure of the map: high values of the U-matrix indicate a cluster borders, uniform areas of low values indicate clusters themselves. It is possible to make decisions about the number of clusters using U-matrix alone, however, when using extensive dataset and U-matrix with a large number of cells it is often difficult to draw clear borders between clusters based on U-matrix only.

A widely adopted definition of optimal clustering is a partitioning that minimizes distances within and maximises distances between clusters. In practice, most clustering methods do not produce a single clustering, but offer several with different number of clusters. To select the best among them, each can be evaluated using some kind of validity index. There is a multitude of different validity indices, when using SOM toolbox Davies-Bouldin Validity Index (Davies and Bouldin, 1979) was implemented and when using MOPED optimal clustering was chosen with the help of Silhouett index (Rousseeuw, 1987).

Davies-Bouldin Validity Index

This index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} \left\{ \frac{S_n(Q_i) + S_n(Q_j)}{S(Q_i, Q_j)} \right\},$$

where n - number of clusters, S_n - average distance of all objects from the cluster to their cluster centre, $S(Q_i, Q_j)$ - distance between clusters centres. Hence the ratio is small if the clusters are compact and far from each other. Consequently, Davies-Bouldin index will have a small value for a good clustering (Davies and Bouldin, 1979).

Silhouette Validation Method

The Silhouette validation technique calculates the silhouette width for each sample, average silhouette width for each cluster and overall average silhouette width for a total data set. Using this approach each cluster could be represented by so-called silhouette, which is based on the comparison of its tightness and separation. The average silhouette width could be applied for evaluation of clustering validity and also could be used to decide how good is the number of selected clusters.

To construct the silhouettes *S(i)* the following formula is used:

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}},$$

where *a(i)* –average dissimilarity of *i*-object to all other objects in the same cluster; *b(i)* – minimum of average dissimilarity *of i*-object to all objects in other cluster (in the closest cluster).

It is followed from the formula that $-1 \leq s(i) \leq 1$. If silhouette value is close to 1, it means that sample is “well-clustered” and it was assigned to a very appropriate cluster. If silhouette value is about zero, it means that that sample could be assign to another closest cluster as well, and the sample lies equally far away from both clusters. If silhouette value is close to –1, it means that sample is “misclassified” and is merely somewhere in between the clusters. The overall average silhouette width for the entire plot is simply the average of the *S(i)* for all objects in the whole dataset.

The largest overall average silhouette indicates the best clustering (number of cluster). Therefore, the number of cluster with maximum overall average silhouette width is taken as the optimal number of the clusters. (Rousseeuw, 1987; http://www.cs.tcd.ie/Nadia.Bolshakova/validation_algorithms.html).

K-means clustering

There is a huge number of different kinds of clustering algorithms available. The two main ways to cluster data are hierarchical and partitive approaches. The example of hierarchical way of clustering is a dendrogramm, which does not provide a unique

clustering, rather a partitioning can be achieved by cutting the dendrogramm at certain levels. Partitive clustering algorithms divide a data set into a number of clusters, typically by trying to minimize some criterion or energy function. If a number of clusters is unknown, the partitive algorithm can be repeated for a set of different values, typically from 2 to $\sqrt{n}$, where n is the number of samples in the data set. The most commonly used partitive algorithm is the k-means (MacQueen, 1967). In simple form k-means algorithm consists of following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated

For this study I used k-means clustering algorithm in order to groups SOM cells together on the base of their similarity.

In some studies I used Mahalonobis distance to evaluate dissimilarity between clusters. Mahalanobis distance is a measure of distance between two points in the space defined by two or more correlated variables. For example, if there are two variables that are uncorrelated, then we could plot points (cases) in a standard two-dimensional scatterplot; the Mahalanobis distances between the points would then be identical to the Euclidean distance; the distance as, for example, measured by a ruler. If there are three uncorrelated variables, we could also simply use a ruler (in a 3-D plot) to determine the distances between points. If there are more than 3 variables, we cannot represent the distances in a plot any more. Also, when the variables are correlated, then the axes in the plots can be thought of as being non-orthogonal that is, they would not be positioned in right angles to each other. In those cases, the simple Euclidean distance is not an appropriate measure, while the Mahalanobis distance will adequately account for the correlations. For each group in our sample, we can determine the location of the point that represents the means for all variables in the multivariate space defined by the variables in the model. These points are called group centroids. For each case we can then compute the Mahalanobis distances (of the respective case) from each of the group centroids (MOPED – Modelling Patterns in Environmental Data, http://www.niwa.com.au/pubs/moped).

Visualisation of data on the map

In some cases is it necessary to visualise location of a particular subset of data (certain catchment for example) on general SOM. This can be achieved using ‘hit’ histograms, which show in graphic form the distribution of the best matching units for a given data set (Vesanto et al., 2000). Multiple histograms may be drawn and these are identified by different colors and /or markers. This makes it possible to compare different data sets by the distribution of their ‘hits’ on a map. See Figure 6.3 for the example of hit diagrams.

Quality of SOM

There are two measurements of SOM quality: average quantization error and topographic error. Average quantization error is simply the average distance (weighted with mask) from each data vector to its best matching unit (BMU). Topographic error gives the percentage of data vectors for which the BMU and the second BMU are not neighbouring map units. In practice, large quantization errors are observed when using non-normalized data due to the wide range of values in environmental variables, after normalization quantization error is usually small (less than 1).

Data analysis using SOM

The following logical sequence in SOM implementation was applied with variations according to the purpose of particular case study.

1) Data preprocessing and normalization. 'Range' normalization has been usually applied, scaling data between 0 and 1 using the following formula (Vesanto et al., 2000): $v'(i) = (v(i) - \min(v)/(\max(v) - \min(v))$

'log' transformation has also been used for some studies: $v'(i) = \ln(v-\min(v) +1)$.

2) SOM built using default settings provided by SOM toolbox.

3) SOM U-matrix and component planes visualised.

4) Several k-means partitioning tried and evaluated using Davies-Bouldin (or Silhouette) index. The clustering with the smallest index (largest in case of Silhouette) is selected (Vesanto and Alhoniemi, 2000).

5) SOM with resulting k-mean clustering is visualised and cluster number for each data unit is exported and used for further analysis or GIS visualisation.

3.2.2 Supervised Neural Networks: Multilayer Perceptron and Feedforward networks

Slightly different architecture of ANNs were used for the different case studies, in general I used generalized feedforward networks, which are a generalization of MLP such that connections can jump over one or more layers. In theory, a MLP can solve any problem that a generalized feedforward network can. In practice, however, generalized feedforward networks often solve the problem much more efficiently (NeuroDimensions, 1999).

When building an MLP model the decision has to be taken on:

- number of hidden layers and number of neurons in hidden layers
- transfer function
- step size, momentum coefficient
- number of iterations

I used MLP with one hidden layer for all studies except the scenario analysis (Chapter 7). In theory, one hidden layer is sufficient to approximate any function (Sovan Lek, personal communication), but in practice the predictive ANN with multiple hidden layers (as modular ANN, see Chapter 7) can offer some advantages in some cases.

Houng (2001) conducted optimisation study using the QLD data and I used a momentum and step size suggested by her, namely: learning rate for hidden layer and for output layer 0.1 and 1 respectively, and momentum coefficient 0.9. Different number of neurons in hidden layer was used for the different case studies, depending on the number of input neurons (the more input neurons the more hidden neurons), the optimum number was decided by trial and error in each case.

Maier and Dandy (1998) have shown that hyperbolic tangent transfer ('tahn', between –1 and 1) produces better performance in terms of root mean square errors (RMSE) between actual and predicted output, and learn quicker than linear or bipolar sigmoid (between 0 and 1). Hyperbolic tangent transfer function was used in all case studies here:
$\tanh z = e^{z} – e^{-z}/e^{z} + e^{-z}$,
where e is the the base of the natural algorithm.

Optimum number of iterations or epochs was determined using 'cross-validation' set, typically 10% of data. The aim of the training of a neural network is to minimize the output error with the respect to the known desired output. This error is defined as mean square error between the network outputs and the actual outputs. Cross-validation is executed in concurrence with the training of the network. Every so often, the network weights are frozen, the cross-validation data is fed through the network, and the results are reported. If error from the cross-validation set is getting larger than the error from the training set, this is the sign that the network has begun to overtrain and training should be stopped to ensure maximum generalization of the model. See Principe et al. (2000) for more on cross-validation and training.

Models used in all case studies were validated using randomly selected independent testing or validation set (typically 30% of data). A comparison between the actual and predicted values was made in order to evaluate performance of the network. The validation results are represented as percentage of correct predictions in case of presence-absence outputs and as Product-Momentum Correlation Coefficient in case of continuous outputs (as taxonomic richness, Salinity Index, etc.). In case of presence-absence predictions continuous output of the network was translated into '0' (absent) and '1'(present) by using cut-off value of '0.5'. In other words, if output is equal or more than 0.5- taxon was counted as present, if output is less than 0.5 – taxon is absent.

It was shown by Manel et al. (2001) that percentage of correct predictions as widespread measure of predictive accuracy is affected systematically by the prevalence (i.e. the frequency of occurrence) of the target organism, and reliance of

this measure using raw data can be misleading. For example, in case with rare taxon which is absent in 96% or cases and present in 4% of cases, ANN can guess all outputs as ‘0’ and be erroneously estimated as highly accurate with 96% of correct predictions. In order to avoid the problem with overrepresented ‘0’ or ‘1’, I equalized the testing or validation data by duplicating the data points so the dataset contains 50% of ‘0’ and 50% of ‘1’ values. In this case we can be sure that percentage of correct predictions is indeed indicative for the true accuracy of the network, as in the cases where network simply ‘guessing’ all outputs as ‘1’ or ‘0’, the total accuracy will never be higher than 50%.

In general, the following logical sequence was used for the majority of case studies:

1) Preprocess the data, normalize and randomise if necessary.
2) Build a variety of networks with a different number of hidden neurons, train with cross-validation and test on validation set.
3) Select the optimum performing architecture, build, train with cross-validation and test on validation set several networks of this architecture (typically 10 models).
4) Select the model with best performance using validation set.
5) Use the best performing model for simulation, sensitivity or scenario analysis.

3.3 Structure and functioning of the Stream Decision Support Framework

Figure 3.4 illustrates the principal structure and the corresponding functionality of the SDSF. The stream databases are structured into physical, biological and risk or water quality variables. The unsupervised ANN models (SOM) can be used for spatial ordination, clustering and diagnosis of stream sites. The supervised ANN models allow the prediction of the occurrence of aquatic macroinvertebrates depending on stream habitat and water quality conditions. In addition they can be used in order to elucidate the relationships between physical and biological variables by means of sensitivity analysis and conducting scenario analysis on potential impacts or restoration measures. Combination of techniques as SOM and MLP and neural networks with statistical methods provides an additional power and an opportunity to extract more information from the data available.

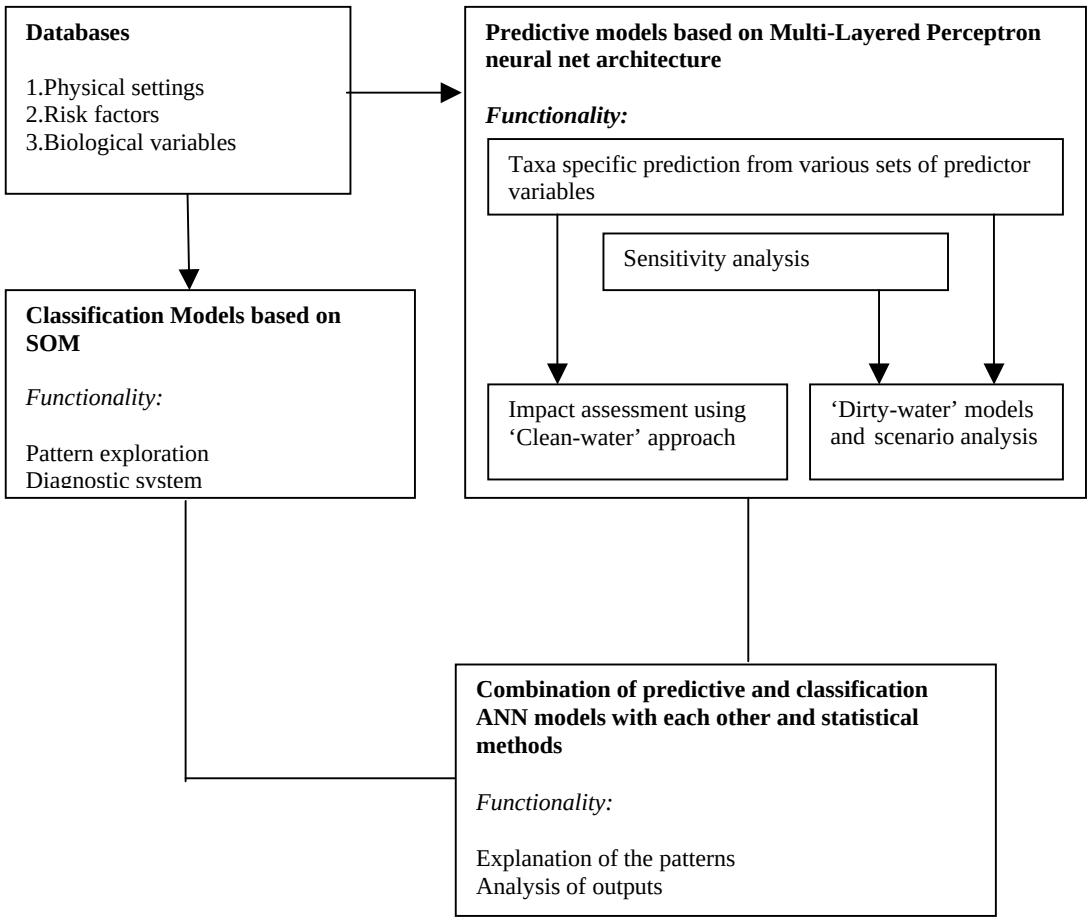


Figure 3.4. Structure of the Stream Decision Support Framework (SDSF).

Chapter 4

Ordination, clustering and correlation hunting using Self Organising Maps (SOM)

4.1 Exploring natural variability with SOM using referential datasets from Victoria and Queensland

Recognizing the existence of natural geographical variations in species distribution is an important consideration in the development of biomonitoring programmes, particularly at large spatial scales (Karr, 1991). Australia is a continent of diverse ecosystems with a variety of environmental conditions. To be able to detect changes in ecological characteristics caused by anthropogenic impacts it is necessary first to distinguish those changes from the natural variability inherent to the particular aquatic ecosystems. Some of the streams can show natural impoverishment due to the variety of reasons, which should not be confused with an impoverishment caused by the human induced stresses. When using the referential approach it is particularly important to be able to recognise systems with similar habitat characteristics and macroinvertebrate communities in order to be able to compare those systems in reference and potential impacted conditions. In AusRivAs UPGMA (see Chapter 2) procedure is used to cluster sites together on the basis of the faunal similarity. Self Organising Map neural networks provide an alternative approach to the clustering sites on the basis of their similarity plus it offers an interesting method to visualise distribution of both biotic and environmental variable on the same spatial scale using so-called component planes (see Chapter 3 for the description).

This study demonstrates the use of both component planes and clustering using two datasets from Victoria (pooled abundance data) and Queensland (presence-absence data) collected only from the sites assessed as reference or minimally impacted sites. The selection of reference sites was based on the Monitoring River Health Initiative method outlined in the River Bioassessment Manual (Davies, 1994). It is very difficult to find completely undisturbed streams in Australia and it is possible that reference sites can be under some kind of anthropogenic influence, however, their

faunal characteristics considered to be as natural as it is possible under modern conditions. With this assumption I used these datasets in order to gain an understanding of the broad spatial patterns in macroinvertebrate communities and try to identify the most important environmental variables structuring these communities in the conditions maximally close to the natural. Methodologically this task can be addressed in a variety of ways. Giraudel and Lek (2002) suggested a method allowing determining the most relevant variables for structuring the obtained map using a Structuring Index. However, according to this method we must use all variables in question as an input for the SOM to determine their contribution to the general clustering. In our case, we are mainly interested in the variability between macroinvertebrate communities. If we to use occurrence pattern of macroinvertebrates and environmental variables as an input, resulted clustering will reflect variability in the environmental variables as well as variability in macroinvertebrate communities. In this chapter I suggest use of statistical procedures as ANOVA and Mahalonobis distance and descriminant analysis as well as SOM. This approach combines flexibility and interesting visualisations of SOM with qualitative expression of statistical methods ('p' values, etc.). I used a slightly different combination and succession of the methods for Victoria and Queensland data in order to demonstrate the flexibility and potential of the Stream Decision Support Framework.

The main hypotheses to be tested in this study are:

1) Using SOM it is possible to identify sites similar on the base of the taxonomic composition of macroinvertebrate communities and isolate them into the clusters.

2) These clusters are similar to the bioregions previously described for Victoria and Queensland.

3) SOM component planes provide valuable information in order to characterize those clusters by the environmental factors.

4) SOM component planes provide valuable information on the relationships between abiotic and biotic variables.

4.1.1 Victoria dataset

Method

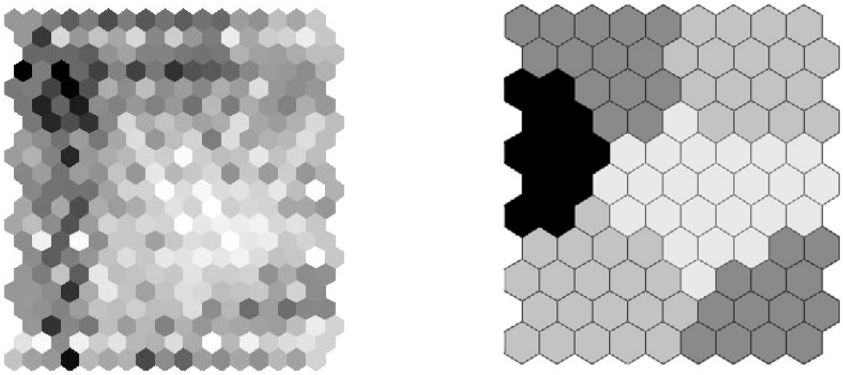
The abundance pattern of 128 macroinvertebrate taxa collected from 407 sites in the edge habitat was examined using SOM. Resulting map was partitioned further by k-means algorithm into a minimum optimum number of clusters using SOM toolbox for Matlab (see Chapter 3 for k-means partitioning). Data were not normalized for this part of the analysis. Clusters were analysed using ANOVA, mean values of the environmental variables in each cluster and Mahalonobis distance between clusters. In order to investigate relationships between variables and discovered patterns using component planes we built the second SOM with 19 environmental variables (see Table 4.2 for the description of variables) as an input, all data normalized between 1

and 0. Data subsets corresponding to the clusters discovered earlier was visualised using hit diagrams (see Chapter 3 for the description).

Result

The first SOM built for the purpose of finding similar pattern in macroinvertebrate communities has the following characteristics: map size 9x11 cells, topographic error: 0.02, quantisation error: 1.09.

The U-matrix of the resulting SOM is shown on Figure 4.1(a) and the six clusters or groups resulting from partitioning by k-means algorithm on Figure 4.1(b). The six clusters or groups roughly corresponding to the five ecological zones defined by the EPA Victoria (Metzeling et al., 2001), (Figure 4.2). Group 1 corresponds to Forest A and Highlands, Group 2 to Cleared Hills and Coastal Plains, Groups 3 and 4 to Forest B, Group 5 to Highlands and Group 6 to Murray and Western Plains.



a b

Figure 4.1. SOM outputs: a) U-matrix, b) Partitioning into 6 clusters by the K-means algorithm.

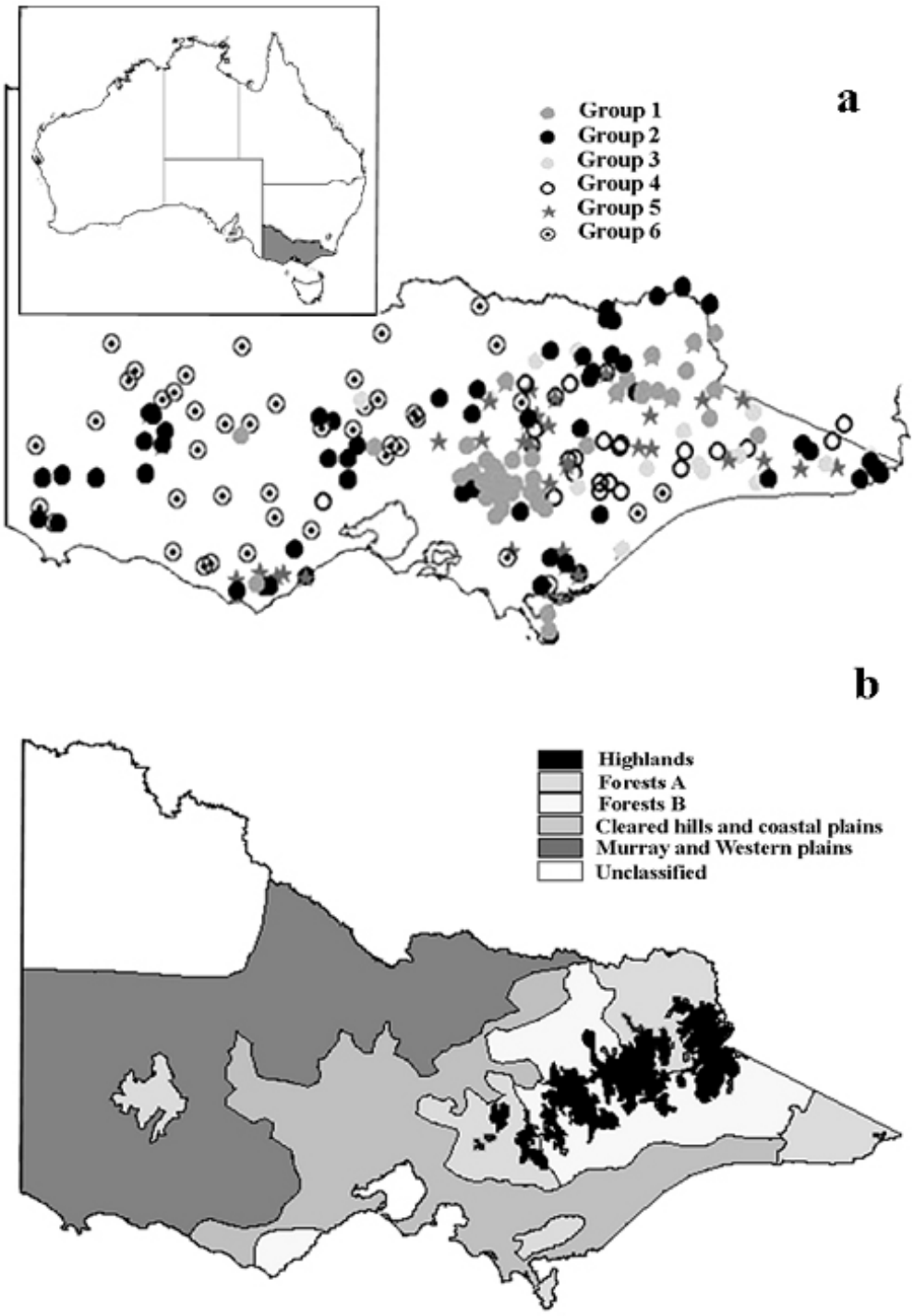
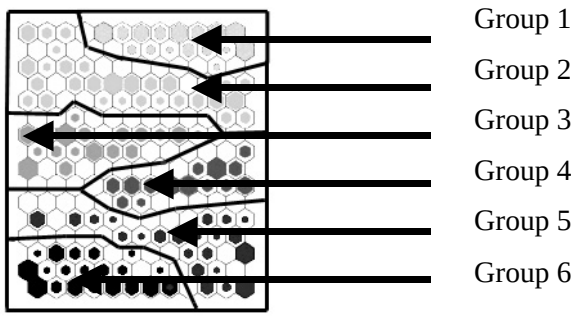
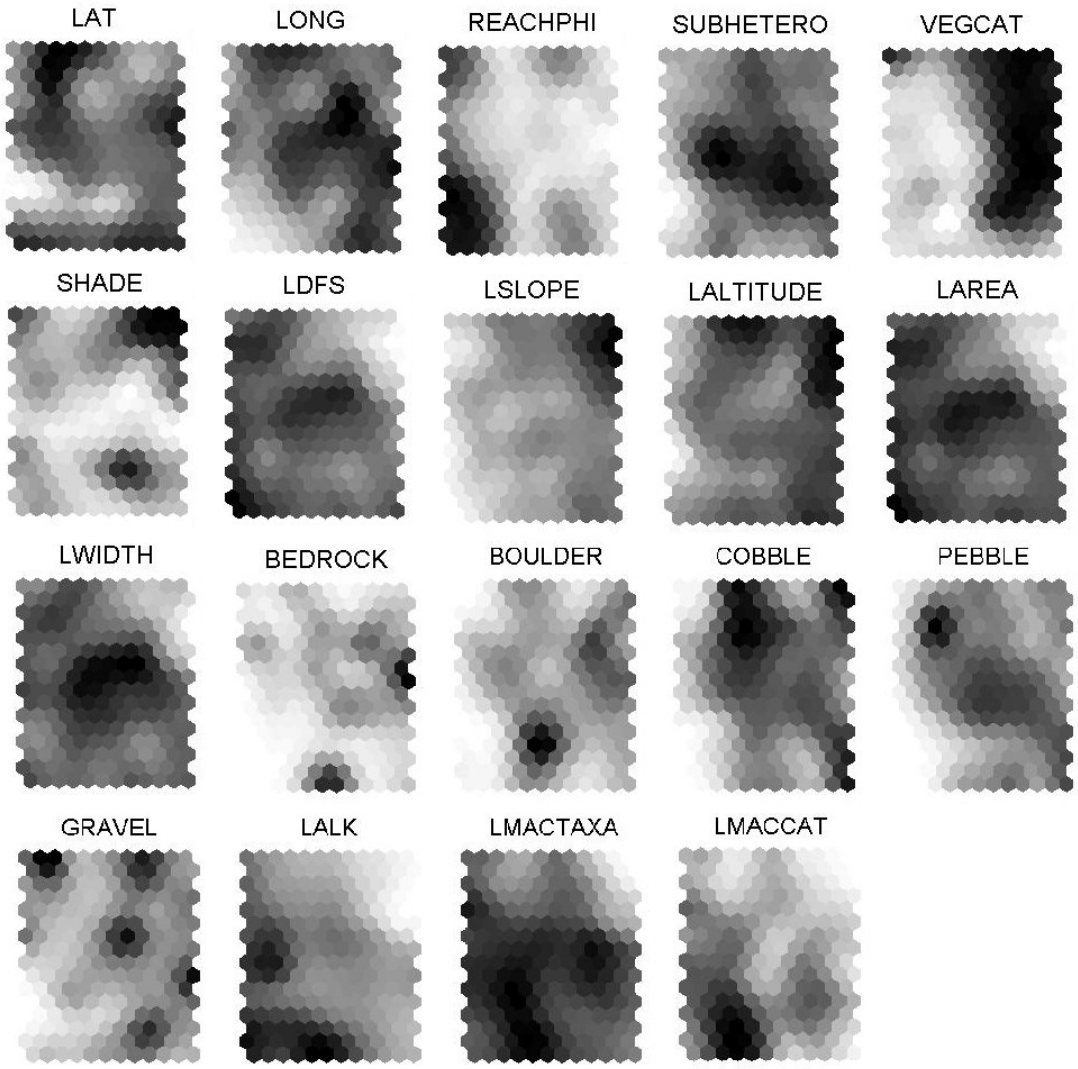


Figure 4.2. a) Distribution of macroinvertebrate groups resulting from SOM (sites belonging to the same group have the same marker), b) biological regions in Victoria based on benthic macroinvertebrates (Metzeling et al., 2001).



a)



b)

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.

The second SOM built for the purpose of viewing the component planes was sized 12x16, topographic error: 0.02, quantisation error: 0.58.

Figure 4.3(a) shows a distribution of 6 groups or clusters on SOM grid (see Chapter 3 for more on hit diagrams). Clusters appear to be well separated from each other without overlap between them. Figure 4.3(b) shows 19 SOM component planes for all environmental variables (see Table 4.2 for abbreviations). Alkalinity is particularly low at the area corresponding to the group 1, and particularly high at the groups 6 and 3. Number of macrophyte taxa is comparatively low at the groups 1 and 2, and comparatively high at all other groups. Distribution of macrophyte category amongst clusters is quite patchy, but distinctively different in the group 1 and 6. Distribution of values for vegetation category does not follow horizontal gradient characteristic for distribution of clusters at first sight, but it might be an important variable to distinguish between groups 1 and 6, and to some extent between groups 3 and 4.

Groups 1 and 6 clearly differ in relation to slope, with high values for this variable at the group 1 and low values at the group 6. Group 1 is also can be characterized by relatively high altitude, although for other groups altitude does not seem to be falling into any distinctive pattern. Other variables do not appear to have any distinctive pattern corresponding to the distribution of clusters (groups) and will not be considered here.

Table 4.1 shows mean values of the continuos environmental variables in each of 6 SOM defined clusters. Using discriminant analysis we have been able to predict correctly 60.0% (244 of 407 sites) of group membership using environmental settings as independent variables with 20.9% of near misses. Practically all variable were significant ($P<0.05$) in discriminating between clusters (Table 4.2) with Alkalinity having the largest F value. The Mahalonobis distance (Table 4.3) was the largest between cluster 1 and cluster 6. This is not surprising as the cluster 1 corresponds to the forest and highland ecosystems and cluster 6 to Murray and Western plains (see Figure 4.2).

Table 4.1. Mean values of the continuous environmental variables in each of 6 SOM defined clusters, total abundance of macroivertebrate and number of macroinvertebrate families are added for comparison. (L) – the variable was provided log-transformed.

Cluster	1	2	3	4	5	6
N of sites in cluster	75	79	54	45	74	80
Latitude	-37.52	-37.46	-37.33	-37.37	-37.59	-37.30
Longitude	146.25	145.71	146.50	147.03	146.13	143.80
Reach phi	-3.10	-2.69	-2.43	-4.89	-1.65	1.39
Substrate heterogeneity	2.68	2.49	2.94	3.33	2.71	2.12
Shade	3.06	2.38	1.75	1.56	2.37	1.85
Distance from source (L)	0.89	1.30	1.76	1.45	1.38	1.76
Slope(L)	1.18	0.71	0.47	0.80	0.76	0.32
Altitude(L)	2.63	2.20	2.07	2.38	2.29	2.13
Catchment area (L)	1.51	2.15	2.91	2.54	2.35	2.84
Width (L)	0.61	0.82	1.16	1.07	0.88	0.93
Bedrock (%)	4.95	9.25	4.39	12.84	3.12	6.80
Boulder (%)	13.81	11.55	11.88	19.17	12.33	7.90
Cobble (%)	26.70	22.50	21.05	22.98	20.69	9.73
Pebble (%)	12.07	12.20	13.20	17.68	12.79	5.84
Gravel (%)	11.47	10.64	10.02	11.50	9.25	6.33
Alkalinity (L)	0.96	1.27	1.74	1.36	1.30	2.05
Number of macrophyte taxa (L)	0.25	0.45	0.65	0.62	0.54	0.63
Total abundance	260.44	413.19	1026.35	661.75	617.12	788.82
Number of macroinvertebrate families	25.18	24.93	30.31	33.17	29.48	26.82

Table 4.2. Description of environmental variables and results of univariate analysis between 6 SOM defined clusters.

Variable	Abbreviation	Mean	Min	Max	F	P
Alkalinity (L)	LALK	1.44	0.39	2.69	89.2	0.000
Catchment area (L)	LAREA	2.35	-0.52	4.13	40.5	0.000
Vegetation category	VEGCAT		1	4	40.4	0.000
Number of macrophyte taxa (L)	LMACTAXA	0.51	0	1.14	38.5	0.000
Slope(L)	LSLOPE	0.7	-0.79	2.51	34.8	0.000
Distance from source (L)	LDFS	1.41	-1	2.73	33.5	0.000
Longitude	LONG		141.24	149.68	28.3	0.000
Macrophyte category (L)	LMACCAT		0	0.69	20.9	0.000
Shade	SHADE		0	5	20.4	0.000
Reach phi	REACHPHI	-1.99	-8.41	8.75	20.1	0.000
Altitude(L)	LALTITUDE	2.28	1	3.2	18.8	0.000
Width (L)	LWIDTH	0.88	-0.52	2	18.7	0.000
Substrate heregeneity	SUBHETERO	2.65	0	5	11.7	0.000
Pebble (%)	PEBBLE	11.77	0	80	10.4	0.000
Cobble (%)	COBBLE	20.3	0	80	10.0	0.000
Boulder (%)	BOULDER	12.28	0	80	5.2	0.000
Becrock (%)	BEDROCK	6.61	0	99	4.5	0.001
Gravel (%)	GRAVEL	9.71	0	70	3.3	0.006
Latitude	LAT		-39.1167	-35.9295	2.1	0.070

Table 4.3. Mahalonobis distances between 6 SOM defined clusters (environmental settings used as independent variables).

	2	**3**	**4**	**5**	**6**
1	2.21	3.71	2.95	2.12	4.37
2		2.07	2.17	1.35	2.86
3			2.15	1.97	2.20
4				1.70	3.46
5					2.94

4.1.2 Queensland dataset

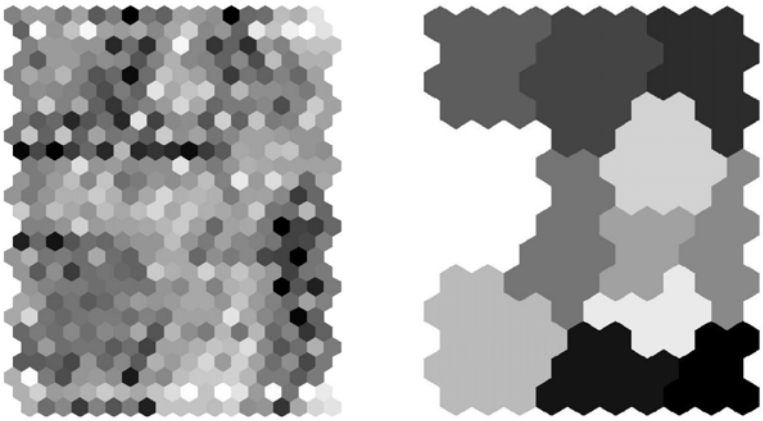
Method

In order to explore distribution patterns of macroinvertebrate communities in Queensland streams I built a SOM using 69 most commonly occurring macroinvertebrate taxa (taxa occurring at more than 5% of sites) from the riffle habitat. A set of 28 physico-chemical variables was used to explain patterns in macroinvertebrate occurrence (see Table 3.1 for the data description). In comparison with previous study (Victoria dataset) I used slightly different approach, as Queensland dataset contains many more environmental variables than the one from Victoria. Also, I wanted to explore a variety of approaches to the solution of similar

problems. First I built SOM using only presence-absence of macroinvertebrate taxa. Then the discriminate analysis has been applied twice using two sets of variables 1) physico-chemical variables, 2) biotic variables (occurrence pattern of macroinvertebrate taxa, taxonomic richness and PET richness). The first ten variables discriminating the best between clusters (10 environmental variables and 10 biotic variables) were used to build a second SOM in order to compare component planes and visualise the major gradients in both abiotic and biotic factors.

Results

Figure 4.4 shows the resulting SOM U-matrix and its partitioning into 12 kmean clusters using standard procedure provided by SOM toolbox for Matlab (see Chapter 3). Table 4.4 shows mean values of the physico-chemical variables for each cluster. It was possible to correctly predict 40.8% (311 of 763) of cluster membership with 21.0% of near misses (160 of 763) using discriminant analysis. All environmental variables with the exception of depth were significant in discriminating between clusters (Table 4.5). Similarly, the majority of macroinvertebrate taxa (with the exception of 5 out of 70) were significant ($p<0.05$) in discriminating between clusters.



a) b)

Figure 4.4. a)U-matrix and, b) k-means partitioning into 12 clusters, QLD reference sites, macroinvertebrates only.

Mahalonobis distance was the largest (more than 3) between clusters 1 and 2, 2 and 5, 1 and 12, 5 and 12, 7 and 12. Figure 4.5 shows spatial distribution of all clusters in groups of 4 (it would make it too difficult to read if all 12 clusters were shown on one plot). There is only a partial correspondence to the QLD bioregions (Figure 4.6) with the majority of the clusters spread over two or more bioregions. As rainfall variables (Table 4.5) were among the most important in discriminating between clusters we expect that the distribution of clusters should to some extent follow the rainfall pattern. This becomes apparent when distribution of clusters is compared with rainfall

distribution throughout the state (Figure 4.7). For example, clusters 2, 3, 8, 9 (Figure 4.5) are found in the areas with high annual rainfall, cluster 7 is mostly found particularly in areas with high mean wet season monthly rainfall and clusters 8 and 12 in the areas with high mean dry season monthly rainfall. Clusters 1, 5, 6 and 11 are mostly found in inland areas with comparatively low mean annual rainfall. Clusters of sites located in the areas with low rainfall are very different (Mahalonobis distance, Table 4.6) from those located in the areas with high rainfall, for example sites within cluster 2 (mean annual rainfall 2518.68 mm) and sites within cluster 1 (mean annual rainfall 966.93 mm).

The second important factor discriminating between the clusters is distance from source. Importance of distance from source and stream order has been explained by River Continuum Concept (RCC) (Vannote et al., 1980), which relates sources of energy inputs into the aquatic system to the river/stream inhabitants. In other words, changes in available resources along the stream continuum from headwater to lowland are reflected by faunal composition. The applicability of RCC in Australia is discussed further in Chapter 4.1.3. Slope is naturally related to distance from source as headwater streams will be mostly found in hilly areas. For example, sites within clusters 2 and 12 are characterized by small distance from source and relatively high slope, when sites located in clusters 1 and 5 are characterized by the opposite, high distance from source and low slope (see Table 4.4).

Latitude and longitude reflect a variety of factors including historical aspect of the development of a site's fauna. Some macroinvertebrates (as insects) are not confined to particular stream and can migrate between unconnected streams and rivers, however many others cannot (as crustaceans, mites and so on). Geographical position and history of each site are naturally important factors contributing into the structure of macroinvertebrate communities.

Flow (measured as maximal velocity) is obviously an important factor affecting structure of macroinvertebrate communities. Many taxa have particular flow preferences and adaptations enabling them to survival in particular flow conditions.

Water quality variables are next in order of importance in discriminating between SOM clusters (Table 4.5). Even though reference sites suppose to be unaffected by human activity and have generally good water quality, there is still significant natural variability between site in relation to parameters like water temperature, pH, turbidity and conductivity.

Substrate composition variables are obviously of local importance and although still significant in discrimination between clusters are positioned in the end of the list (Table 4.5) as well as temporal variables (Year and Season).

The second SOM built in order to visualize the most important environmental gradients in comparison with biotic variables contributing the most to the discrimination between clusters has the following characteristics: size 14x10 cells, final quantization error: 0.56 and final topographic error: 0.03. Figure 4.8 shows 20 component planes for environmental and biotic variables providing the best discrimination between SOM defined clusters.

Rainfall, geographical position and distance from source are three most prominent gradients. Water temperature gradient appears to be related to both geographical position and distance from source. This is naturally explainable by changes in water temperature from cooler south to the warmer north and changes along the river continuum as sites downstream are generally wider and less shaded by riparian vegetation. Distribution of Velocity and Total Nitrogen show irregular pattern with velocity to some extent positively correlated with rainfall variables.

Not surprisingly distribution of taxonomic richness and PET richness appear to be patchy and irregular with so many different gradients affecting macroinvertebrate communities. Interestingly, the highest taxonomic richness is found in the sites positioned in the middle of the map corresponding to the areas with medium values of abovementioned environmental gradients. It appears that extremes in any of the environmental factors are reducing taxonomic richness. However, the highest values of PET richness are shifted towards areas with higher rainfall and flow.

When we consider component planes for the distribution of individual taxa as well, it becomes clear that taxa discriminating best between clusters have either limited geographical distribution or strong preferences towards one or more environmental factors. This tendency is evident from abrupt changes (very dark and very light shades) in frequency of occurrence in different areas. For example *Baetidae* seems to be limited to sites located not far from source and characterized by medium to high maximal velocity. *Hydrometridae*, on the contrary was found mostly downstream, but velocity does not appear to be a limiting factor.

Table 4.4. Mean values of abiotic variables in each of 12 SOM defined clusters.

Cluster	1	2	3	4	5	6	7	8	9	10	11	12
Number of samples in a cluster	77	51	69	68	103	87	71	46	56	38	63	34
Season	1.54	1.33	1.49	1.39	1.41	1.43	1.28	1.56	1.42	1.44	1.33	1.52
Habitat Depth (m)	0.32	0.33	0.33	0.32	0.35	0.38	0.32	0.30	0.31	0.30	0.33	0.31
Max Velocity (m/s)	0.02	0.20	0.08	0.30	0.07	0.10	0.15	0.07	0.14	0.05	0.05	0.13
Bedrock (%)	5.19	3.23	1.44	0.73	4.80	3.39	2.95	4.45	5.26	3.42	5.39	14.55
Boulder (%)	1.36	9.02	2.53	0.95	3.08	3.33	0.88	5.54	3.83	0.78	1.42	5.29
Cobble (%)	1.94	11.56	5.21	3.67	2.20	5.02	2.43	11.41	12.14	2.10	1.66	7.94
Pebble (%)	2.01	5.68	4.20	3.52	4.02	5.51	2.09	11.52	7.58	6.18	1.03	6.02
Gravel (%)	8.44	6.37	12.24	10.66	7.57	6.14	10.28	11.08	8.66	7.36	6.66	10.00
Sand (%)	45.97	43.13	39.27	59.85	46.69	48.11	59.29	34.23	40.53	44.07	36.98	31.17
Silt/Clay (%)	35.06	20.98	35.07	20.58	31.60	28.47	22.04	21.73	21.96	36.05	46.82	25.00
Mean phi	0.69	-1.57	0.67	-0.10	0.22	-0.07	-0.10	-1.66	-1.42	0.88	1.69	-2.14
Latitude	-19.42	-18.45	-20.67	-17.33	-17.83	-22.20	-17.18	-22.96	-19.06	-20.83	-22.83	-23.78
Longitude	145.61	146.56	147.62	145.09	144.10	148.32	144.76	149.21	146.69	147.95	147.01	150.16
Altitude (m)	219.75	148.58	185.24	290.40	154.90	161.31	168.22	284.37	217.67	168.31	229.09	385.88
Slope (m/m)	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.01
Distance From Source (km)	152.16	18.92	78.64	90.07	211.99	67.47	145.39	36.82	34.63	40.43	116.99	11.63
Mean annual rainfall (mm)	966.93	2518.68	1231.06	1414.46	996.95	1414.44	1235.54	1497.97	1792.10	1201.06	956.69	1505.61
Water Temp (°C)	25.05	21.15	22.93	24.23	26.20	22.00	26.4	19.73	21.90	23.58	21.10	18.75
Conductivity (µS/cm)	334.00	120.62	362.51	107.07	273.50	314.86	169.32	158.11	183.96	393.30	264.87	145.24
Dissolved oxygen (mg/L)	7.47	7.91	7.50	7.62	7.31	7.28	7.60	8.46	8.01	8.19	6.96	7.65
pH	7.76	6.97	7.64	7.20	7.67	6.93	7.36	7.31	7.27	7.49	7.35	7.17
Turbidity (NTU)	43.85	2.89	4.11	16.62	21.96	63.65	19.51	3.69	3.55	6.75	136.22	7.16
Alkalinity (mg/L CaCO3)	110.38	34.44	130.29	31.24	102.29	50.58	85.28	39.98	47.79	97.89	77.71	66.78
Total Nitrogen (mg/L as N)	0.61	0.19	0.31	0.20	0.46	0.55	0.34	0.30	0.22	0.44	1.29	0.31
Total Phosphorus (mg/L as P)	0.05	0.01	0.02	0.01	0.03	0.06	0.02	0.02	0.01	0.02	0.11	0.02
Mean wet season monthly rainfall	134.09	341.45	163.74	200.24	143.69	182.39	179.69	190.59	245.79	162.62	120.82	185.01
Mean dry season monthly rainfall	28.48	76.36	39.15	34.42	27.01	54.28	30.85	57.17	52.16	37.01	38.63	64.74

Table 4.5. Univariate analysis of variance between 12 SOM defined clusters using environmental variables as predictors.

Variable	F	P
Mean dry season monthly rainfall	24.3	0.000
Distance From Source (km)	22.4	0.000
Longitude	22.0	0.000
Latitude	22.0	0.000
Mean annual rainfall (mm)	21.9	0.000
Slope (m/m)	20.2	0.000
Mean wet season monthly rainfall	18.6	0.000
Max Velocity (m/s)	18.0	0.000
Water Temp (°C)	16.8	0.000
Total Nitrogen (mg/L as N)	13.4	0.000
Total Phosphorus (mg/L as P)	10.3	0.000
pH	10.1	0.000
Turbidity (NTU)	10.0	0.000
Cobble (%)	8.9	0.000
Conductivity (µS/cm)	8.4	0.000
Alkalinity (mg/L CaCO3)	8.3	0.000
Altitude (m)	6.4	0.000
Pebble (%)	5.7	0.000
Mean phi	5.4	0.000
Boulder (%)	5.0	0.000
Sand (%)	4.7	0.000
Silt/Clay (%)	4.4	0.000
Dissolved oxygen (mg/L)	3.3	0.000
Bedrock (%)	2.2	0.015
Gravel (%)	2.0	0.024
Habitat depth (m)	1.5	0.131

Table 4.6. Mahalanobis distance between groups using set of environmental variables as predictors.

	2	3	4	5	6	7	8	9	10	11	12
1	3.08	1.50	2.38	1.09	2.10	1.51	2.39	2.18	1.78	1.76	3.02
2		2.59	2.48	3.25	2.07	2.80	2.31	1.63	2.88	2.87	2.39
3			2.02	1.92	1.99	1.71	2.02	1.62	1.49	2.09	2.64
4				2.46	2.32	1.64	2.39	1.81	2.39	2.65	2.73
5					2.51	1.39	2.84	2.55	2.23	2.27	3.43
6						2.14	1.81	1.74	1.99	1.50	2.10
7							2.55	2.05	1.87	2.29	3.13
8								1.52	2.09	2.36	1.56
9									1.95	2.34	2.22
10										2.21	2.88
11											2.66

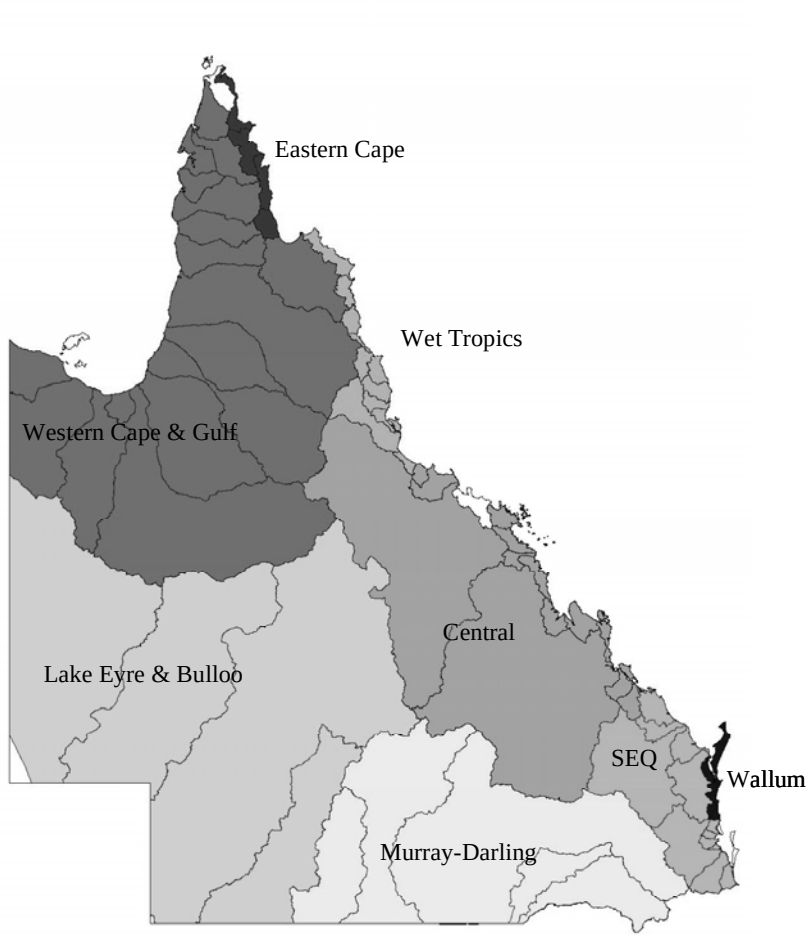


Figure 4.6. Bioregions of Queensland based on aquatic macroinvertebrates, defined by NR&M (in preparation).

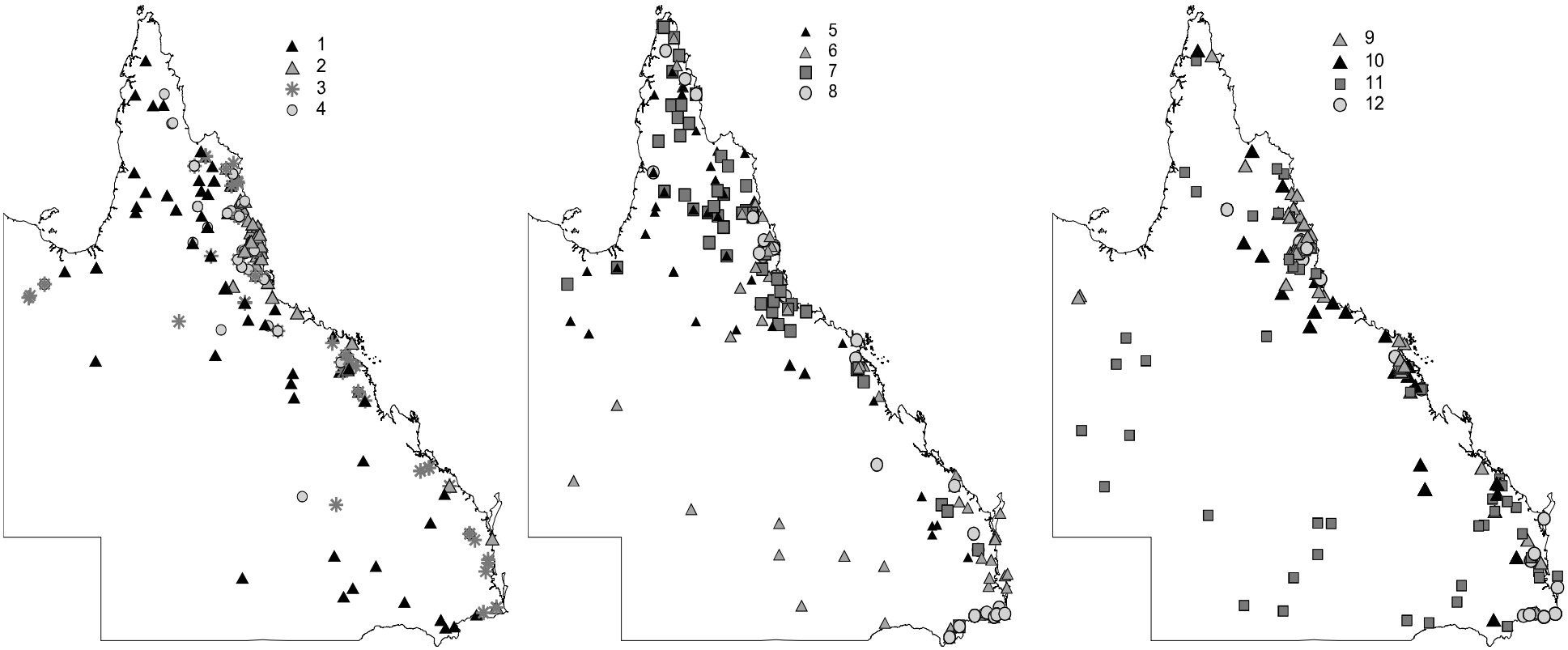
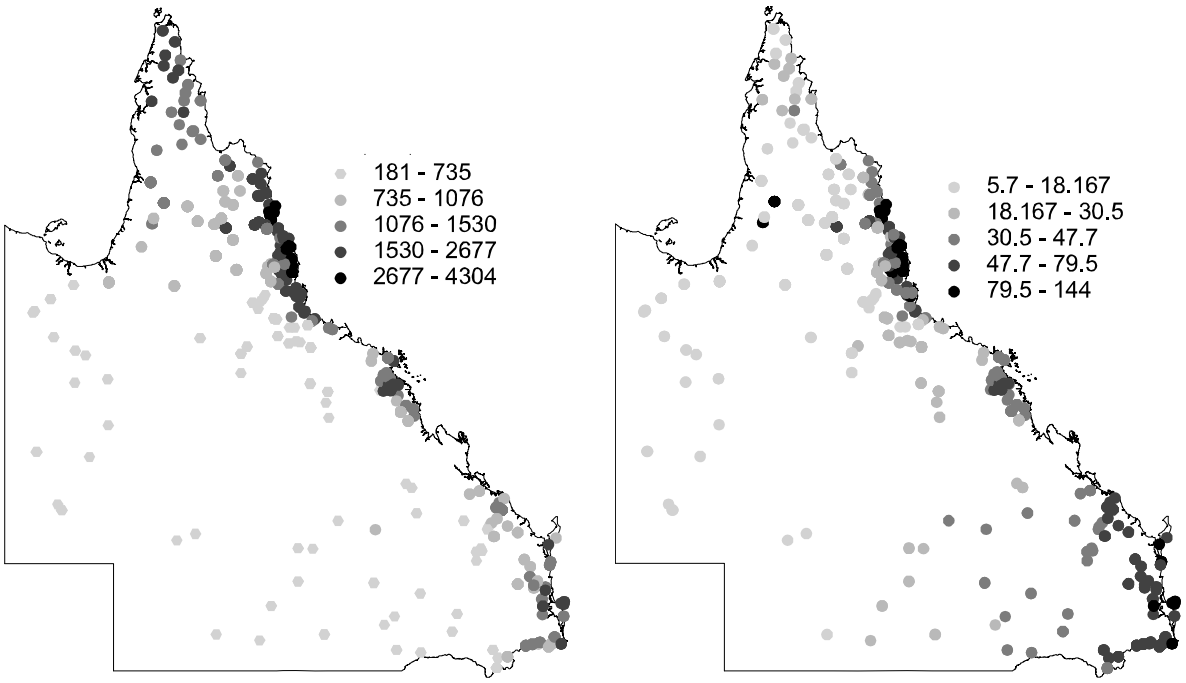
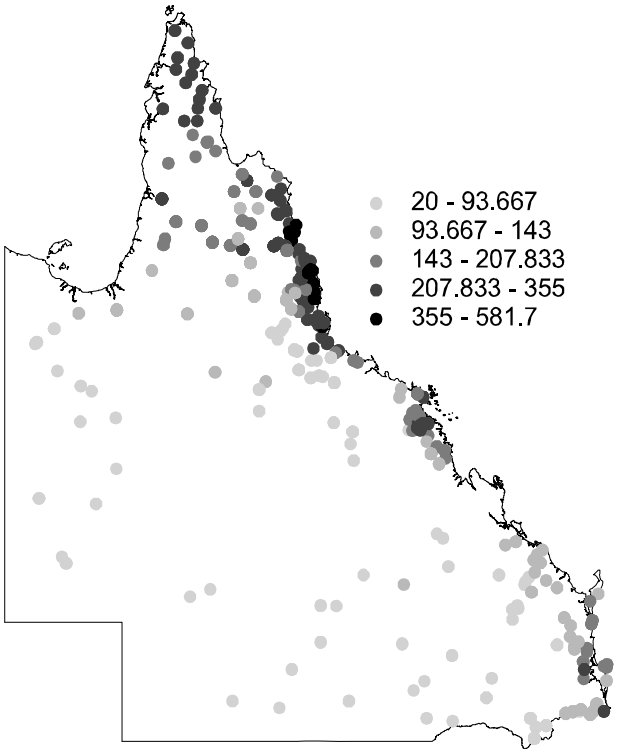


Figure 4.5. Distribution of 12 SOM defined clusters (clusters are shown in groups of 4 for readability), reference sites, only macroinvertebrates.



a)

b)



c)

Figure 4.7. Distribution of rainfall pattern in QLD: a) mean annual rainfall, b) mean dry season rainfall, c) mean wet season rainfall.

Table 4.7. Results of the univariate analysis of variables between SOM defined clusters, macroinvertebrate taxa (only first 20 are shown).

Variable	F	P
Taxonomic Richness	119.6	0.000
Hydrometridae	51.2	0.000
Baetidae	36.8	0.000
Isostictidae	31.4	0.000
Helicopsychidae	30.4	0.000
Psephenidae	29.6	0.000
Elmidae	27.6	0.000
PET Richness	27.1	0.000
Ostracoda	24.1	0.000
Cladocera	24.0	0.000
Ancylidae	22.3	0.000
Corduliidae	22.1	0.000
Notonectidae	21.4	0.000
Naucoridae	20.4	0.000
Ptilodactylidae	20.4	0.000
Hydropsychidae	19.2	0.000
Caenidae	17.2	0.000
Aeshnidae	16.2	0.000
Pleidae	16.2	0.000
s-f Chironominae	16.0	0.000

Discussion and Conclusion

In this chapter I investigated natural variability within macroinvertebrate communities using two datasets from Victoria and Queensland using SOM as clustering and visualisation tool. Component planes have been used to investigate the major environmental gradients affecting macroinvertebrate communities in two states. Slightly different approaches in combining statistical methods and SOM were used for each dataset. Both approaches provided interesting insights into the relationships between environmental factors and structure of macroinvertebrate communities.

Clusters discovered by SOM were meaningful and easily explainable (hypothesis 1). Hypothesis 2 is accepted in case of Victorian data as clusters of sites with similar macroinvertebrate assemblages discovered by SOM were largely in accordance with previously defined bioregions of state of Victoria (Metzeling et al., 2001). SOM component planes provided easy and highly visual way to assess relationship between variables and define the ones, which are likely to be of importance in shaping macroinvertebrate assemblages within each group (hypotheses 3 and 4). Discriminant analysis has been used to provide further, more quantitative insight into the relative importance of environmental factors for structuring the macroinvertebrate communities.

In the case of Queensland dataset SOM model of occurrence pattern of 69 macroinvertebrate taxa produced largely meaningful and explainable clustering. However, the pattern discovered only vaguely resembled biological regions defined by NR&M, so in case of Queensland data hypothesis 2 is not true. This can be explained by differences in the methods and data used to define the bioregions. I used

only riffle habitat and only common taxa when bioregions were outlined using data from all habitats and all taxa sampled, plus bioregions were identified using statistical clustering techniques (unpublished results). Still, SOM identified changes in community composition from south to north and from coastal areas to inland, the major directions in the distribution of bioregions.

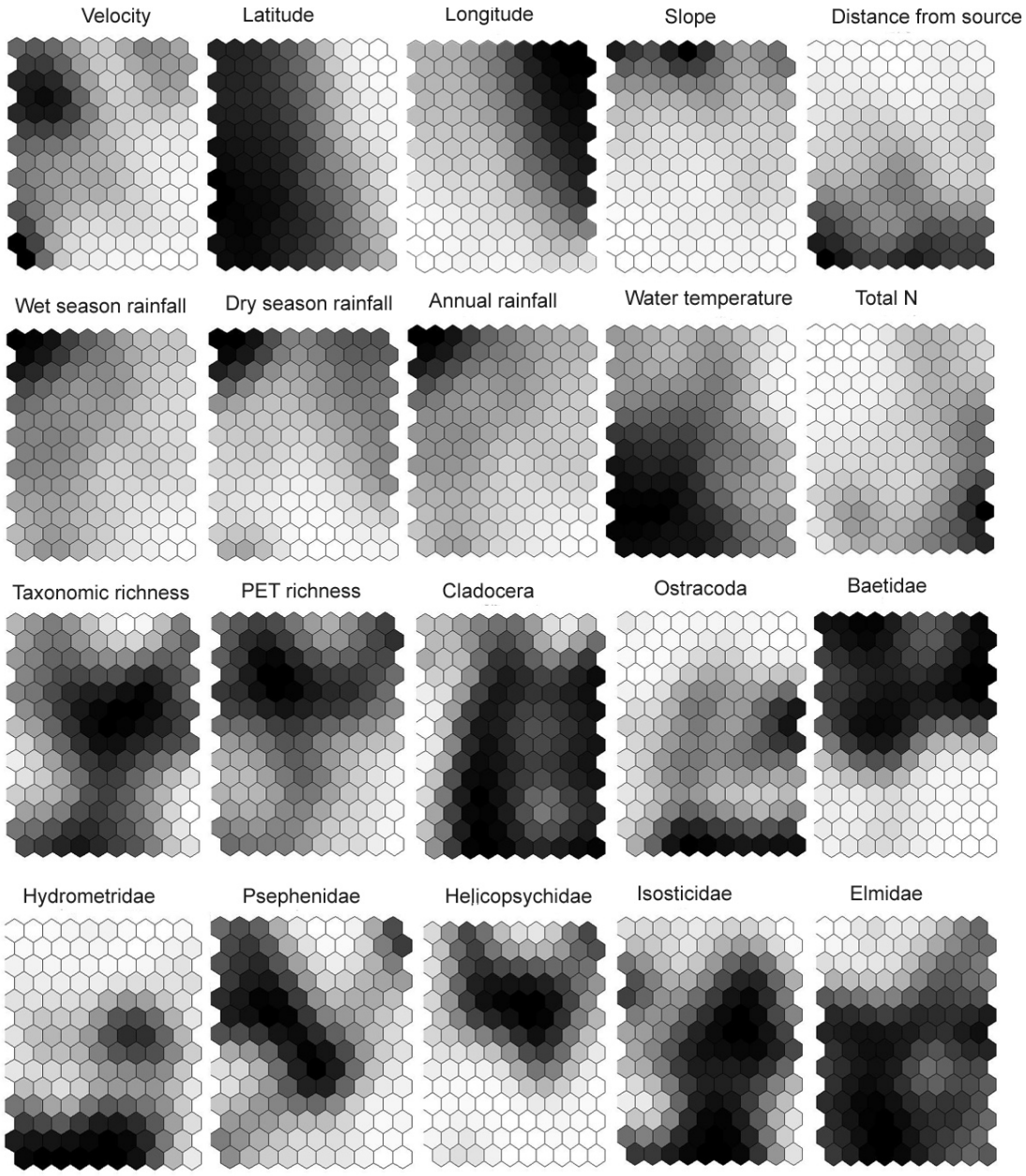


Figure 4.8. SOM component planes of the first 10 abiotic and first 10 biotic variables best discriminating between SOM defined clusters (all variables normalized between 1 and 0).

Three main environmental gradients have been identified using ANOVA and comparison of SOM component planes namely: geographical location, rainfall pattern

and distance from source. Water quality, temporal variables and substrate composition appear to be less important but still significant in discriminating between types of macroinvertebrate communities defines as SOM clusters. The structure of macroinvertebrate community is affected by all those gradients, which in combination can produce infinite number of possible environmental conditions. However, we need to keep in mind that the environmental variables provide explanation only to a certain degree. Biotic interactions such as competition and predation are the other significant factors structuring macroinvertebrate communities. Bun and Davies (2000) showed that an assumption that changes in macroinvertebrate communities are always reflection of changes in environmental factors overrides importance of biological processes and can lead to erroneous conclusions about causes of these changes. For example, Power (1990) showed that in the presence of predatory fish, smaller predators were reduced, tube-weaving chironomids larvae proliferated, and the benthic substrate was reduced to a midge-infested residue. Predation effect of this kind can be a major cause of spatial and temporal variation in stream community structure. Biomonitoring models based entirely on abiotic variables would be unable to predict such marked changes in the nature of the stream. Predator-caused shift in the community structure can be mistakenly attributed to some form of anthropogenic disturbance. Modelling biotic interactions within macroinvertebrate communities can be an interesting direction for the future research.

4.2 Exploring relationships between environmental variables and diversity of stream biota in four NSW catchments

Introduction

When an ecologist is presented with a new dataset for the analysis the first logical step is to have a general view of the possible major relationships between variables. Traditionally this is done using scatter plots, box-plots, various correlations and spatial distributions on GIS maps. SOM component planes are an interesting and underutilised method for visualising relationships between variables. It provides an opportunity to see the possible correlations, no matter whether linear or non-linear and occurring on broad or local scale. It also allows visualising spatial and temporal gradients using the very simple and quick procedure of building SOM and choosing the option of viewing the input variables as component planes, which is provided by the majority of software packages implementing SOM.

Traditionally, component planes are visualised using hexagonal or square cells coloured in accordance with the mean value of the variable in question in each cell. The cells are arranged in two-dimensional grid resulting in variable coloration of the whole grid in accordance with the spatial or temporal pattern. The information reflected by the colour can be quantified as mean value within cell, however, this is not particularly useful in the initial stage of the data analysis where we are mostly interested in broad general patterns, identifying areas containing extreme values or risk values, areas or high biodiversity values, etc. Visual information provided by

colouring patterns of component planes is often enough to identify possible relationships, which could be further explored using a variety of methods.

This chapter aims to demonstrate the use of the SOM component planes on dataset provided by Bruce Chessman (Land and Water Conservation, NSW) and collected in four catchment of NSW as a part of MARA survey (see Chapter 3 for the data description). Even though, NSW data is somewhat limited in term that occurrence pattern or abundance of individual taxa was not provided, it is still very interesting as it contains biotic variables reflecting not only taxonomical richness of macroinvertebrate communities but also richness of native fish, macrophytes and diatoms, plus a range of environmental variables including water chemistry. This provides an interesting opportunity to use SOM component planes to gain an insight into relationships between both environmental factors and biotic variables.

The main hypotheses for this study are:

1) SOM component planes is a very useful tool for the initial analysis of data containing a number of biotic and abiotic variables.

2) It is possible to make certain assumptions about the environmental status of certain areas and suggest hypotheses for the further testing using the SOM component planes only.

3) SOM component planes allow detecting both linear and non-linear relationships between variables.

Material and methods

I used a total of 22 variables (Table 4.8), both biotic and abiotic to demonstrate use of SOM component planes for understanding complex and often non-linear relationships in ecological datasets. In this case we are not interested in clustering but visualising component planes as an initial stage of the data analysis. For this purpose I used 22 variable as an input for SOM. Standard procedure with default parameters offered by SOM toolbox for Matlab was used to build, train and estimate the quality of SOM. All data was normalised prior the analysis using “log” transformation.

Table 4.8. Description of the variables from NSW dataset.

Variable	Abbreviation	MIN	MAX	MEAN
Site elevation (m ASL)	Elev	5.00	780.00	268.16
Site slope (m/km)	Slope	0.10	88.89	8.52
Site discharge (m^3/s)	Flow	0.00	6.13	0.28
Average of maximum and minimum stream width per quadrat (m)	Width	0.22	44.38	7.27
Average of maximum stream depth per quadrat (m)	Depth	0.01	3.19	0.72
Number of diatom species per site	DiatomSp	6.00	78.00	38.31
Number of native aquatic macrophyte species per site	MacrNaSp	2.00	29.00	11.17
Number of native macroinvertebrate families per site	InveNaFa	13.00	59.00	31.12
Number of native fish species per site	FishNaSp	0.00	9.00	2.17
Macroinvertebrate family biotic index (SIGNAL 1995 version) (range 1-10)	SIGNAL95	4.35	7.02	5.42
Number of native fish individuals per hour of electrofishing	FishNNPH	0.00	1630.00	201.16
Water temperature at 0.2 m (C)	TempSur	6.40	38.00	20.14
Turbidity at 0.2 m (NTU)	Turb	0.40	64.70	12.30
Electrical conductivity at 0.2 m (uS/cm)	EC	33.00	2330.00	370.45
pH at 0.2 m	pH	4.42	8.70	7.42
Ammoniacal nitrogen at 0.2 m (mg/L)	NH^3	0.01	1.60	0.06
Oxidised (nitrate plus nitrite) nitrogen at 0.2 m (mg/L)	NOx	0.01	1.00	0.05
Filterable phosphorus at 0.2 m (mg/L)	FiltP	0.00	0.85	0.03
Bank erosion score (range 0-100)	Erosion	0.00	96.43	8.37
Number of alien fish individuals per hour of electrofishing	FishANPH	0.00	4736.67	273.12
Stock damage score (range 0-100)	Stock	0.00	78.13	13.64
Catchment area above site (km)	CatArea	1.00	1815.75	231.75

Results

The parameters of the resulted SOM were: map size 9x 6, final quantization error: 0.54, final topographic error: 0.017. Figure 4.9 shows hit diagram showing location

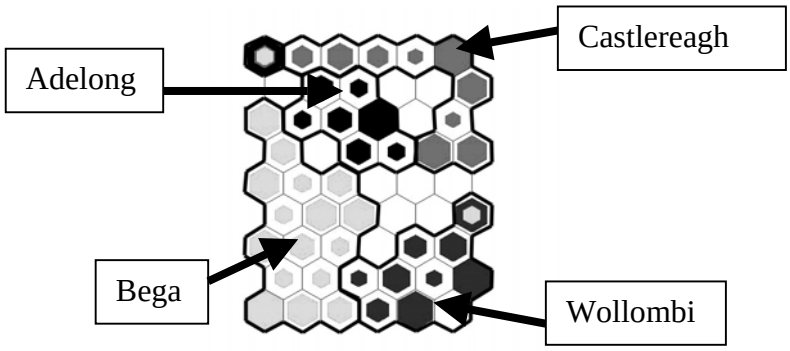


Figure 4.9. Hit diagram, hits for four catchments shown in different color.

of the catchment specific subsets on the general SOM. It is obvious that all four subcatchments are quite different from each other with very little overlap between them.

Figure 4.10 shows component planes for the physical properties: elevation, slope, flow, width, depth, catchment area above the site and risk factors as: conductivity, pH, NO3, stock damage score, NOx, phosphorus, erosion score, turbidity and water temperature at the surface. Castlereagh and to some extent Adelong catchment are characterised by higher elevations than the other two catchments. Width and depth are generally high in Bega and Wollombi, with some deep and wide sites located in Castlereagh. Some sites in Bega and Castlereagh have relatively high flow, when in Wollombi flow is mostly uniformly low.

When we look further at the risk factors (Figure 4.10) and biotic variables (Figure 4.11) it becomes clear which areas are most likely experiencing some kind of anthropogenic impact. To provide some comparison between the information extracted only from study of component planes by a researcher totally unfamiliar with an area and reality we inserted short description of each subcatchment taken from "Assessing the conservation value and health of New South Wales rivers. The PBH (Pressure-Biota-Habitat)" report prepared by Bruce Chessman, Land and Water Conservation, NSW, 2002.

Castlereagh

Component planes:

Most of the subcatchment is characterised by relatively high elevation with some slopy sites. Flow is generally low. Sites are generally not wide, depth is variable. The potential problems with water quality indicated by very high phosphorus, erosion at places and turbidity in some parts. Conductivity is highly variable indicating potential problems with salinisation at some places, pH mostly high. Temperature is generally not too high, indicating that at least some parts of the catchment have a good riparian vegatation cover. Variable, and generally not high number of native macroinvertebrate species which seems to be decreasing along the gradient of increasing turbidity, which is variable throughout the catchment. SIGNAL95 index vary (seems to be decreasing in areas with high turbidity as well). Very low number of native fish species, but fish abundance is high in some places. Generally not too high abundance of introduced fish. The area is most likely quite heterogeneous with some sites in good condition and some are significantly degraded. The most likely problems are water extraction, salinity, turbidity, nutrients (phosphorus) and erosion.

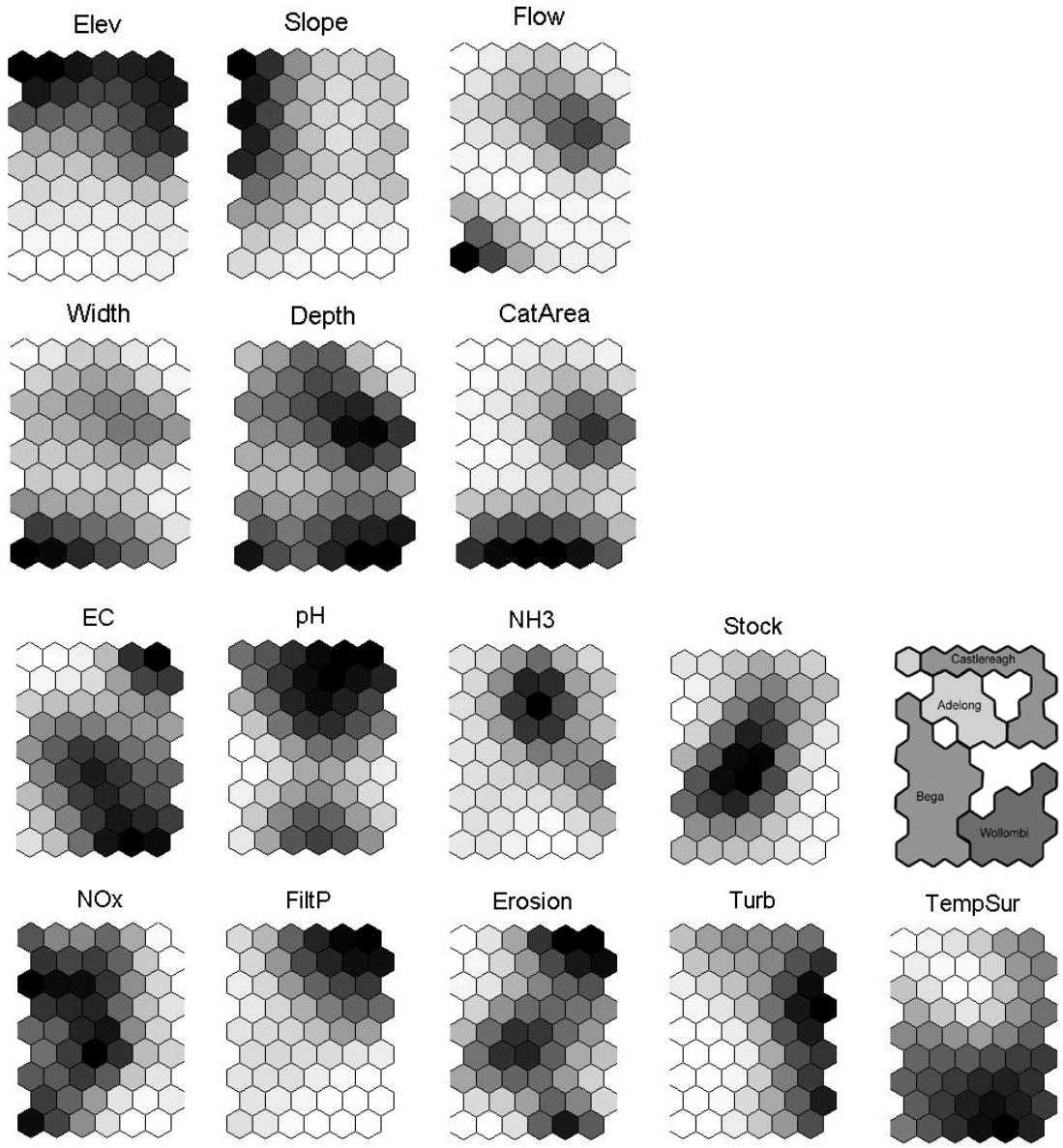


Figure 4.10. SOM component planes for the natural settings and risk factors in NSW dataset.

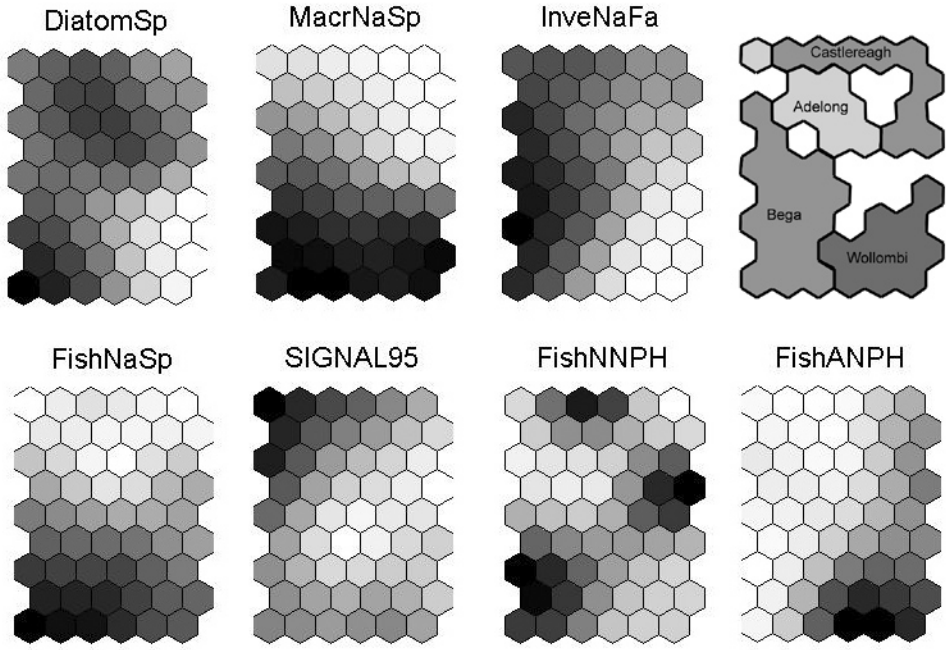


Figure 4.11. SOM component planes for the biotic variables in NSW dataset.

From Chessman, (2002):

"The Castlereagh River rises on Westons Mountain in the east of the Warrumbungle Range at the elevation of about 900 m. Most of the Castlereagh subcatchment has been cleared for grazing of sheep and cattle, but substantial areas of native forest remains, especially in the west. The Castlereagh River is impounded approximately 9 km downstream of its source by Timor Dam. Castleagh river above Binnaway was classified as having a high level of both water extraction and environmental stress. Riparian vegetation was rated in the medium stress category and both channel stability an in-stream structures in the high stress category."

Adelong

Component planes:

The subcatchment is characterised by medium to high elevation, mostly low flow but some streams can be wide and many are comparatively deep. Generally low conductivity and high pH, high level of nutrients, NH^3 and NOx almost everywhere and phosphorus at some places. Erosion and turbidity are probably not a major concern. Mostly low water temperature indicating a good riparian vegetation cover. Number of macroinvertebrate species, native fish species and abundance of native species are low to medium, but uniformly low abundance of introduced fish species. Surprisingly high diversity of diatom species.

From Chessman, (2002):

"Most of the Adelong subcatchment has been cleared for grazing of sheep and cattle, but extensive radiata pine forests and scattered orchards lie in the south. In the Stressed River assessment Adelong was characterised by high level of water extraction and a medium level of environmental stress. Poor environmental ratings were given for a bank stability index, the level of development, percent of the sub-catchment with non-conservation uses, a tree shortfall index and total phosphorus level in Adelong Creek. Good ratings were given for riparian vegetation cover, width, stream bed condition and salinity and very good ratings for lack of overcropping and for stream pH."

Bega

Component planes.

The area with high variability regarding to both geomorphological and water quality factors. Low elevation, variable flow, very high to very low. Some sites are very wide and deep. Some sites are characterised by high conductivity but mostly conductivity is low to medium. Low pH indicates possible acidification problems. Many sites sustained heavy stock damage. Mostly high level of NOx, but phosphorus is low, so the turbidity. Heavy erosion at places. It is likely that not many sites have a good riparian vegetation cover as water temperature is medium to high, exceeded only by sites at Wollombi subcatchment. Medium to high number of macroinvertebrate species and diatoms, same for the number of macrophyte species, however SIGNAL is mostly low to medium. Very high diversity and abundance of native fish species at some sites, and very low abundance of introduced species. Most likely problems are water extraction, acidification, nutrients, stock damage and clearance of riparian vegetaion, however, et least some parts of the subcatchment are in good condition as high diversity and high abundance of native fish are observed.

From Chessman, (2002).

"About half of the Bega River catchment is forested, with the remainder cleared for grazing. Two tributaries are impounded. Water extraction was considered from low to high in various parts of the sub catchment, leading to unresolved management classification. The various parts of the subcatchment is variable in regards to other risk factors with many of them experiencing degradation of riparian vegetation and bank and bed instability, however, high species diversity and threatened fish species have been found throughout the subcatchment."

Wollombi

Component planes

Flow and conductivity are two most obvious risk factors in Wollombi subcatchment. Elevated nutrients are found only at some sites, however, there are many sites with high turbidity and erosion is also high at some places. Very high water temperature

indicating that most of the subcatchment is most likely cleared of the riparian vegetation. Very low diversity of native macroinvertebrates and diatoms, variable diversity of native fish, averaging to medium, high at some sites. Very high abundance of introduced species, and uniformly low abundance of native species. This all indicates that this subcatchment is probably experiencing the highest degradation in comparison with previous three with major problems as water extraction, clearance of riparian vegetation, sedimentation (turbidity) and possibly secondary salinisation.

From Chessman, (2002).

"Wollombi Creek characterised by a high level of both water extraction and environmental stress. The primary environmental stress factors listed were bank instability, degradation and sedimentation of the stream bed and macroinvertebrates in poor condition, however high diversity of fish species is found there contributing to conservation value of the subcatchment."

The exercise above shows that is it possible to characterise geographical areas just by using SOM component planes and make an initial assumptions which can be further tested. It is also possible to detect possible correlations between variable and suggest the causal relationships. For example, we can see from the component planes that number of native fish species and elevation appear to be in negative relationship, and this relationship appears to be linear. The Product –Momentum correlation between these two variables is: $R=-0.56$, $p<0.05$, and has been observed that diversity of native species in NSW is negatively correlated with altitude and tends to be the highest in the lowland streams (Chessman, 2002).

Number of native macroinvertebrates appears to be low at the areas with high turbidity, high water temperature and high abundance of introduced fishes. R between turbidity and macroinvertebrate diversity is -0.35 , $p<0.05$, between water temperature and macroinvertebrate diversity is -0.28, $p<0.28$. We can assume that in case with water temperature relationship is likely to be unimodal as the highest diversity of macroinvertebrates is found in the areas of medium water temperature and decreasing towards extemes. Indeed, the scatterplot shown on Figure 4.12 confirmes this trend. In case of turbidity relationship is likely to be non-linear (logarithmic), as diversity of macroinvertebrates decreasing not uniformly along the turbidity trend. If we examine scatterplot (Figure 4.12) we can see that there is an obvious non-linear relationship between these two variables.

Discussion and conclusion

The aim of this study was to demonstrate use of SOM component planes for the analysis of datasets containing a number of biotic and abiotic variables. Component planes provide quick and simple method for evaluation of the geographic distribution, possible gradients and correlations in the data (hypothesis 1). It has been shown that it is possible to quickly characterise the distribution and the extent of anthropogenic impact in various areas with reasonable accuracy and define possible direction for the further data analysis (hypothesis 2). It is also possible to detect possible non-linear relationships between variables and suggest the nature of the relationship (as linear,

unimodal, etc.), which has been demonstrated using example of fish diversity and elevation (hypothesis 3), macroinvertebrate diversity versus turbidity and water temperature.

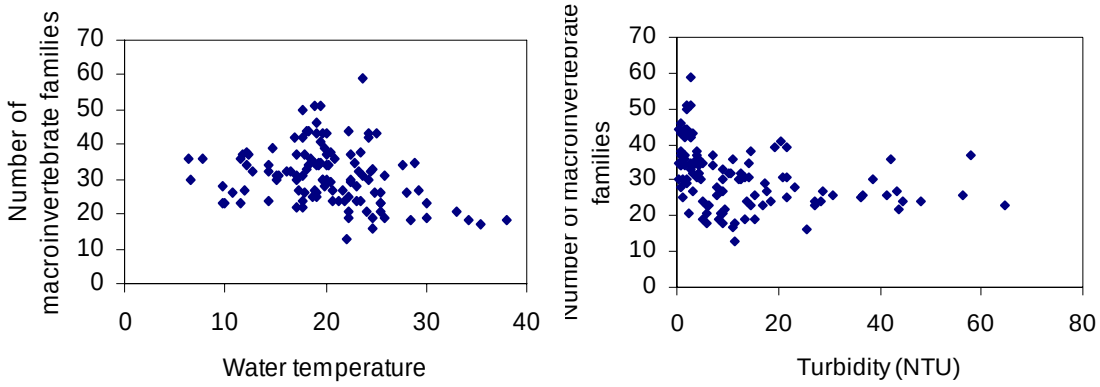


Figure 4.12. Scatterplots for the number of macroinvertebrate families versus water temperature and turbidity.

Use of component planes provides flexible, highly visual and easily understandable approach for the data mining and exploration. Closer look at the component planes for NSW dataset is likely to provide much more information than has been described here. For example SOM built for each subcatchment would provide more spatial resolution, and possibility to detect gradients on a finer scale, but for the purpose of this study we consider it not necessary.

4.3 Exploring the effect of water quality on trophic structure of the macroinvertebrate communities using SOM in combination with Canonical Correspondence Analysis (CCA)

Introduction

An understanding of the ecological processes within streams and rivers is an important prerequisite for the development of management strategies and guidelines. Analysis of the trophic structure based on functional feeding groups (FFG) is commonly used to assess the health and integrity of aquatic ecosystems. FFG include taxa of the same feeding type, using similar sources of energy and matter. There are different partitioning systems and approaches to the division of macroinvertebrates into functional feeding groups with the 5 main categories most commonly recognising (Cummins, 1973; Chessman, 1986):

- Shredders feed on living or decomposing plant tissues, including wood, which they chew or gouge;

- Collectors feed on a fine particulate organic matter by filtering particles from suspension or fine detritus from sediment;
- Scraper or grazers feed on attached algae and diatoms by grazing solid surfaces;
- Predators feed on cells of living animal tissues by engulfing and eating the whole or parts of animals or piercing prey and sucking body fluids;
- Filterers strain the water column for fine organic matter such as micro-algae and some tiny fragments of plant and animal material.

Invertebrate functional group analysis is known to be sensitive to natural geomorphic features and anthropogenic disturbances that occur along a river continuum (Vannote et al., 1980; Cummins, 1973).

The River Continuum Concept (RCC) relates the sources of energy inputs into the aquatic system with the aquatic biota that inhabit that system. In the headwaters of a catchment, streams are often heavily shaded and light is low, photosynthesis is restricted and energy derives from high inputs of materials such as leaves, woods etc., shredders tend to predominate, because they can break up large matter into finer particles. Downstream from the headwaters of a catchment, collectors are often found to dominate community structure as they are able to filter the fine particles generated upstream and themselves add particles (faeces for example) to the current. Where the waterway becomes broader, with increased available light allowing photosynthesis in the middle reaches, algae, diatoms and macrophytes develop and serve as food for grazers. Predators tend only to track the localized abundance of food resources (Vannote et al., 1980).

Although a valuable theoretical model, the RCC was developed for northern hemisphere and does not seem to apply to many Australian streams (Boulton and Brock, 1999). One of the reasons for this is that the leaf litter is relatively tough and shredders are rare. Stream flow is highly variable and regimes of flooding and disturbance appear to be more important in structuring riverine communities. Lake et al. (1986) suggest that in Australia, patterns in feeding group representation are likely to be more complicated than the RCC suggests.

When in good condition communities of macroinvertebrates are known to have high diversity and stable dynamic of proportional distribution of FFG for the given natural conditions. Disturbances caused by anthropogenic influences as contamination, suspended solids, nutrients load often result in a decrease in diversity within macroinvertebrate assemblages and changes in the structure of communities including changes in the proportional composition of trophic groups. Widely recognised changes in FFG associated with human activities include reduction in shredders with loss of riparian habitat, and consequent reduction in autochthonous inputs. Another possible change reflecting anthropogenic impact is an increase in grazers with increased periphyton (algae and diatom) resulting from enhanced light and nutrients entry.

I am aware of only one application of SOM to investigate the spatial distribution of macroinvertebrate functional groups. Cereghino (in press) used SOM to model differences in macroinvertebrate functional structure among microhabitat types in low-ordered streams in Southwest France. Water depth, current velocity, substratum,

and particulate organic matter were used to characterize the microhabitats of macroinvertebrates partitioned into functional feeding groups (FFGs) based on the nature of the food typically ingested and the feeding behaviour. Four clusters was identified on the SOM map (k-means algorithm) according to eight habitat variables, this classification being chiefly related to depth and mean current velocity, and to the size of the mineral particles. Similarly, four subsets were derived from the SOM according to the proportions of the various FFGs. Gathering-collectors and predators dominated in deeper areas, with cobbles and pebbles subjected to high current velocities. Shredders, filtering-collectors, and grazer-scrapers dominated in sandy to stony areas, at low depth and current velocity. Correlation coefficients between observed and predicted values of each FFG were highly significant. The percentage of gathering-collectors was negatively correlated with the percentage of all other FFGs, for both observed and predicted data. Significant relationships were also obtained between shredders and grazer-scrapers, and between predators and grazer-scrapers for predicted data only. On a local scale, different areas were the template for different ecological functions. Energy and habitat use by FFGs may be regulated by the patchiness of the habitat mosaics, and, subsequently the microdistribution of FFGs can be related to habitat template theory.

The current study aims to explore the relationships between trophic structure of macroinvertebrate communities and water quality factors as temperature, dissolved oxygen, turbidity, conductivity, pH and nutrient concentrations using Self Organising Map (SOM) neural networks as described by Kohonen (1995). In order to provide more insight into possible structuring forces for the assemblages with different trophic structure we used a combination of SOM and Canonical Correspondence Analysis (CCA).

CCA is a popular method for relating environmental gradients to species distribution. It was introduced by ter Braak in 1995 as an extension of commonly used Canonical Analysis (see Chapter 2 for more on CCA). Output from CCA algorithm includes axis scores for sampling units and species and vectors representing the correlations between the environmental variables and principal axes can also be included on these plots, creating a biplot (Quinn and Keough, 2002). Traditionally sampling units or species as centroids of all sampling units where the species has occurred are visualised in biplot with vectors representing gradients in environmental variables. However, trying to analyse large datasets this way can be very complex and visualising all sites will produce very busy and ultimately not informative plots. One possible solution to this problem is to reduce dimensionality of the data prior to application of CCA by clustering sites in similar groups and use these groups instead of sites as an input to CCA. In this chapter I explore this option by clustering sites on the basis of their similarity in regards to trophic structure using SOM and then apply CCA to the resulting clusters in order to explore the input of various environmental factors into structuring of these clusters.

The main hypotheses for this study are:

1) RCC is not applicable for the Queensland stream systems.

2) Water quality variables affect trophic structure of macroinvertebrate communities. Different proportions of FFG are likely to be found in the areas with different water quality.

3) Using SOM it is possible to identify typical FFG structures and explain them using environmental variables.

4) SOM component planes are useful tool for the understanding relationships between FFG and water quality variables.

5) It is possible to extract additional information by using combination of SOM and CCA.

Data and methods

NR&M dataset was used for this study, including two habitats: riffle (1333 samples) and edge (2442 samples). Only common taxa found in more than 5% of sites were used (69 taxa in total).

Each taxon (mostly family level) was assigned a number of functional feeding group (as described by Chessman, (1986)). Five functional groups were used: collector, grazer, predator, filterer and shredder. Some taxa (like Leptophlebiidae) include species with membership to more than one functional group and were considered as Leptophlebiidae 1, Leptophlebiidae 2, etc. The resulted total number of taxa/functional group considered for the analysis was 77. We calculated the percentage of taxa belonging to each functional group relative to the total number of taxa present in each sample. The eight water quality variables and seven geoclimatic variables were used for the analysis.

The first SOM was built in order to find possible correlations between the functional groups themselves and water quality variables using component planes. Component planes show the value of one variable in each map unit. For this purpose we used both percentage values of 5 functional groups and 8 water quality variables as an input. The second SOM was built specifically for the purpose of clustering data matrix containing percentage values of FFGs in order to find similar spatial patterns and relate them to the water quality variables. For this purpose we only used FFGs as an input. SOM was partitioned into a minimal optimum number of clusters using k-means partitioning algorithm implemented in SOM toolbox for Matlab. The resulting clusters were then analysed using Mahalonobis distances between groups and box and whiskers plots.

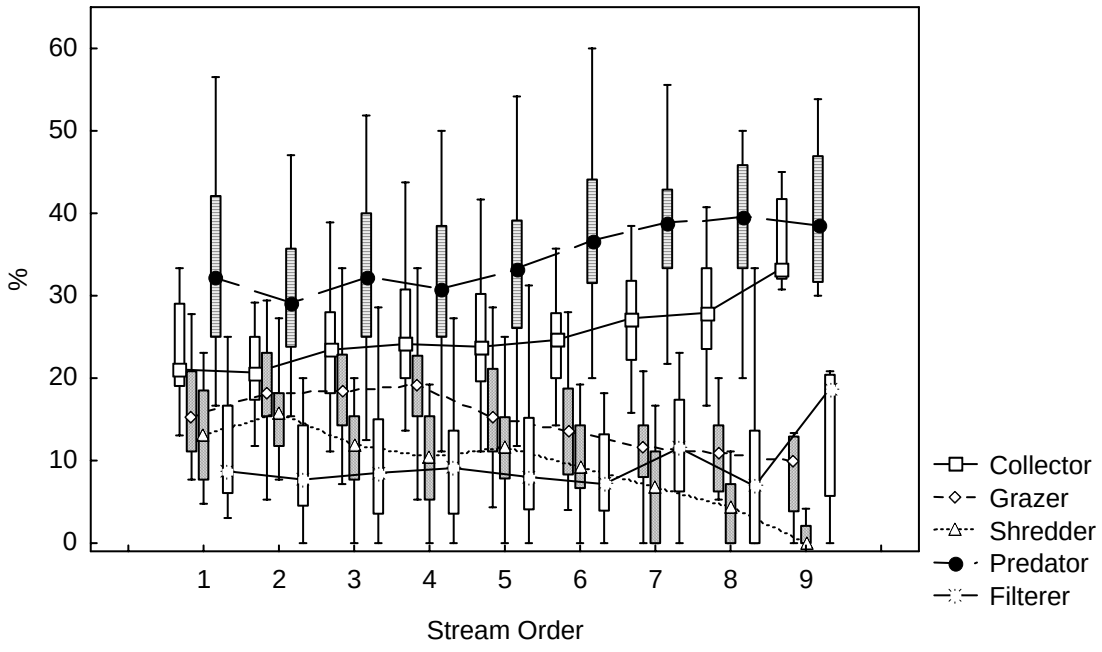
For the first SOM, data was log-transformed to avoid dominance of large values of conductivity and turbidity. We did not transform data for the second SOM as all values were expressed as percentage values between 0 and 100.

In order to provide more insight into the relationship between trophic structure and water quality we applied CCA to the data matrix containing SOM defined clusters and environmental variables. Significance of axis and each variable has been evaluated using Monte Carlo test (999 permutations) and only significant variables ($p < 0.05$) were used for the final model.

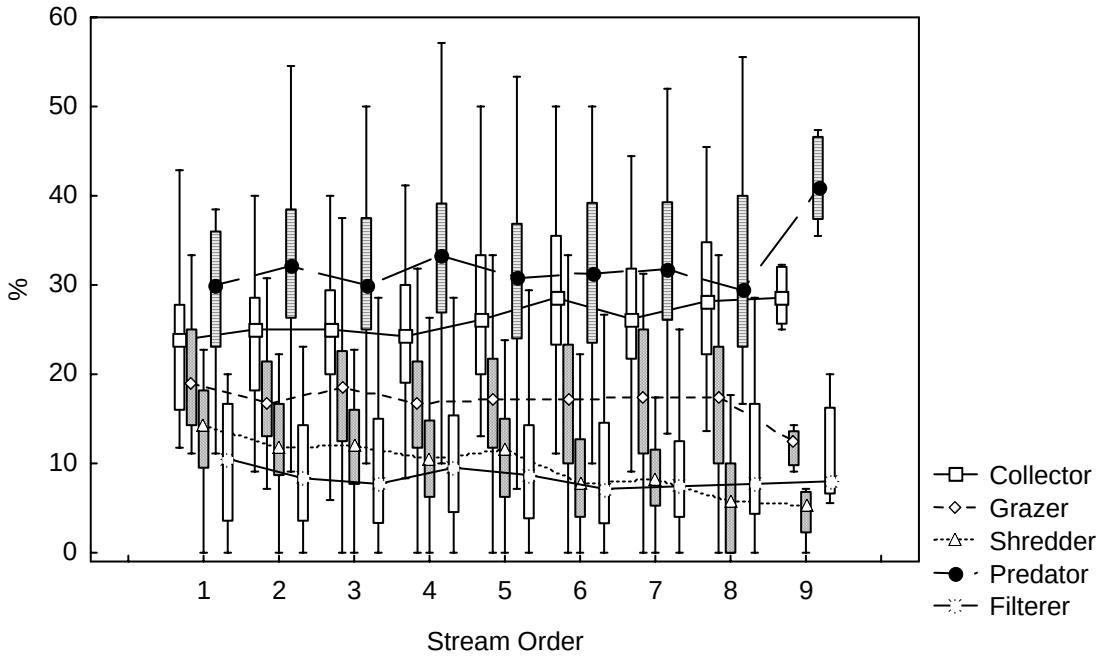
Results

Succession of FFGs along the stream order gradient in reference and test conditions.

Figure 4.13 shows box plots for the distribution of proportional FFG along stream order gradient for reference and test sites in the riffle habitat. In general, under the close to natural conditions in Queensland proportion of collectors and predators steadily increase downstream. Proportion of grazers slightly increases in mid-order streams and decreases further down and proportion of shredders decreases along the gradient which in general terms agrees with RCC. Filterers appear to become slightly more common in lowlands but this trend is not very pronounced. These trends become much less obvious when the wide range of sites in different conditions is considered. Shredders appear to be following the natural pattern to a certain extent and percentage of collectors slightly increases downstream, otherwise there is no distinctive pattern evident. This suggests that water quality is a contributing factor into the changes in the natural succession of FFG along the stream order gradient.

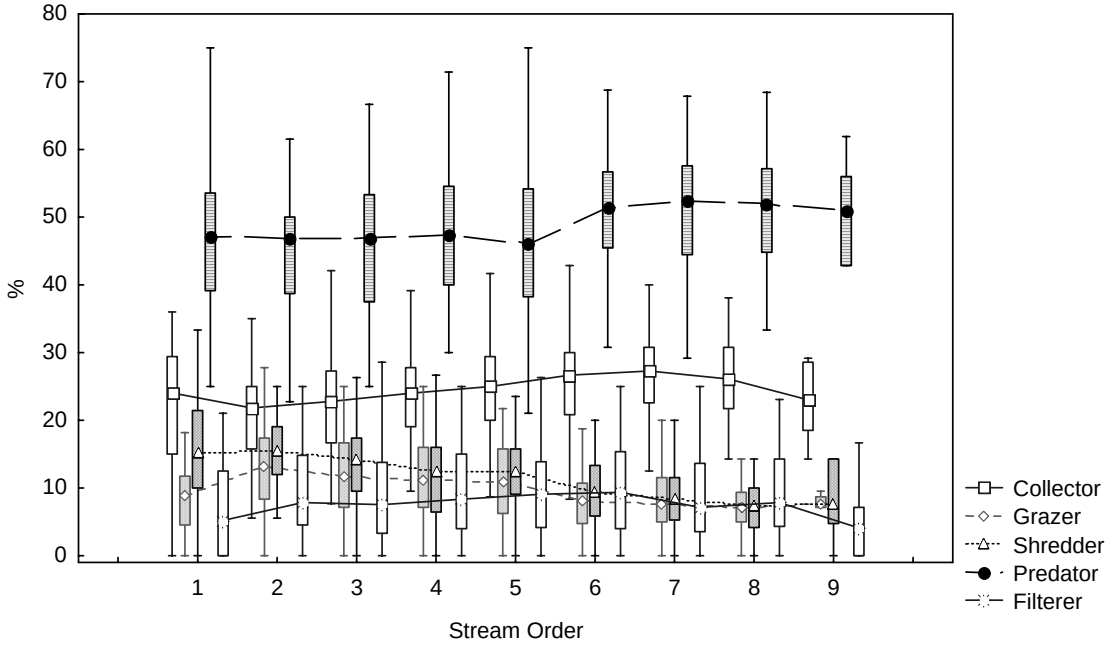


a)

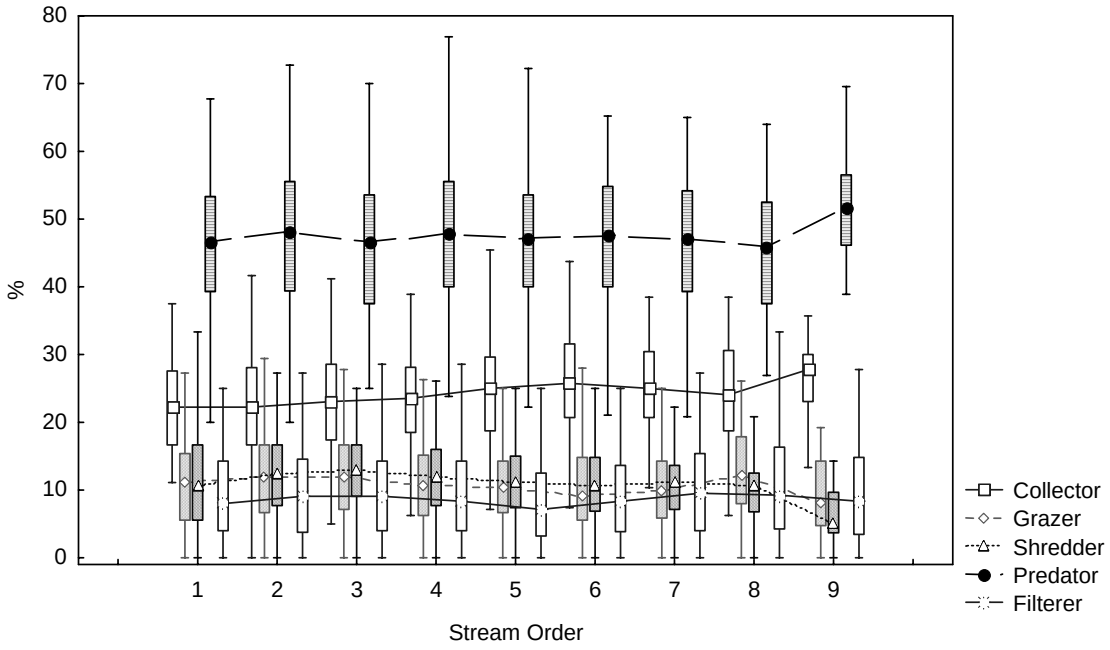


b)

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.



a)



b)

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.

In case of the edge habitat there is slight decrease in the proportion of grazers and shredders along the stream order gradient and uneven increase in collectors and predators in reference sites, but much less pronounced than at the riffle habitat. In case of test sites succession of FFG does not appear to follow any particular trend (Figure 4.14).

Using SOM component planes for visualisation of the relationship between water quality variables and feeding functional groups

Riffle

The specifications of the SOM built and accepted for this task were: size 17x11, final quantisation error: 0.25, final topographic error: 0.07.
Figure 4.15 shows component planes for 5 trophic groups and eight water quality variables. Darker shades represent higher values. The component planes for collectors and predators show increases in the proportion of both groups from the top to the bottom of the map, although, this trend is not so clear in the case of collectors with a patch of medium values in the top right corner. The component planes for grazers and shredders show the opposite trend with proportional values decreasing from top to bottom, whereas the filterers do not seem to follow this trend.

The component planes for turbidity, nutrients and water temperature show increases in value from the top to bottom, corresponding to increases in the proportional values of collectors and predators. Not surprisingly dissolved oxygen (DO) follows the opposite trend decreasing from top to bottom, although there are still some patches with high DO corresponding to areas with high water temperature. Conductivity, alkalinity and pH show a trend of uneven increase to the right but do not resemble a trend in any of the functional groups or other water quality variables.

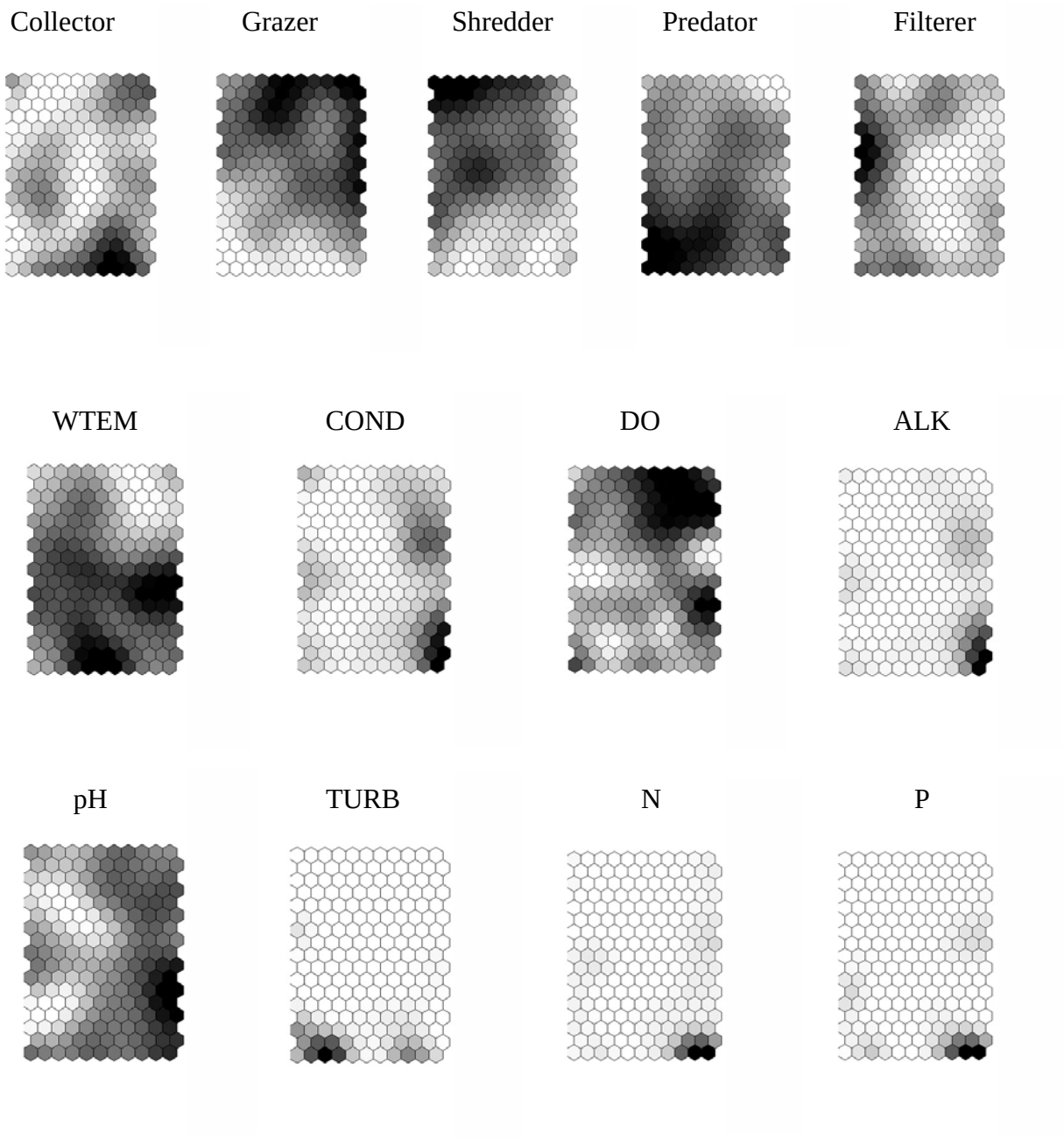


Figure 4.15. SOM component planes for the proportional values of functional feeding groups and water quality variables: water temperature (WTEM), conductivity (COND), dissolved oxygen (DO), alkalinity (ALK), turbidity (TURB), total nitrogen (N) and total phosphorus (P), see Table 1 for units.

Edge

The dataset for the edge habitat contained 2442 samples, many more then the dataset for the riffle habitat and it was quite challenging to build an SOM, which would clearly reflect pattern in this highly heterogeneous data set. SOM map accepted for this task was sized 20x13, final quantization error: 1.16, final topographic error: 0.07.

In comparison with riffle habitat patterns in distribution of FFG in relation to water quality variables reflected by SOM component planes for the edge (Figure 4.16) look more scattered and unclear. However, middle part of the right side of the planes for turbidity, nutrients and dissolved oxygen deserve particular attention. This areas with the highest concentration of sites with high turbidity and nutrients also have the high proportion of collectors and predators and low proportion of grazers and shredders, it also have comparatively low dissolved oxygen. The similar pattern was reflected by SOM component planes for the riffle habitat.

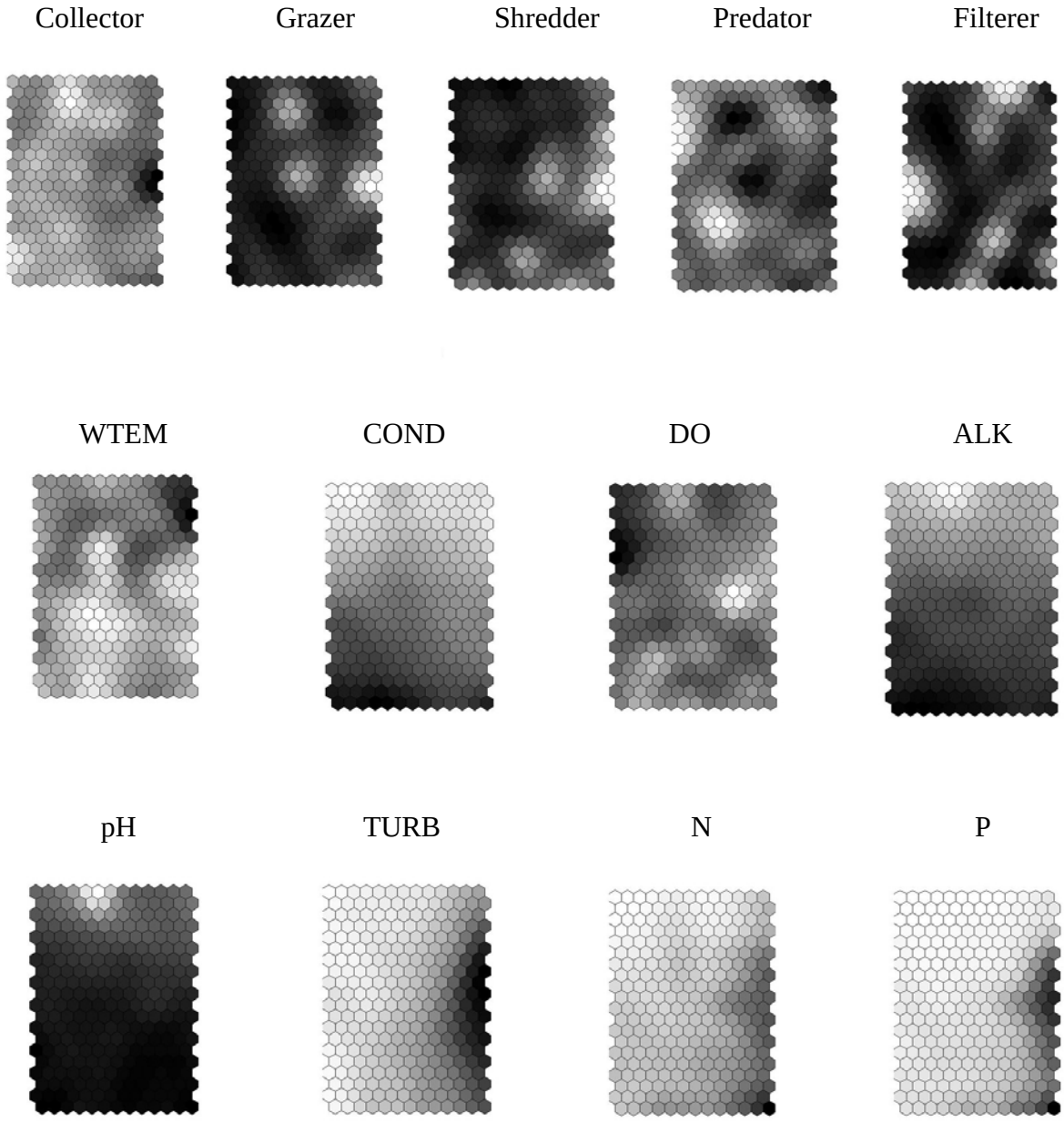


Figure 4.16. SOM component planes for proportional values of feeding functional groups and water quality variables, edge habitat.

Using SOM for clustering

Riffle

Resulting SOM sized 17x11 (final quantisation error: 6.32, final topographic error: 0.08) was partitioned into 7 clusters. Box and whiskers plots for the median percentage of FFGs in each cluster are shown in Figure 4.17. The mean values for each of the FFGs and all the variables for each of the seven resulting clusters are shown in Table 4.9.

Cluster 6 has the largest proportion of collectors and generally low proportion of all the other groups. Cluster 7 has a large proportion of collectors and the largest proportion of predators with lowest proportion of grazers and shredders. Cluster 5 has the largest proportion of grazers and the lowest proportion of predators in comparison with other clusters. Cluster 1 has the largest proportion of shredders and a large proportion of grazers. Cluster 4 is characterised by a relatively high proportion of filterers with a large proportion of predators present. Clusters 6 and 7 are generally characterised by a much larger distance from stream source than all the other clusters and are associated with greater turbidity. These two clusters are also characterised by lower rainfall (cluster 7 is lower then cluster 6). Cluster 7 also has the highest water temperature and the highest alkalinity and cluster 6 the highest values for both nitrogen and phosphorus. Conversely, Cluster 1 is characterised by the highest rainfall, lowest distance from source and lowest turbidity. Clusters 2 and 5 have similar characteristic with some variability and most likely include sites in high rainfall areas with good/moderate water quality. Clusters 3 and 4 include a variety of sites with different conditions in between and are not easy to characterise according to either water quality or the natural settings.

Table 4.9. Mean values of FFG and water quality variables for 7 SOM defined clusters, riffle habitat.

Cluster number	1	2	3	4	5	6	7
Number of samples in a cluster	204	234	206	125	170	129	265
Collector	20.16	20.31	28.35	22.80	27.57	35.60	27.31
Grazer	20.68	18.47	15.49	14.66	22.81	16.57	10.32
Shredder	15.03	10.85	10.05	10.05	12.74	8.28	5.95
Predator	28.59	35.08	32.07	38.37	21.28	22.63	42.42
Filterer	10.17	4.82	10.17	19.69	13.85	4.81	6.83
Depth	0.19	0.18	0.18	0.17	0.19	0.18	0.16
Velocity	0.65	0.63	0.60	0.63	0.73	0.63	0.58
Mean phi	-5.3	-5.17	-5.27	-4.87	-5.92	-5.58	-4.36
Altitude	181.79	164.42	162.03	177.12	167.05	179.27	174.08
Slope	0.01	0.0	0.00	0.00	0.00	0.00	0.00
Distance From Source	36.43	55.03	74.15	72.01	63.12	120.20	147.23
Mean annual rainfall	1979.75	1560.96	1402.86	1390.23	1730.54	1266.81	1000.73
Water Temperature	20.68	21.99	22.86	22.17	20.62	21.68	23.23
Conductivity	227.60	306.40	314.32	335.04	285.50	386.51	307.45
Dissolved oxygen	8.026	8.054	7.787	7.680	7.86	8.04	7.49
pH	7.34	7.48	7.48	7.48	7.44	7.60	7.49
Turbidity	4.50	6.18	6.51	8.17	9.3	15.74	34.93
Alkalinity	64.63	71.25	85.72	82.21	72.34	89.60	110.76
Total N	0.35	0.40	0.43	0.34	0.49	0.85	0.44
Total P	0.03	0.04	0.0	0.02	0.03	0.09	0.05

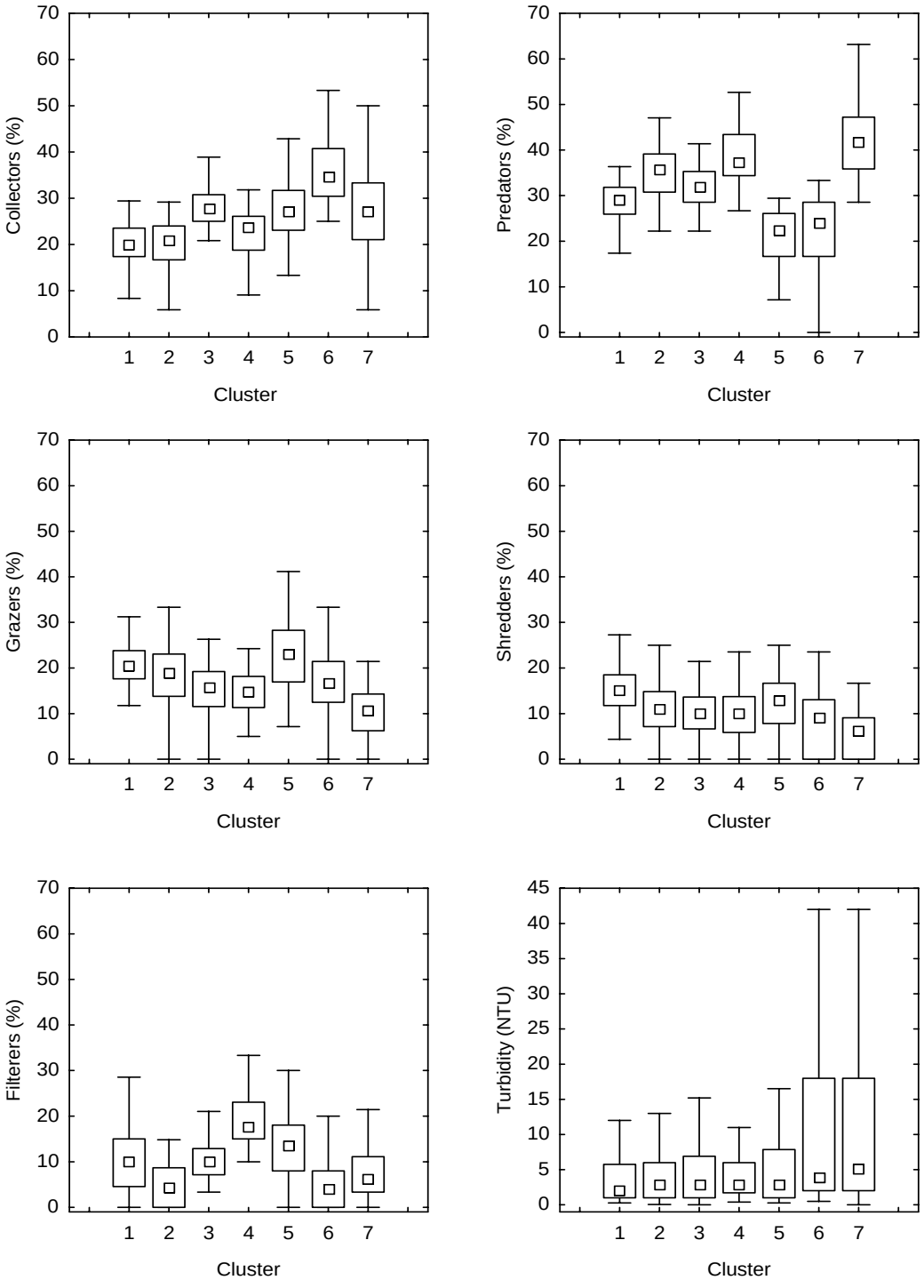


Figure 4.17. Box and whiskers plots for median values of FFG and turbidity. Box – 20-80%, whiskers – minimum and maximum.

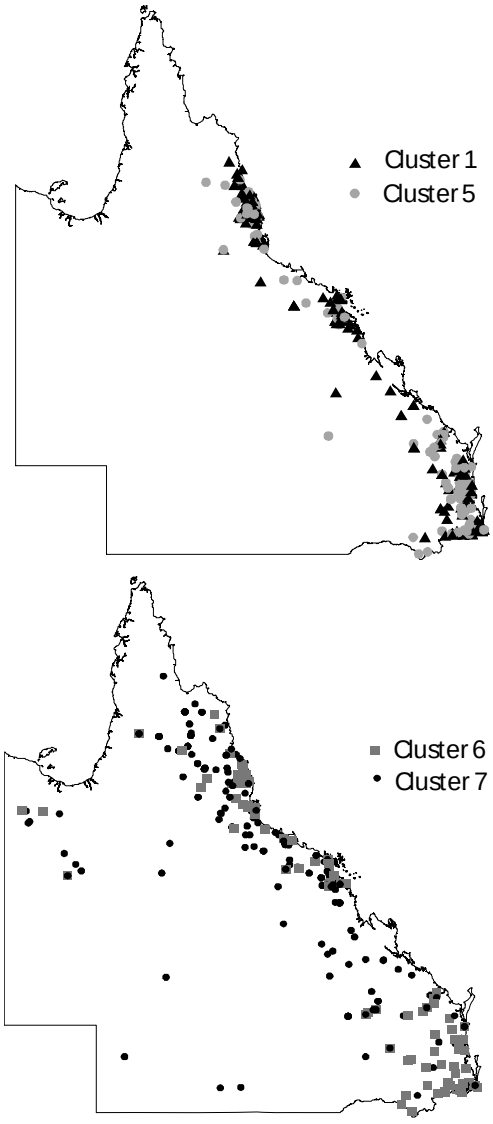


Table 4.10 shows results of univariate analysis between groups using only FFG. Predators and collectors distinguish the most between groups, although all FFG significantly distinguish between groups. Mahalonobis distance between groups (Table 4.11) using FFG only was the greatest between groups 1, 4, 5 and groups 6 and 7, and group 6 and 7 between themselves.

Table 4.10. Univariate analysis of variance between groups using FFG only.

Variable	F	P
Predator	397.4	0.000
Collector	178.0	0.000
Filterer	160.5	0.000
Grazer	143.4	0.000
Shredder	85.3	0.000

Figure 4.18. Spatial position of sites within clusters 1, 5, 6 and 7.

Table 4.11. Mahalanobis distance between groups using FFG only.

Cluster	2	3	4	5	6	7
1	2.11	2.23	2.65	1.99	3.58	4.31
2		1.80	2.87	3.58	3.34	2.59
3			2.28	2.73	2.23	2.54
4				3.38	4.37	3.47
5					3.15	5.24
6						3.80

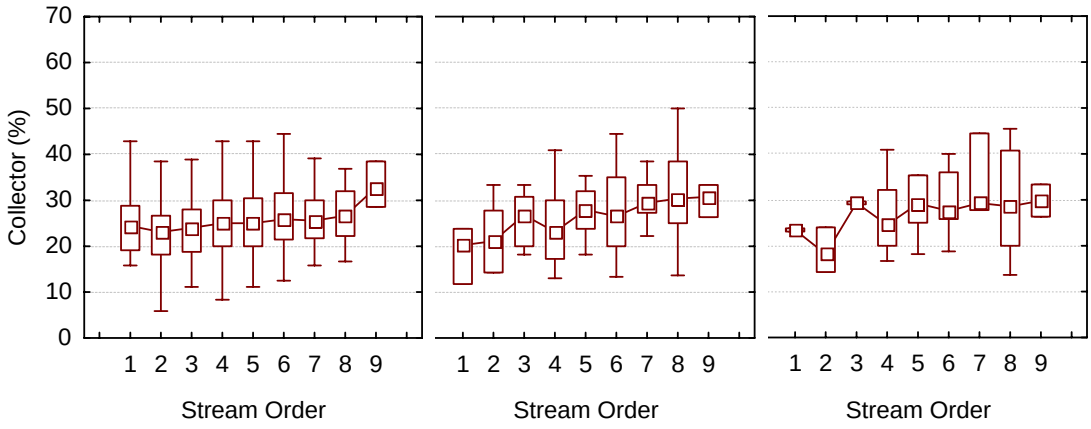
Similarly to the case with FFG, Mahalonobis distance between clusters based on water quality variables (Table 4.12) was the largest between groups 1, 5 and 7.

Table 4.12. Mahalanobis distance between groups using combination of natural settings and water quality variables.

Cluster	2	3	4	5	6	7
1	0.80	1.01	1.03	0.78	1.26	1.75
2		0.46	0.50	0.82	0.92	1.32
3			0.46	0.95	0.79	1.05
4				1.00	0.88	1.00
5					0.96	1.68
6						1.15

Figure 4.18 shows spatial distributions of sites belonging to the two most distinctive groups of clusters: 1, 5 and 6, 7. Clusters 1 and 5 are located mostly along the coastline in the areas characterised by high rainfall, comparatively low temperatures and high relief. Sites belonging to Clusters 6 and 7 are more widely distributed with many located further inland in the flatter areas, characterised by lower rainfall and higher average temperatures.

Figure 4.19 demonstrates the effect of turbidity on proportional distribution of collectors along the stream order gradient. All sites used for these plots had relatively ordinary nutrients concentrations (total N <0.71 and P<0.05 mg/L). The plots show clear increase in median percentage values of collectors, particularly in mid order streams.



a, b, c)

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. Box – 20-80%, whiskers – minimum and maximum.

Edge

Resulted SOM map 20x13 (final quantization error: 6.32, final topographic error: 0.08) was partitioned into 12 clusters. Box plots for median values of proportional values of FFG are shown at Figure 4.20. Trophic groups are in exactly the same order regarding their discriminating ability as for the riffle habitat, with predators and

collectors being the groups responsible for the most of the difference between clusters (Table 4.13).

Mahalonobis distances between clusters using FFG only and both natural settings and water quality variables are shown at Table 4.14 and Table 4.15 respectively. In general clusters 10, 11, 12 and to some extent cluster 8 are being the most different from the rest of the clusters both using FFG and all variables. Mean values of FFG and all variables for each of the seven resulted clusters are shown in Table 4.16. Cluster 12 is dominated by collectors, with low proportion of predators, grazers and shredders. Cluster 11 has moderate proportion of collectors but second high proportion of predators and the lowest proportions of grazers and shredders. Both clusters 11 and 12 are characterised by the high values for turbidity, nutrients and water temperature, largest distance from source and relatively low mean annual rainfall. Cluster 8 is dominated by predators with very low proportion of collectors, and low to medium proportions of other FFG. It is characterised by high conductivity and relatively high nutrients values. Cluster 1 is being very different from clusters 10, 11 and 12 and characterised by lowest proportion of collectors and highest proportion of both grazers and shredders. It is also characterised by high rainfall, low distance from source and low turbidity. Other clusters include sites with variable characteristics regarding both FFGs, natural settings and water quality characteristics. Similarly to the case with riffle habitat, proportion of filterer doesn't seem to follow any easily defined pattern.

Table 4.13. Univariate analysis of variance between clusters using FFG only, edge.

Variable	F	P
Predator	844.0	0.000
Collector	339.8	0.000
Filterer	261.5	0.000
Grazer	255.9	0.000
Shredder	92.3	0.000

Table 4.14. Mahalanobis distance between 12 SOM defined clusters using FFG, edge habitat.

	2	3	4	5	6	7	8	9	10	11	12
1	2.40	3.79	3.51	3.15	3.82	3.57	4.94	3.95	4.38	5.61	5.84
2		3.94	2.44	2.21	2.66	2.10	4.55	2.52	2.12	3.86	3.64
3			2.38	5.51	3.56	4.32	8.19	5.97	5.61	7.35	5.12
4				3.71	3.27	2.11	6.63	4.67	3.59	5.13	3.15
5					4.82	1.99	3.04	3.25	2.49	2.61	4.67
6						4.31	6.76	3.37	3.73	6.01	4.04
7							4.96	4.07	2.49	3.23	3.23
8								3.96	4.30	3.52	7.06
9									2.30	3.91	4.81
10										2.33	2.95
11											4.37

Table 4.15. Mahalanobis distance between 12 SOM defined clusters using physical settings and water quality variables, edge habitat.

	2	3	4	5	6	7	8	9	10	11	12
1	0.73	0.87	0.83	0.74	0.83	0.87	0.96	0.93	1.35	1.57	1.59
2		1.00	0.52	0.44	0.66	0.49	0.76	0.66	0.98	1.31	1.23
3			0.79	1.07	0.94	1.06	1.28	1.20	1.51	1.79	1.56
4				0.56	0.72	0.57	0.97	0.72	1.05	1.39	1.20
5					0.64	0.41	0.65	0.40	0.79	1.11	1.15
6						0.64	0.95	0.83	1.06	1.29	1.23
7							0.83	0.43	0.74	1.07	1.06
8								0.72	0.97	1.29	1.28
9									0.68	1.03	1.09
10										0.77	0.84
11											0.81

Table 4.16. Mean values of FFG and water quality variables for 12 SOM defined clusters, edge habitat.

Name	1	2	3	4	5	6	7	8	9	**10**	**11**	**12**
Number of samples in each cluster	150	237	239	322	369	131	264	127	163	153	168	118
Collector	14.54	23.51	24.96	27.91	19.56	28.06	25.69	13.33	22.97	28.68	26.99	38.26
Grazer	15.35	10.48	18.86	13.53	9.52	12.33	10.00	6.14	9.75	6.64	3.53	4.82
Shredder	18.07	13.40	15.60	11.79	10.93	13.63	10.41	12.38	8.63	7.89	6.23	8.47
Predator	46.03	46.89	33.18	39.89	53.77	39.90	47.47	64.19	53.29	51.48	58.51	41.93
Water Temperature	21.42	22.22	20.75	21.42	22.26	21.78	23.21	22.89	23.06	23.74	23.47	22.42
Conductivity	346.5	379.28	357.01	352.34	377.51	266.51	300.99	553.13	414.52	279.30	234.01	289.97
Dissolved oxygen	7.31	7.63	8.03	7.93	7.32	7.32	7.45	7.40	7.21	7.13	7.10	7.65
pH	7.26	7.43	7.41	7.49	7.45	7.44	7.55	7.35	7.51	7.44	7.52	7.55
Turbidity	11.95	12.71	15.84	19.69	28.41	30.14	35.93	38.02	44.90	70.97	131.51	137.27
Alkalinity	95.24	93.80	78.02	83.50	92.68	63.93	92.75	81.60	107.08	79.97	78.60	73.11
Total N	0.49	0.49	0.48	0.62	0.56	0.45	0.57	0.66	0.70	1.00	0.78	0.98
Total P	0.03	0.04	0.06	0.05	0.06	0.05	0.06	0.09	0.08	0.13	0.12	0.18
Depth	0.3	0.35	0.35	0.33	0.34	0.35	0.33	0.34	0.31	0.35	0.37	0.37
Maximal	0.13	0.08	0.19	0.14	0.10	0.11	0.10	0.06	0.09	0.06	0.04	0.08
Bedrock	4.16	3.58	3.76	2.74	3.72	2.78	2.36	2.83	2.48	3.61	4.55	1.75
Boulder	3.36	1.73	5.24	2.03	1.26	3.22	2.44	1.89	1.52	1.28	0.83	1.06
Cobble	8.56	5.71	12.85	6.91	3.74	6.54	5.87	2.37	4.61	2.36	2.28	3.39
Pebble	5.56	4.1	8.97	5.35	4.52	5.72	4.84	4.99	5.21	3.91	1.39	2.39
Gravel	8.23	10.1	11.78	8.50	7.65	9.65	8.50	8.46	7.45	7.50	4.23	8.37
Sand	37.7	40.52	29.91	37.65	40.50	39.96	39.71	38.07	36.07	35.49	39.30	34.76
Silt/Clay	32.33	34.09	27.45	36.78	38.58	32.17	36.25	41.37	42.63	45.82	47.39	48.24
Mean phi	-0.11	0.42	-1.09	0.63	0.96	0.13	0.68	1.28	1.34	1.66	1.89	2.08
Altitude	117.22	177.04	162.77	203.49	153.06	156.11	175.82	117.64	151.60	180.21	155.86	205.57
Slope	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Distance From Source	47.71	91.10	79.71	104.20	113.01	110.86	118.79	85.52	141.95	164.82	200.42	199.74
Mean annual rainfall	1617.87	1411.52	1518.42	1355.62	1280.08	1447.41	1277.86	1342.92	1176.869	1016.440	959.235	1081.803

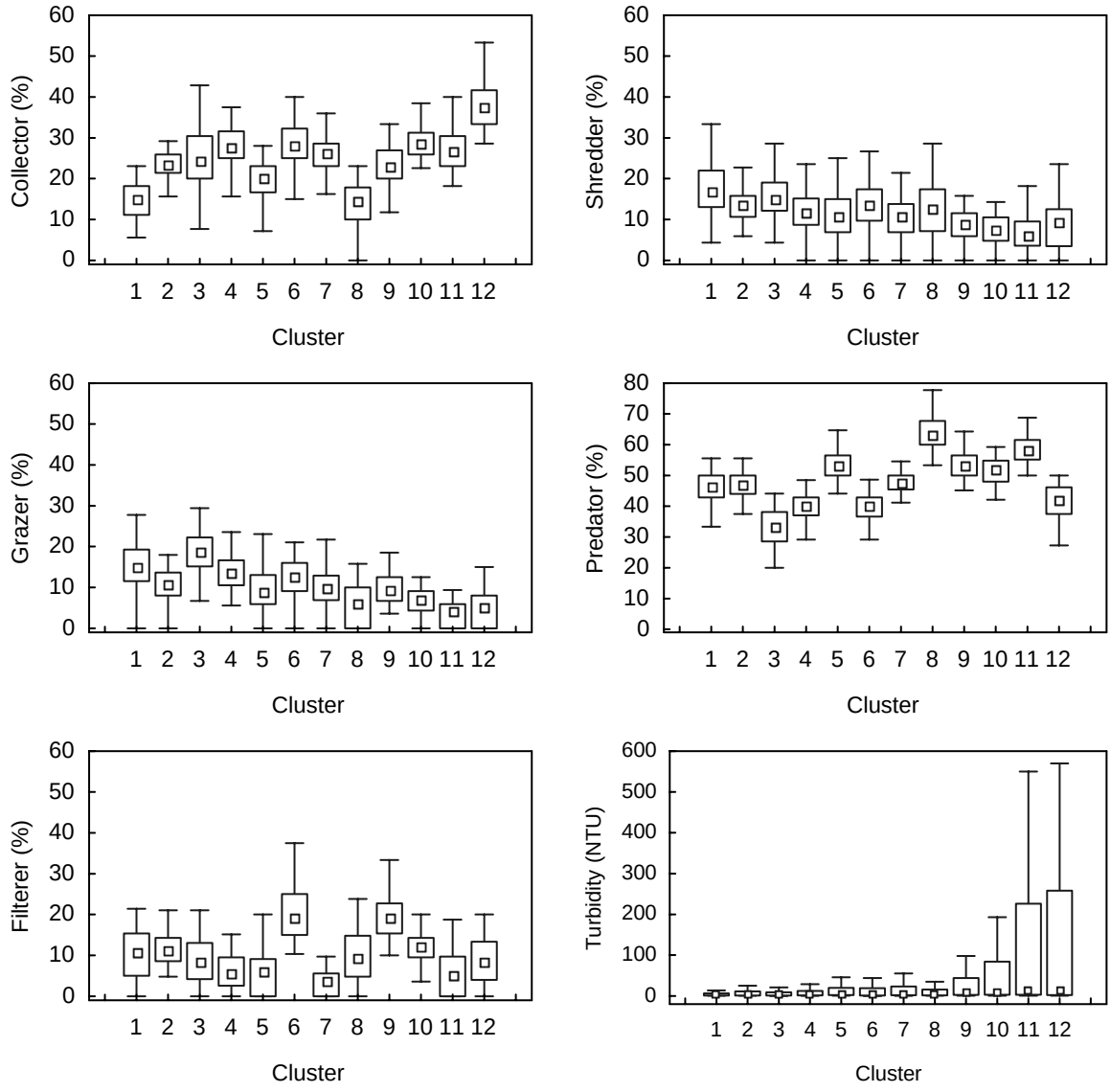


Figure 4.20. Box plots for median values of proportion of collectors and predators FFG in each of 12 SOM defined clusters, edge habitat. Box – 20-80%, whiskers – minimum and maximum.

Using a combination of SOM and CCA

Riffle

First two axes explained cumulative 4.7 percentage variance between trophic clusters and 86 % of cluster-environmental relation. Significance of all axis was evaluated by Monte Carlo test (999 unrestricted permutations). All axis were significant ($p<0.05$). Table 4.17 shows cumulative variance explained, F value and significance of water

quality and geoclimatic variables tested. Only significant variables ($p<0.05$) were used for the final model.

Figure 4.21 shows the resulting CCA bi-plot with SOM defined clusters in relation to the 9 significant variables. CCA provides a convenient visualisation of the relationships which were already mostly evident from the analysis of the mean values

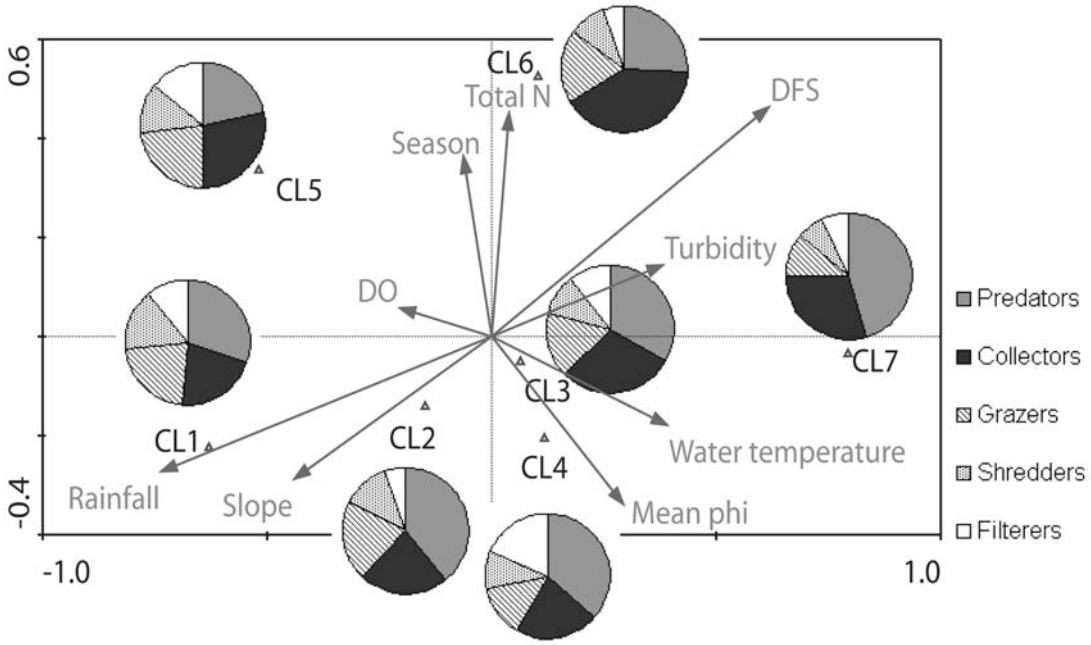


Figure 4.21. Bi-plot resulted from CCA with SOM defined clusters and eight environmental variables. Pie charts show mean percentages of each FFG within a cluster.

Table 4.17. Multivariate effects (in order of model selection) of CCA using SOM defined clusters as "species".

Variable	Cumulative variance explained	F	P
Mean annual rainfall	0.13	29.19	0.002
Distance From Source	0.18	11.62	0.002
Water Temperature	0.22	9.19	0.002
Season	0.26	9.66	0.002
Mean phi	0.28	4.26	0.002
Turbidity	0.29	3.31	0.006
Total N	0.31	3.49	0.004
Slope	0.32	2.72	0.01
DO	0.33	2.23	0.026
pH	0.34	1.7	0.11
Alkalinity	0.35	2.07	0.06
Conductivity	0.35	1.67	0.13
Maximal velocity	0.36	1.66	0.14
Depth	0.37	1.2	0.29
Total P	0.37	0.89	0.5
Altitude	0.37	0.67	0.66

within each of the SOM defined clusters. In general, assemblages located at the right hand side (clusters 1, 2 and 5) are characterised by a higher proportion of grazers and shredders and clusters 1 and 5 in particular are characterised by the lowest combined proportion of predators and collectors. These assemblages are associated with high rainfall and high dissolved oxygen, low water temperature, turbidity and conductivity.

Clusters 6 and 7, on the contrary, are characterised by the highest combined proportion of collectors and predators and located on the left side associated with generally decreasing water quality along the gradient of increasing distance from source. Clusters 3, 4 and 2 are located close to each other and closer to the middle section of the plot.

Edge

First two axes explained cumulative 2.1 percentage variance between trophic clusters and 72 % of cluster-environmental relation. All axis were significant ($p<0.05$) as evaluated by Monte Carlo test (999 unrestricted permutations). Table 4.18 shows cumulative variance explained, F value and significance of water quality and geo-climatic variables tested. Only significant variables ($p<0.05$) were used for the final model. Similarly to the riffle habitat, turbidity, distance from source and mean annual rainfall are the variable best discriminating between clusters. Maximal water velocity is more important in distinguishing between groups in case of edge habitat than riffle because edge habitat data set includes sites with no flow, when riffle habitat requires some flow by definition

Figure 4.22 shows CCA biplot for FFG clusters and significant variables, both water quality and geoclimatic. The overall picture is similar to that discovered in the case of riffle variable with the exception that conductivity and velocity are significant and included in the model. In generall, clusters located far from the center on the right hand side (10, 11, 12) are dominated by collectors or predators and have low proportion of grazers and shredders. Clusters 1 and 3 located directly opposite and characterised by relatively high proportions of grazers and shredders. The other clusters are located in between those two extremes and characterised by various combinations of FFG. The cluster 8 stands aside from the rest and associated with water temperature and conductivity gradient. It is strongly dominated by predators.

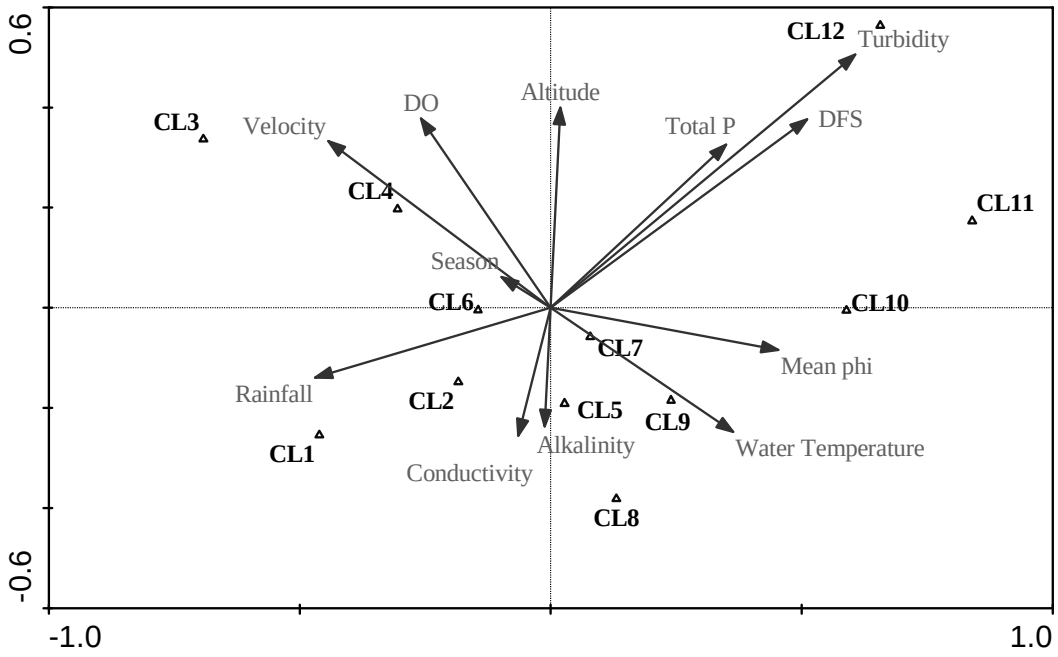


Figure 4.22. Bi-plot resulted from CCA with SOM defined clusters and environmental variables.

Table 4.18. Multivariate effects (in order of model selection) of CCA using SOM defined clusters as "species", edge habitat.

Variable	Cumulative variance explained	F	P
Turbidity	0.08	18.82	0.002
Max Velocity	0.12	8.36	0.002
Water Temperature	0.15	7.36	0.002
Season	0.18	6.63	0.002
Mean phi	0.21	4.83	0.002
Altitude	0.23	4.7	0.002
Distance from source	0.24	4.08	0.002
Dissolved oxigen	0.26	3.96	0.002
Conductivity	0.27	2.2	0.02
Total P	0.28	2.19	0.02
Rainfall	0.29	2.14	0.018
Alkalinity	0.3	1.94	0.03
Depth	0.31	1.68	0.08
pH	0.32	1.39	0.15
Slope	0.32	1.11	0.3

Discussion and conclusions

This study explored broad patterns of proportional changes in trophic structure of macroinvertebrate communities in relation to water quality and geo-climatic variables. Self Organising Maps built both for the purpose of comparison between gradient changes in proportion of different FFGs and clustering FFGs into similar groups produced interesting and mostly logically explainable results (hypothesis 3). Results were broadly similar for both riffle and edge habitats, however there were some difference. In the case of riffle, distance from source and mean annual rainfall appear to be the most important geo-climatic variables for the discrimination between clusters of FFG. The importance of distance from source is naturally explained by the river continuum concept. It appears that in a case of riffle habitat in Queensland succession of FFG along stream order gradient to some extent follows the assumptions of RCC when stream conditions are close to natural (hypothesis 1 is not true in this case). However, this trend was practically undetectable when a variety of streams experiencing some kind of antropogenic impact has been analysed.

In the case of edge habitat we could not observe any detectable changes in FFG along the river continuum neither in reference nor in test sites. One of the possible explanations it that edge habitat dataset includes many more sites from inland areas with low rainfall and intermittent flow, when riffle habitat by definition requires flowing water. The effect of flow might be masking natural gradients along the river continuum as ephemeral streams would be dominated by highly resilient fauna adapted to the surviving extended no flow periods despite of the site's stream order. This in fact is confirmed by the fact that water velocity has more significance variable discriminating between FFG clusters in the case of edge habitats than in riffle habitat. The difference in faunal composition between intermittent flow streams and stream flowing most of the time is likely to be stronger in comparison with difference between communities sampled in streams flowing most of the time (in case of riffle habitat).

Season when the site was sampled appears to be important in both riffle and edge habitat cases. As sites were sampled only in spring and autumn we assume that the difference is the previous season, which in the case of Queensland is wet season (summer) or dry season (winter). The effect of season can be explained by the influence of the previous season, in other words whether system is coming from wet or dry season. However, it is difficult to make any definite conclusions about the effect of seasonality on the trophic structure, this effect might be quite different depending on the geographic position of the site (as seasonal effect might be very different in wet tropic and in dry inland areas), plus high irregularity of seasonal rainfall in many areas can make it difficult to discriminate the effects with high confidence. This might be an interesting area for the future research.

As turbidity, water temperature and nutrients have been found the most important water quality variables affecting trophic structure of macroinvertebrate communities for both riffle and edge habitats (hypothesis 2 is true), is it very logical to explain the difference in FFG succession between reference and test sites in the case of riffle habitat by effect of these variables in particular. Overseas studies shown that nutrient and sediment load as a result of land use practices overrides natural feature. In Lapwai Creek, an agriculturally impaired stream in Northern Idaho, functional groups of

macroinvertebrates were similar among sites despite expectations of differences along a river continuum, and the assemblage composition was markedly different from that found in less-impaired stream (Delong and Brusven, 1998). Despite substantial variation in terrain and the extent of riparian vegetation, the relative homogeneity of the macroinvertebrate assemblages of these sites was interpreted, via increasing sedimentation and the dominance of periphyton as an energy source, as evidence of the overwhelming effect of agricultural land use.

Increase in water temperature is one of the result of degradation or clearance of riparian vegetation. It should mean increased light penetration and more favourable conditions in growth of algae and result in an increase in the proportion of grazers. In fact, whether this assumption is true is highly dependent on particular local conditions as substrate composition, turbidity, availability of nutrients, etc. Clearance of the riparian vegetation is also a factor contributing to the sedimentation and nutrients load. In our case we could not detect increase in the proportion of grazers associated with increase in water temperature. This can be explained by the fact that the FFG clusters associated with high water temperature also were characterised by high turbidity, which makes it difficult to make any definitive conclusions about effect of water temperature on trophic structure.

Clearance of the riparian vegetation can also result in the reduction of leaf litter and subsequent reduction in proportion of shredders. This was confirmed by our results. The clusters associated with high water temperature were also characterised by relatively low proportion of shredders.

The effect of turbidity on trophic structure is the easiest to explain. Turbidity prevents light penetration into the water column and slows growth of water plants and algae decreasing primary production and food quality (Allan, 2004). This shifts primary autotrophic pathways of energy transfer to heterotrophic ones with the subsequent change in the trophic structure of macroinvertebrate communities. Excess deposition of sediments also slows down break down of leaf litter by smothering leaves and reducing availability of oxygen to the leaf surface. This is likely to reduce the number of oxygen loving microbes and shredders macroinvertebrates, impairs substrate suitability for periphyton and biofilm production and reduces stream depth heterogeneity leading to decrease in pool species. Our results shown that turbidity causes reduction in the proportion of grazers and shredders and increase in the proportion of collectors and to some extent predators. This effect is most likely visible in mid-order streams were turbidity is naturally low. It confirmed by our results, when we compare succession of the FFG along the river continuum for reference and test sites, the natural pattern of succession becomes indistinct as turbidity and other water quality factors as water temperature and nutrients become artificially elevated.

Effect of nutrients is increase in autotrophic biomass and production, accelerated litter breakdown rates and may cause decrease in dissolved oxygen and shift from sensitive species to more tolerant, often non-native species (Allan, 2004). Our results have shown that increase in nutrients (total nitrogen and total phosphorus) is associated with increase in proportion of predators and collectors at the expense of the other groups. We could not detect any pronounced increase in grazers which might be explained that grazers in the case of Queensland is generally more sensitive to a variety of stress than collectors and predators, or that the many sites with elevated

nutrients were also characterised by high turbidity as well (which is true in the case of edge habitat).

It is not easy to make any assumptions about effect of conductivity on trophic structure of macroinvertebrate assemblages as clusters with high conductivity were also characterised with elevated nitrogen and phosphorus and comparatively high water temperatures. Cluster 8 (edge habitat) had the highest mean conductivity and was characterised by the highest proportion of predators and relatively low proportion of collectors. This might reflect complex pattern of replacement of salinity sensitive taxa by salinity tolerant. There is a possibility that the high conductivity might have different effect depending whether it associated with natural conditions (specific types of soil) or secondary salinisation also associated with variety of other factors as input of nutrients and sediments through degraded riparian vegetation.

Our results have shown that Self Organising Maps and combination of SOM and CCA are useful methods for the analysis of patterns in trophic structure of macroinvertebrate communities (hypothesis 4 is true). It provided simultaneous reduction of data dimensionality and the visualisation of relationships between different types of assemblages and environmental variables. The aim of this study was to demonstrate use of the methods in order to reveal patterns in trophic structure of macroinvertebrate communities and their relation to both geo-climatic and water quality factors. Our results reveal a number of interesting relationships, which certainly can be further, investigated and improve our understanding of Australian freshwater ecosystems. In general, it was shown that trophic structure is affected by a number of both natural geo-climatic characteristics and water quality parameters. Increase in water temperature, turbidity and elevated nutrients can affect natural succession of FFG along the stream order gradients. In particular, elevated proportion of collectors and predators at the expense of the other trophic groups can be expected at the sites experiencing some kind of anthropogenic impact.

Chapter 5

Predicting macroinvertebrate taxa and macroinvertebrate communities in freshwater streams by MLP

5.1 Using the clean-water (or referential) approach and Victorian dataset

Introduction

The central idea of the referential approach is to study the relationship between habitat conditions and biota in near pristine sites and then apply this relationship to predict the fauna at impacted sites as if they were unimpacted (Reynoldson et al., 1997). RIVPAC and AusRivAs, the two most widely used assessment systems in Australia and UK are based on referential approach using a combination of statistical methods. Huong (2001) conducted an extensive study comparing the performance of the predictive neural network models with AusRivAs using the same dataset from NR&M (Queensland). The results of this study showed the ANN were able to predict the occurrence of stream macroinvertebrates with high accuracy and their performance was superior to that of AusRivAs (Huoung et al., 2001). In order to further investigate applicability of ANNs to the prediction of the occurrence of stream macroinvertebrate in Australia in accordance with referential approach I used the dataset provided by Victorian EPA. The main distinction of this data from the one from NR&M is limited number of predictor variables and the fact that macroinvertebrate occurrence data have been pooled together from the autumn and spring sampling events.

The main hypothesis for this study is:

ANN models can be used to predict occurrence of macroinvertebrates according to the referential approach in Victoria streams.

Methods

I used 21 variables for physical and biological habitat properties as inputs and binary data for the occurrence of 15 macroinvertebrate taxa as an output. The 15 output taxa

were randomly chosen in order to validate the models' accuracy for common and rare taxa, where 5 taxa were considered to be very common (present at more than 70% of sites), 5 taxa to be common (present at about 50% of sites) and 5 taxa uncommon (present at less than 30% of all sites). The accuracy of the ANNs predictions was estimated as the percentage of correct predictions in testing dataset, using randomly chosen 30% of data not used for training.

It was shown by Manel et al. (2001) that percentage of correct predictions as widespread measure of predictive accuracy is affected systematically by the prevalence (i.e. the frequency of occurrence) of the target organism, and reliance of this measure using raw can be misleading. In order to avoid the problem with overrepresented 0 or 1, I equalized the data by duplicating the data points so the dataset contains 50% of '0' and 50% of '1' values (see Chapter 3). The models have been developed using the Neuro Solutions 4 software. The cross-validation technique with 10% of data has been used to determine the optimum architecture of the ANN and prevent overtraining.

Results

The average accuracy of all models estimated using a testing subset was 77.7%. Average accuracy for very common taxa was 75.6%, for common 75.9% and for uncommon 81.6%. In general the majority of the models were accurate, approaching or higher than 70% of correct predictions (Table 5.1).

Table 5.1. Percent of correct predictions of occurrence of macroinvertebrates in streams of Victoria (testing 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

Discussion and conclusion

Huong et al. (2001) compared performance of multi-layered perceptron models for 37 macroinvertebrate taxa based on 896 stream data sets from the Queensland stream system with that of AusRivAs. The ANN model validation by means of 167 independent data sets revealed 73% as lowest rate and 82% as average rate of correct ANN predictions of stream site habitats. The average rate of correct predictions was slightly higher than the one resulted from our study using Victorian dataset (77.7%), however this difference is marginal, and might be accounted to the fact that we used equalized data in terms of number of 1 and 0, when Huong et al. (2001) used the raw data, which could influence the measurements of the accuracy.

I selected three types of taxa with different frequency of occurrence (very common, common and uncommon) in order to test whether ANN is able to deal with prevalence of presence or absence data points in the dataset. The models developed with very common and uncommon taxa were as accurate as those developed for the common taxa. In general, ANNs developed for this study predicted the occurrence of stream macroinvertebrates in Victoria almost as well as those previously developed for the Queensland streams using NR&M data.

5.2 Using the dirty-water approach and NSW dataset

Introduction

Biotic communities in streams are influenced by a large number of environmental factors such as the geological history of the area, environmental stability, ecosystem productivity, habitat heterogeneity, competition and predation (Compin and Cereghino, 2003). The taxonomic richness is an integrative descriptor of the biotic community is also strongly influenced by anthropogenic disturbances, which may lead to losses of taxa (Brittain and Saltveit, 1989). Therefore, taxonomic richness is often used as a biological indicator of disturbance. Park et al. (2003) used counterpropagation neural network to predict species richness and Shannon diversity index of benthic macroinvertebrate communities using 34 environmental variables. The model showed a high accuracy of the prediction ($r > 0.9$ and 0.67 for learning and testing process, respectively).

However, the sensitivity of taxonomic diversity in a given geographical region must be assessed with respect to its biotic and abiotic specificity. Dirty-water models are particularly interesting in this respect as they utilise a wide range of input variables, including those potentially altered by anthropogenic impacts. The most important application of the dirty-water approach is the simulation of various scenarios in order to predict the ecological consequences of altering input variables.

In this study I attempted the prediction of two biotic indices: Number of Macroinvertebrate Families and Number of Native Macrophyte Species with dirty-water models using relatively small dataset from NSW. Traditionally, models like RIVPACs and AusRivAs use taxa specific predictions. In Australia, there is a large number of even relatively common taxa, which often makes the process of taxa-specific modelling slow and very complex. In some cases it might be advantages to be

able to quickly run multiple scenarios of anthropogenic impact or remediation using small number of biological indices as taxonomic richness, Shannon index, PET richness, etc. as an output, given sufficient accuracy of the predictive models.

The main hypothesis for this study is:

SOM component planes is a convenient tool for the detection of the relationships between abiotic and biotic variables.

Methods

I used a combined set of 20 physical, chemical and biological predictor variables (Table 5.2). The Number of Macroinvertebrate Families and Number of Native Macrophyte Species were used as outputs variables. A number of models were built and the best performing models were selected. The architecture of both models used for the prediction of Number of Macroinvertebrate Families and Number of Native Macrophyte Species was the same: 20 neurons in input layes, 6 neurons in hidden layer and 1 neuron in output layer, with 'tahn' transfer function. 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 using randomly selected 30% of the dataset as validation subset, which was not used for training.

Table 5.2. The list of predictor variables used for the development of dirty-water models.

Variable	MIN	MAX	MEAN
Site elevation	5.00	780.00	268.16
Site slope	0.10	88.89	8.52
Site discharge	0.00	6.13	0.28
Average of maximum and minimum stream width per quadrat	0.22	44.38	7.27
Average of maximum stream depth per quadrat	0.01	3.19	0.72
Water temperature at 0.2 m	6.40	38.00	20.14
Turbidity at 0.2 m	0.40	64.70	12.30
Electrical conductivity at 0.2 m	33.00	2330.00	370.45
pH at 0.2 m	4.42	8.70	7.42
Ammoniacal nitrogen at 0.2 m	0.01	1.60	0.06
Oxidised (nitrate plus nitrite) nitrogen at 0.2 m	0.01	1.00	0.05
Filterable phosphorus at 0.2 m	0.00	0.85	0.03
Bank erosion score (range 0-100)	0.00	96.43	8.37
Catchment area above site	1.00	1815.75	231.75

Results

The 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 5.1).

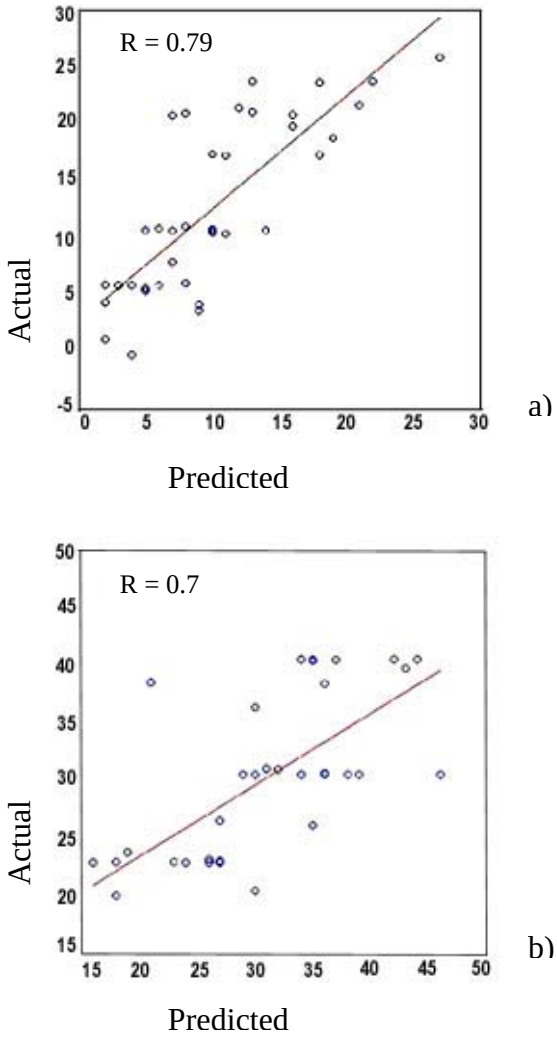


Figure 5.1. Predicted output versus actual output for the validation set (30% of the database not used for training) for: a) number of native macrophytes species, b) number of families of stream macroinvertebrates from NSW.

Discussion and conclusion

In this chapter I tested the applicability of ANNs for modelling 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 modelled were overall taxonomic richness rather than single taxa occurrence.

Even though there were only 122 samples of the NSW streams system available, results of the predictive modelling of two biological variables for the 'dirty-water' approach demonstrated that supervised ANNs can 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 NR&M) collected over many years over vast areas are most likely to contain the information on various conditions including extreme events. 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. Scenario analysis including the combined effect of salinisation and nutrients on the structure of macroinvertebrate communities is described in Chapter 7.

5.3 Optimisation of the modelling design in respect to the cost efficiency of environmental monitoring.

This chapter contains several small studies designed to answer specific practical questions formulated by the staff of NR&M. These questions mainly concern the practical efficiency and cost in relation to the design of field survey and the utilisation of available data.

5.3.1 How many predictor variables is enough?

The relationships between variables in ecology are almost always very complicated and highly non-linear (Gevrey et al., 2003). The accuracy of predictive models is highly dependent on the availability of variables explaining the observed patterns in biota distribution. The choice of the predictor variables is of particular importance, as some variables might be more important than others. Inclusion of many unimportant predictors can make the models slow, cumbersome and even less accurate. Huong et al. (2001) showed that exclusion of redundant inputs could improve the performance of the ANN models. However, finding the optimum number of predictor variables for each taxa-specific model can be a very lengthy and complex process. Using the same number of predictor variables for all individual models appears to be a more practical solution, however it is unclear how much accuracy can be lost and what is the minimum optimum number of variables needed. In order to answer this question I designed a following study.

Methods

In order to investigate the relationship between the number of predictor variables and accuracy of the model I prepared 4 sets of variables (Table 5.3) starting from full set of 50 variables used in previous studies (Huong et al., 2001) and reducing a number of variable on logical basic in relation to the cost of field sampling and measurements.

Since we are trying to come up with a definite number of variables optimum for all the taxa specific models, we can not use sensitivity analysis for each models as conducted by Huong et al. (2001), and had to use previous knowledge or considerations of cost when deciding on which variable to exclude. For example, for the minimal reduction set I excluded some of the variables describing rainfall pattern, some water quality variables as total hardness, which is correlated with conductivity. For the maximal reduction subset I excluded variables describing substrate composition and the concentrations of ions and cations. For the extreme reduction set I excluded all rainfall variables as the influence of rainfall can be still partially explained by the altitude and conductivity. I also excluded nutrients as they are to some extent correlated with water temperature. The infinite combination of variables is possible, but this study is only intended to provide some insight into the loss of accuracy from excluding a variety of parameters, and the combinations of variables were chosen arbitrarily.

I selected 5 taxa generally considered to be sensitive to a variety of anthropogenic stresses (taxa with high SIGNAL score, see Chessman, (2003)), as accuracy of the model is more important for the prediction of sensitive taxa than opportunistic in order to be able to detect the anthropogenic impact. I build 5 taxa specific models for the occurrence of Leptophlebiidae, Gomphidae, Calamoceratidae, Philopotamidae and Helicopsychidae, using 4 sets of predictor variables (totally 20 models). Statewide data collected from all 5 habitats were used, 30% of data were used for the validation purposes and 10% of data were used to control overtraining or as cross-validation subset. The number of neurons in the input layer was respective to the number of predictor variables, ten neurons was used in a hidden layer in all models.

In order to avoid confounding by the effect of prevalence of presence or absence (see Chapter 3) the validation subset was equalised by duplicating values until the number of ‘1’ and ‘0’ was equal. Percentage of correct predictions resulted from the simulation on validation subset was used as the measurement of accuracy.

Table 5.3. Subsets of predictor variables used to investigate the relationship between the number of predictor variables and accuracy of the model.

Full set (50 variables)	Minimal Reduction (38 varibles)	Maximal Reduction (19 va ıbles)	Extreme Reduction (9 variables)
Habitat	Habitat	Habitat	Habitat
Season	Season	Season	Season
Width (m)	Width (m)	Depth (m)	Latitude
Depth (m)	Depth (m)	Velocity - max (m/s)	Longitude
Velocity - max (m/s)	Velocity - max (m/s)	Mean phi	Altitude (m)
Bedrock (%)	Bedrock (%)	Latitude	Water Temp (°C)
Boulder (%)	Boulder (%)	Longitude	Conductivity (µS/cm)
Cobble (%)	Cobble (%)	Altitude (m)	DO (mg/L)
Pebble (%)	Pebble (%)	Stream Order	pH
Gravel (%)	Gravel (%)	Slope (m/m)	
Sand (%)	Sand (%)	Distance From Source (km)	
Silt/Clay (%)	Silt/Clay (%)	Ratio of a / b	
Detrital cover (%)	Detrital cover (%)	Water Temp (°C)	
Mean phi	Mean phi	Conductivity (µS/cm)	
Latitude	Latitude	DO (mg/L)	
Longitude	Longitude	pH	
Altitude (m)	Altitude (m)	Turbidity (NTU)	
Stream Order	Stream Order	Total N (mg/L as N)	
Slope (km/m)	Slope (km/m)	Total P (mg/L as P)	
Distance From Source (km)	Distance From Source (km)		
0-2 Reach	0-2 Reach		
Ratio of a/b	Ratio of a / b		
Range in wet season monthly means	Mean annual rainfall (mm)		
Range in dry season monthly means	Soil Type Number		
Percentage rainfall in wet season	Vegetation Type Number		
Mean annual rainfall (mm)	Water Temp (°C)		
Mean daily max temp (^{o}C)	Conductivity (µS/cm)		
Mean daily min temp (^{o}C)	DO (mg/L)		
Soil Type Number	pH		
Vegetation Type Number	Turbidity (NTU)		
Water Temp (°C)	Alkalinity (mg/L CaCO3)		
Conductivity (µS/cm)	Total N (mg/L as N)		
DO (mg/L)	Total P (mg/L as P)		
pH	K^{+} (mg/L)		
Turbidity (NTU)	$CO3^{--}$ (mg/L)		
Alkalinity (mg/L CaCO3)	$SO4^{--}$ (mg/L)		
Total Hardness (mg/L CaCO3)	0-4. Habitats		
Total N (mg/L as N)	0-8. substrate categories		
Total P (mg/L as P)			
K^{+} (mg/L)			
Ca^{++} (mg/L)			
Mg^{++} (mg/L)			
$HCO3^{-}$ (mg/L)			
$CO3^{--}$ (mg/L)			
$SO4^{--}$ (mg/L)			
0-4. Habitats			
0-8. substrate categories			
Mean Channel Width (m)			
Mean Depth (m)			
Instantaneous Discharge (cumec)			

Results

On average, the models built using full set of predictor variables (50) had the highest predictive accuracy, however, the decrease in accuracy for the reduced subsets was not very big and even models with extremely reduced set of input variables were still quite accurate approaching or exceeding threshold of 70% of correct predictions. The highest decrease in predictive power with the reduction of number of input variables was observed for Leptophlebiidae, from 75.31 % for full set to 69.24 % for the extreme reduction subset. This can be explained by the fact that many environmental variables are correlated to some degree and the exclusion of many variables has little effect as far as the essential factors are kept. Exclusion of the variables describing substrate composition as percentage of boulder, gravel, sand, etc. appear to have very little effect in particular, as the difference between the predictive accuracy of minimal and maximal reduction subsets is very small and the performance of the model for Gomphidae has actually improved.

Table 5.4. Comparative accuracy (% of correct predictions) of taxa specific models trained using different sets of predictor variables.

Taxa	Full set (50 var)	Minimal Reduction (38 var)	Maximal Reduction (19 var)	Extreme reduction (9 var)
Leptophlebiidae	75.31	72.86	72.71	69.24
Gomphidae	70.28	67.89	69.45	67.39
Calamoceratidae	72.18	70.22	69.07	67.59
Philopotamidae	86.15	85.34	85.31	82.94
Helicopsychidae	84.57	81.61	75.48	72.62
Average	77.70	75.58	74.40	71.96

Discussion and conclusion

Our results showed that it is possible to build accurate ANN models using very limited sets of the predictor variables. Use of many variables does improve the predictive accuracy but the rate of improvement is not very high. Use of medium sized sets of predictor variables (19-38) might be the most practical solution.

5.3.2 Generic models versus local models

Consideration of the modelling accuracy and practicality are always an important aspect in environmental decision-making. When trying to come up with standardised framework, the possible trade-offs between accuracy, complexity and practicality have to be considered. It is much easier and faster to train one state-wide generic model and use it for the subsequent predictions on a local or state-wide scale when necessary, however, Australian and the state of Queensland's aquatic ecosystems in particular are characterised by extremely diverse conditions and natural variability can

play a significant role. I am not aware of any studies attempting to understand the loss of accuracy when comparing models built using local or wide scale data. It is also unclear to what degree ANNs can cope with the natural variability on the state-wide scale. To address the question whether local models are more accurate then generic ones and whether the models trained on the data from one bioregion (within state of Queensland) can be applied to the neighbouring bioregions I designed the following study.

Methods

In order to investigate the degree to which natural variability can affect the predictive accuracy of ANN models I used data provided by NR&M from three geographic scales:

1) Statewide data: only riffle habitat was used in order to reduce the local variability between different habitats.

2) Bioregion scale subsets. I used three bioregions defined by the Aquatic Ecosystem Health Unit, NR&M, namely: Central, South-Eastern Queensland (SEQ) and Wet tropics and Cape (WT&Cape). See Figure 4.6 for the map of QLD bioregions.

3) Catchment level subsets. I used 2 catchments: Brisbane and Mitchell, these were the catchments with the largest number of observations.

I built 5 taxa specific models using state-wide data, data from three bioregions and data from two catchments. I chose taxa generally sensitive to the variety of anthropogenic impacts as it is more important to get accurate models for these taxa than for the opportunistic taxa.

I used 50 predictor variables (see Table 5.3, full set for the list of variables) for all models with 10 neurons in hidden layer and 1 neuron in output layer for each taxon. All data subsets were split on training (70%) and validation subsets (30%). Cross-validation (10% of all data) was only used in the initial stages in order to determine the optimum architecture and the number of training epochs when overtraining does not occur. All the results reported are based on simulations using validation set, 30% of all data not used for training. Validation data has been equalized in order to avoid the effect of prevalence of ‘1’ or ‘0’.

To address the question whether bioregion specific models can be used for the neighbouring geographic regions, the models developed for the previous three bioregions were tested on neighbouring bioregions (Central and SEQ, WT&Cape and Western).

Results

Table 5.5 shows the accuracy of the models developed using data on the different geographical scales in comparison with the accuracy of the generic model. Accuracy

of the generic model varied from taxon to taxon, ranging from 65.21% for Calamoceratidae to 89.46% for Philopotamidae, being 77.16% on average.

Models developed and tested on the same bioregion were slightly more accurate on average, but some taxon-specific models were less accurate than the generic ones, for example, Leptophlebiidae models for the Cental bioregion and SEQ were significantly less accurate that the generic model. The same stands for Gomphidae in WT&Cape. This can be explained by uneven geographical distribution of these particular taxa. It is possible that the taxon was not present often enough in these bioregions for the model to derive a predictable pattern. However, bioregional models for Calamoceratidae, Philopotamidae and Helicopsychidae were more accurate than generic models when trained and tested on the data from the same bioregion.

When tested on the geographical area different to that on which they were trained, bioregional models were generally less accurate than the generic models, with the average percent of correct predictions ranging from 57.03 % to 67.66%. However, Leptophlebiidae and Philopotamidae models developed for the Central bioregion were still very accurate when tested on the data from SEQ (97.61% and 84.56% respectively), however, the models trained on data from SEQ and tested on Cental were not accurate at all with the exception of Philopotamidae (73.83% of correct predictions).

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.

Trained on:	Tested on:	Leptophlebiidae	Gomphidae	Calamoceratidae	Philopotamidae	Helicopsychidae	Average
State-wide	**State-wide**	**79.75**	**66.89**	**65.21**	**89.46**	**84.5**	**77.16**
Central	Central	55.26	68.29	91.89	95.57	95.59	81.32
SEQ	SEQ	67.83	70.87	89.19	82.49	93.02	80.68
WT&Cape	WT&Cape	80.42	56.69	88.89	93.29	91.32	82.12
Mitchell catchment	Mitchell catchment	62.5	68.83	94.34	92.52	94.34	82.74
Brisbane catchment	Brisbane catchment	86.96	70.37	77.5	93.88	95	84.74
Central	SEQ	97.61	55.95	52.85	84.56	47.32	67.66
SEQ	Central	54.98	54.41	56.28	73.89	57.91	59.49
WT&Cape	Western	70.91	29.09	66.67	47.37	71.13	57.03

Catchment specific models were the most accurate on average, more accurate than generic and slightly more accurate than bioregional models. Only one taxon specific model (Leptophlebiidae) was less accurate than the generic model (62.5% and 79.75% respectively).

Discussion and conclusion

On average, bioregional models perform better in comparison with generic models when tested on the data from the same bioregion, however improvement in the performance is not consistent from taxon to taxon. Depending on the amount of data and specific distributional pattern, bioregional models can be actually less accurate than the generic ones. Not surprisingly, bioregional models perform poorly when applied to the data from the different bioregion, however, exceptions are possible depending on the distribution of specific taxa. In general, catchment models are even more accurate, however exceptions are possible again.

Even though local models are more accurate than generic, the rate of improvement is not very high (about 3 to 7%). Depending on the task at hand, and the available dataset, both approaches (generic and localised) can be used successfully. Generic models can be used in the cases where not much data available from the region of interest, when in cases where, for example, very accurate model is needed for the scenario analysis on the catchments scale and data are available, it would be better to train and utilise bioregional or catchment model.

5.3.3 Matter of time

Temporal variability is an inherent feature of freshwater systems. In many parts of Australia rainfall and daily temperature pattern vary from year to year. The area covered by state of Queensland has two relatively pronounced seasons: wet summer and dry winter. Sampling of macroinvertebrates is conducted in autumn and spring. It is generally accepted that there is no significant seasonal difference in macroinvertebrate communities in relation to autumn and spring (Jon Marshall, NR&M, personal communication). To investigate the possible effect of temporal and seasonal variability on the predictive accuracy of ANN models I designed the following study.

Methods

To address question whether models developed for the data collected in one season can still be applicable for the data collected in the other season, I developed 5 taxa specific models for the same taxa used in the previous chapter using three subsets of data:

1) All data from riffle habitat, all years, both seasons mixed.
2) Only data collected in the season 1(autumn).
3) Only data collected in the season 2 (spring).

Full set (50 predictor variables) was used. All models had 50-10-1 architecture with ‘tahn’transfer function. The models were trained using three abovementioned subsets. The models trained on the data with mixed seasons was tested on randomly selected 30% of the data. Models trained on data collected in season 1 were tested on season 2

and vice versa. All validation data were equalized to 50-50% ratio to account for the effect of prevalence of '1' or '0' values.

To address question whether models developed for the data collected in one year can still be applicable for the data collected in the subsequent years, I developed 5 taxa specific models (full set, riffle habitat) using only data collected in 1994 and then tested the model on: 1) 30% of data collected in 1994 but not used for training, 2) all subsequent years from 1995 to 2001. All validation data were equalized to 50/50 ratio.

Results

All models tested on the season different from the one they were trained on had inferior performance comparatively to the models trained on the data with mixed seasons (Table 5.6). However, more than half of the models trained and tested on different seasons were still quite accurate achieving the 70% threshold of the correct predictions.

Table 5.6. Comparative accuracy of the season specific and mixed-seasons models (% of correct predictions).

Taxa	Mixed seasons	Trained on season 1 Tested on season 2	Trained on season 2 Tested on season 1
Leptophlebiidae	75.31	71.13	73.10
Gomphidae	70.28	67.35	64.02
Calamoceratidae	72.19	59.57	65.52
Philopotamidae	86.16	85.02	79.65
Helicopsychidae	84.58	75.03	75.28
Average	**77.70**	**71.62**	**71.52**

Figure 5.2 demonstrates the accuracy of the models built using only data collected in 1994 when tested on validation subset from 1994 and all data collected in each subsequent year from 1995 to 2001. It is obvious that predictive accuracy of the 1994 models is not depended on the year used for the testing. In case of Gomphidae accuracy of the 1994 model was even higher when tested on 2000. It appears that on average the models were slightly less accurate when tested on the data from 1996 and 1997 in comparison with all the other years.

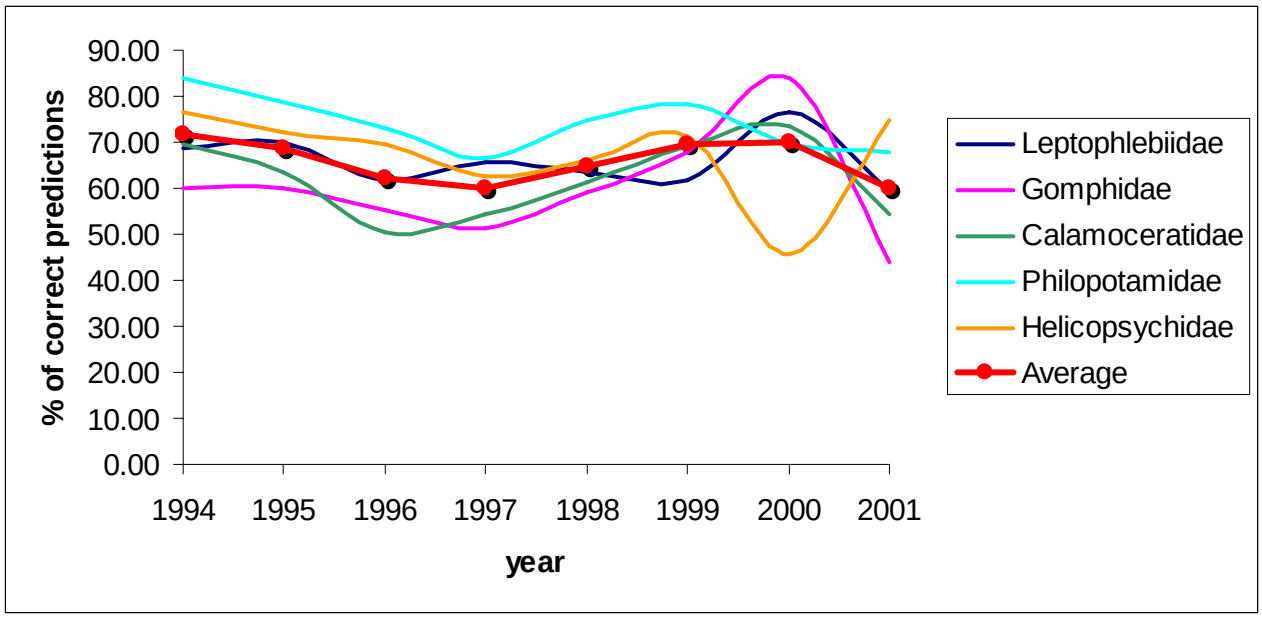


Figure 5.2. Accuracy of the year 1994 model when tested on data collected during other years.

Discussion and conclusions

In this chapter I tried to answer the question on how temporal variability and namely seasonal and annual changes affect the accuracy of the predictive ANN models. My results have shown that seasonal variations appear to be more important than annual, however, ANN models are still capable of achieving high rate of correct prediction when trained on one season and tested on the other. Use of season specific models or inclusion of descriptive variable for the season in which data was collection is likely to improve the accuracy of prediction.

A model developed at earlier years can be used for the subsequent years given that the data was sampled over the same geographical region. However, it is possible that some dramatic changes (as flood or draught) in certain year can produce drastically different habitat conditions and the models developed for the years with different conditions can be less applicable. In this respect, models trained on data collected over several years can be capable of better generalisation.

5.3.4 Habitat issue

Macroinvertebrates occupy a variety of habitats in stream and have different adaptations for the specific conditions. According to the AusRivAs approach, models were developed separately for several habitats as riffle, edge, pool, etc. as it has been found that significantly different macroinvertebrate communities inhabit different habitats (Humphries et al., 1996), and within a given region, the differences among habitats are greater than differences between sites. Unless comparisons between sites are based on the same habitats, they may be confounded by the occurrence of different

habitats at each site (Parsons and Norris, 1996). Therefore, we can expect that models developed for one habitat will perform poorly when tested on another habitat. However, in theory there might be a situation when the decision must be taken but there is no exact match between the model available and simulation data. In order to test the assumption that models developed for one habitat are not applicable for another, I designed a following study.

Method

For this study I used two subsets of data, one collected from the riffle habitat and the other collected from the edge habitat. These two habitats are most often used for the assessment and prediction of macroinvertebrate communities and contain the largest number of data points in the data available. The goal of this study was to determine whether the model built on data collected from one habitat can be applicable for the simulations using data collected from the other habitat. I used the same 5 taxa as in previous studies and same architecture of ANNs, namely 50-10-1 architecture with 'tahn'transfer function. 10% of data were used as cross-validation subset to control overtraining and 30% of data were used for the validation of the models. All validation data were equalized to 50/50 ratio of presence and absence values.

Results

On average, models developed for each habitat had similar accuracy when tested on the subset from the same habitat as that for which they were built (75.59% and 77.16% for edge and riffle habitats respectively). When tested on the data from the habitat different to which they were built all models showed inferior performance. On average, reduction in the accuracy was 9.39%. However, models for Helicopsychidae were still capable of achieving 70% threshold of correct predictions when tested on the data set from the different habitat.

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.

Taxa	Trained: Edge Tested: Edge	Trained: Riffle Tested: Riffle	Trained: Riffle Tested: Edge	Trained: Edge Tested: Riffle
Leptophlebiidae	75.31	79.75	67.8	76.07
Gomphidae	70.28	66.89	65.2	65.35
Calamoceratidae	72.18	65.21	62.76	54.74
Philopotamidae	86.15	89.45	64.49	60.6
Helicopsychidae	84.58	84.49	76.84	75.99
Average	**75.59**	**77.16**	**67.42**	**66.55**

Discussion and conclusions

Macroinvertebrate communities collected from the different stream habitats have different characteristics. This fact was known previously and our research confirmed that this needs to be taken into consideration when building predictive ANN models. When models developed for one habitat are tested on another, accuracy decreases for about 10% in comparison when the models are trained and tested on the data from the same habitat. Habitat specific models are more accurate when simulated on the data from the same habitat, however, in some cases relatively accurate predictions are still possible using data from the different habitat.

5.4 Prediction of SOM defined groups: case study for the comparison of the evolutionary algorithms and supervised neural networks.

Even though ANN have clearly demonstrated their potential for ecological applications in terms of classification and prediction they store learned models in a highly distributed manner by means of connection weights, which bear little resemblance to human understanding of rules or concepts. By contrast, GA can be used for knowledge discovery by deriving predictive models or rule sets, which can easily be understood (Recknagel, 2001). Recknagel et al. (2002) compared applications of ANN and GA in terms of forecasting and understanding of algal blooms in Lake Kasumigaura (Japan). It was demonstrated that models explicitly synthesized by GA not only performed better in seven-days-ahead predictions of algal blooms than ANN models, but provided more transparency for explanation as well.

This study demonstrates and compares the use of both ANN and GA for the prediction of macroinvertebrate spatial assemblages in the stream system of Victoria. Both ANN and GA are applied in order to demonstrate prediction and explanation of patterns in spatial distribution of macroinvertebrate communities discovered by SOM (previously described in Chapter 4.1.2). The predictive and explanatory performance of both ANN and GA are also compared.

The main hypotheses for this study are:

1) It is possible to predict SOM defined clusters using MLP models.

2) GA based models are more accurate than MLP based models.

Methods

The stream database for this study was provided by the Victorian Environment Protection Authority, Australia. It contained abundances of 128 macroinvertebrate families sampled at 407 stream sites (only reference sites) between March 1990 and November 1998. Only the dataset collected in the edge habitats was used for this study. MLP and GA were applied in order to predict and explain occurrences of

macroinvertebrate assemblages based on 19 environmental variables utilising all environmental and macroinvertebrate data of the 407 stream sites.

The conceptual framework for this study is shown on Figure 5.3. The underlying GA was designed and implemented in C^{++} by Jason Bobbin (Department of Science, Defence and Technology (DSTO), Adelaide).

Prediction of the assemblage types by ANN

In order to predict the types of macroinvertebrate assemblages by means of a MLP a 25 x 407 data matrix was created. It considered the 19 environmental variables as inputs and the 6 spatial groups derived from SOM (see Chapter 4.1.2) for each of the sites as outputs. All these data were normalized into the range between 0 and 1. The MLP contained 19 neurons in the input layer, 10 neurons in the hidden layer and 6 neurons in the output layer. The sigmoid function was used as transfer.

The dataset was randomly subdivided into training subset (65% of the data), cross-validation subset (10%) and testing subset (25%). The accuracy of MLP reported in this paper is obtained from the simulation on the testing subset, which was not used for training purposes. The optimum training error was achieved by 1500 iterations.

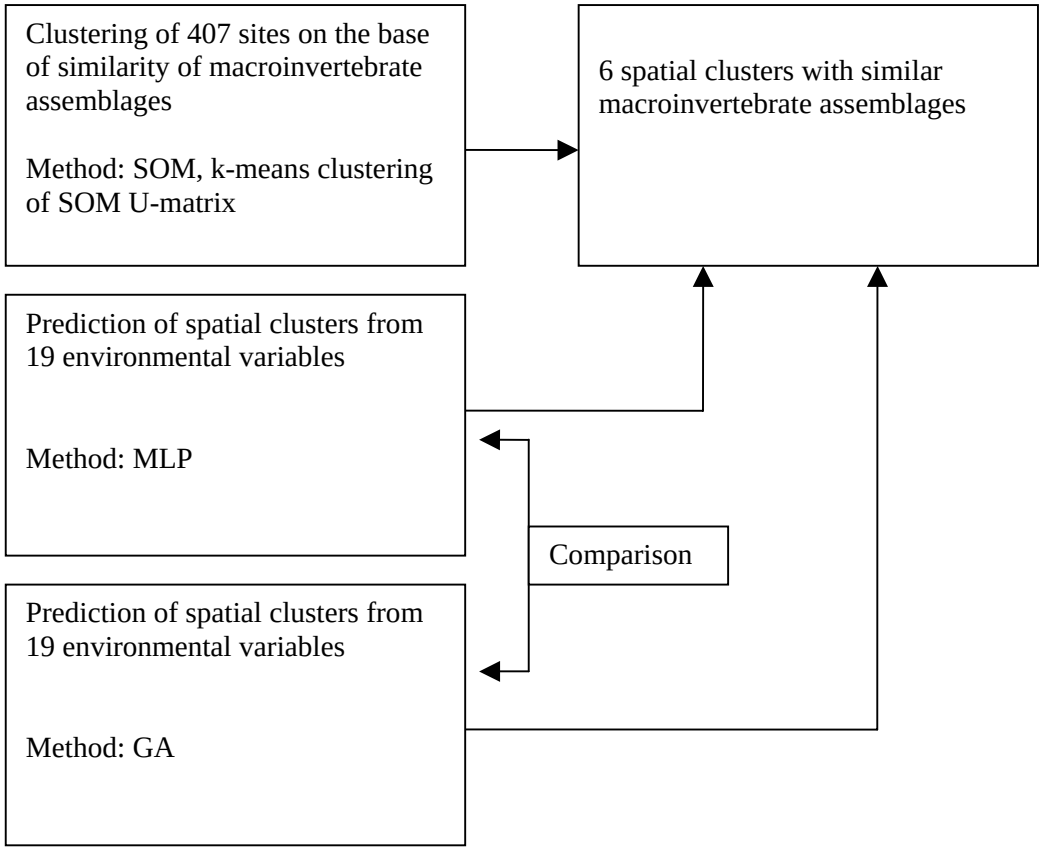


Figure 5.3. Conceptual framework for the study involving prediction and explanation of SOM defined groups using MLP and GA.

Prediction of the assemblage types by GA

A genetic algorithm GA consists of a population of individuals where each individual represents a model. Individuals are modified by mutation and crossover and the best individuals are selected to form a new population. Each new population is called a generation. In the context of the present paper a GA is used to evolve associations between physical and chemical properties of streams (attributes) and spatial clusters derived from partitioning of SOM U-matrix (outputs) based on similarities of macroinvertebrate assemblages. Attributes are associated with outputs by means of a classifier or rule.

Rules are combined to the rule sets by using a ripple-down structure shown in Figure 5.4. When a rule is true any consecutive horizontal rule is immediately tested. If a rule is not true then the consecutive vertical rule is tested. Horizontal arrows in Figure 5.4 represent exceptions to the rule to their left, and vertical arrows point to the rule to be tested if the current rule is not true. The last rule found to be true has its action implemented. If no true rule is found then the evolved default action is performed. Rule D in Figure 2 would have its action performed if and only if rule A is true, rule B

is not true and rule D is true. The approach used by the GA facilitates gradual evolution of the model by allowing mutation processes to slightly modify the model behaviour with exceptions to current rules. Information contained in the rules is

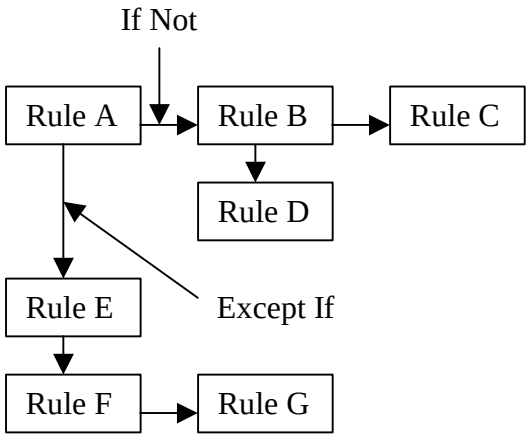


Figure 5.4. Structure of an evolved rule tree.

represented symbolically, where the symbols are associated with values in a parameter vector that is co-evolved alongside the rulesets. Each individual in the population is a complete ruleset. During each generation the structure of the ruleset is evolved by means of discrete operators (addition, subtraction and modification of the rules), and the parameters which define the values on the rules are modified by means of a self adaptive evolutionary algorithm (Schwefel, 1995; Baeck, 1996).

Results

The average percentage of correct predictions by the ANN of the six assemblage groups of macroinvertebrates as discovered by the SOM was 88.56 % while the average percentage of correct predictions by GA was 77.1 %. The mean squared error (MSE) and the percentage of correct predictions (PCP) for each group by the MLP and the GA is shown in Table 5.8.

Table 5.8. Mean square error (MSE) and percentage of correct predictions (PCP) by applications of ANN and GA for each of 6 groups.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
MSE (ANN)	0.07	0.13	0.05	0.15	0.11	0.03
PCP (ANN) %	91.17	82.35	94.11	83.33	84.31	96.07
PCP (GA) %	93.42	96.97	84.44	89.47	53.17	80.90

All the values considered by GA ruleset fall into the minimum and maximum range within SOM cluster (Table 5.9, only group 1 is considered here). However, taking into the consideration that group 1 is heterogeneous and spatially scattered (see. Figure 4.2) the range and averaged values for the predictor variables can give only very approximate and rough idea about the combinations of variables contributing to the occurrence of this particular macroinvertebrate assemblage. On the contrary, rules resulted from GA application give more detailed and directional description of the physical variables, allowing for the spatial heterogeneity. For example, average value of variable Latitude is –37.52. In the ruleset it takes two directions: Latitude IS NOT between -38.2 and -37.4 (Latitude >-38.25) and Latitude is between -37.4 and -36.4. This is the case for the other variables as well. Vegetation category has average value of 3.52, in the ruleset its two directions are Vegetation Category >2.07 and Vegetation Category < 2.07. The same stands for Shade and Alkalinity. Variables Altitude and Macrophyte category have only one direction each and in agreement with averaged values for the cluster.

Discussion and conclusions

In this study two machine learning methods have been applied and compared: Multy-Layer Perceptron neural network and Genetic Algorithm. Traditionally in ecological applications neural network models are used for the prediction of taxa occurrence or abundance from a set of environmental variables. In this study we explored the question if it is possible to predict occurrence of the unit larger than separate taxon, in our case defined pattern in abundance and co-occurrence of all taxa recorded. Even though, separating these patterns as distinct clusters or groups might be artificial it appears that both MLP and GA are well capable of predicting these groups from the set of environmental variables.

Contrary to the previous finding of Recknagel et al. (2002) showing that in time-series case predictive power of GA was higher than that of ANN, in our case ANN outperformed GA on approximately 10% (hypothesis 2 is not true), although both methods were able to meet commonly acceptable 70% threshold of correct predictions (hypothesis 1 is true).

The sites used for the analysis were reference sites, presumably least affected by the agricultural practices or urban developments. We assume that in this case it should be easy to explain patterns in abundance and co-occuring of various macroinvertebrate families by a range of environmental variables. We tried to do this by examining simply data ranges within SOM clusters and GA rule set for the environmental variables likely to be important in distinguishing Group 1 from the other groups. In terms of explanatory power GA is commonly considered as offering more transparency by the generation of rules providing the directional explanation for environmental heterogeneity, while neural networks are thought to be ‘black’ or ‘grey’ box technique.

SOM component planes (see Chapter 4.1.2) provided easy and highly visual way to assess relationship between variables and suggest which ones are likely to be of importance in shaping macroinvertebrate assemblages within each group. Although, this approach can be criticized as to a certain extent qualitative and intuitive, we suggest that it still can be valuable when quick and visual assessment of data is

needed. GA rules provide a qualitative approach but are not easy to follow and understand, it might be useful where more in-depth assessment is needed.

The prediction of defined macroinvertebrate assemblages instead of separate taxon can be used as an extension of the referential approach outlined in the introduction. If the type of macroinvertebrate assemblage predicted under particular environmental conditions do not match the one actually found, it might be then compared against other possible assemblages indicative for various stresses as increased salinity, turbidity, nutrient load, etc.

In conclusion, this study demonstrated that ANN and GA provide different approaches to the defined problem and neither of them could be clearly favored in the context. Both methods were able to predict spatial groups of macroinvenvertebrates from environmental variables with high efficiency and provide an explanation from slightly different angles. We recommend the use of the both methods in combination for achieving the most accurate predictions and the highest explanatory power.

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.

	VARIABLE	MEAN (MIN/MAX)	VARIABLE	MEAN (MIN/MAX)
Descriptive Statistics of Group 1 by SOM	LATITUDE	-37.52 (-39.072/ -36.36)	WIDTH	0.61 (-0.20/ 1.48)
			BEDROCK%	4.95 (0.00 / 60.00)
	LONGITUDE	146.25 (143.28/148.47)	BOULDER %	13.81 (0.00 / 65.00)
			COBBLE %	26.70 (0.00 /70.00)
	REACH PHI	-3.10 (-7.22 / 4.03)	PEBBLE %	12.07 (0.00 / 50.00)
	SUBSTRATE HETEROGENITY	2.68 (1.25 / 5.00)	GRAVEL %	11.47 (0.00 / 42.50)
	VEGETATION CATEGORY	3.52 (2.00 / 4.00)	ALKALINITY	0.96 (0.57/ 1.71)
			MACROPHYTE TAXA	0.25 (0.00 / 0.88)
	SHADE	3.07 (1.00 / 5.00)		
	DISTANCE FROM SOURCE	0.89 (-0.70/ 2.00)	MACROPHYTE CATEGORY	0.06 (0.00 / 0.60)
	SLOPE	1.19 (-0.30/ 2.52)		
	ALTITUDE	2.63 (1.30/ 3.23)		
	CATCHMENT AREA	1.51 (-0.10 / 3.23)		
Rule Set for Group 1 by GA	IF 0.51 < CATCHMENT AREA < 1.02 OR IF MACROPHYTE CATEGORY < 0.13 AND VEGETATION CATEGORY > 2.07 AND ALTITUDE > 1.88 AND SHADE < 0.323 OR SHADE > 1.17 AND LATITUDE > -38.2 OR LAT < -37.4 OR IF MACROPHYTE CATEGORY < 0.13 AND VEGETATION CATEGORY < 2.07 AND LATITUDE > -38.25 AND ALTITUDE > 1.88 AND ALKALINITY < 1.08 AND -37.4 < LATITUDE < -36.4 OR IF MACROPHYTE CATEGORY < 0.13 AND VEGETATION CATEGORY > 2.07 AND ALTITUDE > 1.88 AND 0.323 > SHADE > 1.17 AND -37.4 < LATITUDE< -36.4 AND 1.09 < ALKALINITY < 1.39 OR IF ALKALINITY > 1.087 AND VEGETATION CATEGORY > 2.07 AND -3.38 < REACH PHI < -2.01 THAN ASSEMBLAGE 1 ELSE ASSEMBLAGES 2 TO 6			

Chapter 6

Defining the relationships between water quality and macroinvertebrates using sensitivity analysis with MLP and SOM component planes

6.1 Investigation into stability and quantitative applicability of the sensitivity analysis using supervised neural networks

The sensitivity analysis with predictive ANNs is most commonly used for two purposes: 1) to study contribution of input variables in the network in order to determine the most important inputs and reduce the complexity of the network, 2) to study the response of biotic variables (as taxa distribution or changes in taxonomic richness) to the changes in environmental parameter (sensitivity curves).

A variety of methods has been proposed and compared to study the contribution of the variables in ANN models. Gevrey et al. (2003) compared seven methods finding Partial Derivative method (see Chapter 2 for description) most useful followed by the 'Profile' method. Dedecker et al. (in press) compared three methods including Partial Derivatives and 'Profile' and found that the difference in sensitivity and stability of the methods ware rather small. However, among all other methods, the 'Profile' method is the only technique that provides two elements of information on the contribution of the variables. This method presents the order of contribution of the different environmental variables, and gives direct interpretation of the effect of environmental variables as river characteristics on the occurrence of taxon. The other methods only able to classify the variables by the order of their importance, in other words, to reveal their contribution to the output (Dedecker et al., in press).

Only very few studies have been conducted with the purpose to study variability of the sensitivity analysis. Olden et al. (2004) compared several methods for quantifying variable importance in ANNs as interpretation of connection weights, Garson's algorithm, Partial derivatives, Input perturbation, etc. using Monte Carlo simulation experiments on simulated data. It has been shown that accuracy of the methods estimated as percentage of average similarity between true and estimated outputs

varied from 92% to less than 50%. Sensitivity analysis involving varying each input variable and keeping all other variables at fixed values among other methods was only successful at identifying the true importance of the two most influential variables out of 5.

Sensitivity analysis using predictive neural networks has been previously used to investigate relationships between environmental variables and the occurrence of macroinvertebrates. A high correspondence between relationships discovered by sensitivity analysis and those previously know from the application of the other methods has been shown (Marshall et al., 2002; Huong et al., 2001, 2003).

Sensitivity curves produced by plotting the predicted output against environmental gradient in question often used to study the response of biotic variables (as taxa occurrence or changes in taxonomic richness) to the changes in environmental parameter. It is an attractive method with a potential to identify taxa specifically sensitive to various anthropogenic stressors as organic pollution, salinity, turbidity, etc. However, the accuracy and consistency of sensitivity curves has not been consistently studies. It has been observed (Peter Noble, Civil and Environmental Engineering, University of Washington, USA, personal communication) that individual neural networks with the same architecture trained on the same dataset can produce very different sensitivity curves. Random nature of weights in ANN means that individual models have different sets of weights and indeed can vary in their sensitivity and predictive ability. Predictability and quantitative consistency of the sensitivity analysis using real ecological data is still unclear.

In this chapter I applied ‘Profile’ method to investigate stability and quantitative applicability of the sensitivity analysis using supervised neural networks. Is it appropriate to use sensitivity analysis as a quantitative tool? How much variability is associated with random nature of weighting in the neural network? In order to answer this question I designed a following study.

Data and Method

In order to estimate the degree of variability associated with the sensitivity analysis, I built, trained and tested 10 models of the same architecture for 3 randomly chosen taxa with medium frequency of occurrence (40-50%). I utilised Brisbane catchment subset (150 samples), collected from the edge habitat with 18 predictor variables previously used for the Optimisation study (see Chapter 5.3). Standard default architecture was used (18 input nodes, 6 hidden nodes, 1 output node, tahn transfer, training with momentum), cross validation was applied only for the first run for each taxa to estimate optimum number of epochs before overtraining is likely to occur. For the subsequent runs I used 70% of data for training and 30% for testing. Sensitivity analysis was performed on training subset but accuracy of each model was estimated using testing subset. For this analysis I investigated sensitivity of three taxa (Cladocera, Gripopterigidae and Leptophlebiidae) to turbidity (NTU) and conductivity ($\mu S\ cm^{-1}$) by changing mean value of the predictor variable by 10 standard deviations and calculating output 50 steps in each direction. The sensitivity analysis implemented in Neuro Solution 4.0 software package provides two main types of outputs: sensitivity curves and the estimation of the predictor importance expressed as

the percentage of change in the output in response to the changes in the particular input.

Results

Variability in sensitivity curves

All individual models were reasonably accurate when simulated on testing subset (30% of data), with the average percentage of correct predictions 72.9%, and accuracy was largely consistent from model to model (Table 6.1) with relatively small standard deviation.

Table 6.1. Estimation of mean accuracy and standard deviation of individual models.

Taxa	Mean accuracy over 10 runs (% correct predictions)	Standard deviation
Leptophlebiidae	66.95	3.94
Gripopterigidae	80	3.73
Cladocera	71.78	5.74

However, there was a significant variability between curves produced by the individual models. As can be observed in Figure 6.1, both box (20-80%) and whisker (minimum and maximum values) spreads are very large on the majority of the plots with the exception of response of Leptophlebiidae to the conductivity.

Traditionally outputs with value less than 0.5 are interpreted as absence and those higher or equal of 0.5 as presence. In this respect, the majority of the graphs does not make any sense quantitatively. Output value for Cladocera actually never reaches 0.5 mark indicating that it is never present, which is not true in the reality. Minimum and maximum outputs of the individual models range from way below 0.5 to approaching 1 in many cases. In this sense use of the sensitivity curves to determine conductivity or turbidity thresholds exceeding taxon-specific tolerance would be highly erroneous. However, median values of all outputs are consistently increasing or decreasing in all cases except for Cladocera against conductivity, where no any definite response can be observed. The same is true for the maximal and minimal values. It appears that individual models produced sensitivity curves of the same shape (decreasing or increasing along the turbidity/conductivity gradient), but the quantitative range of those curves was different in each particular case.

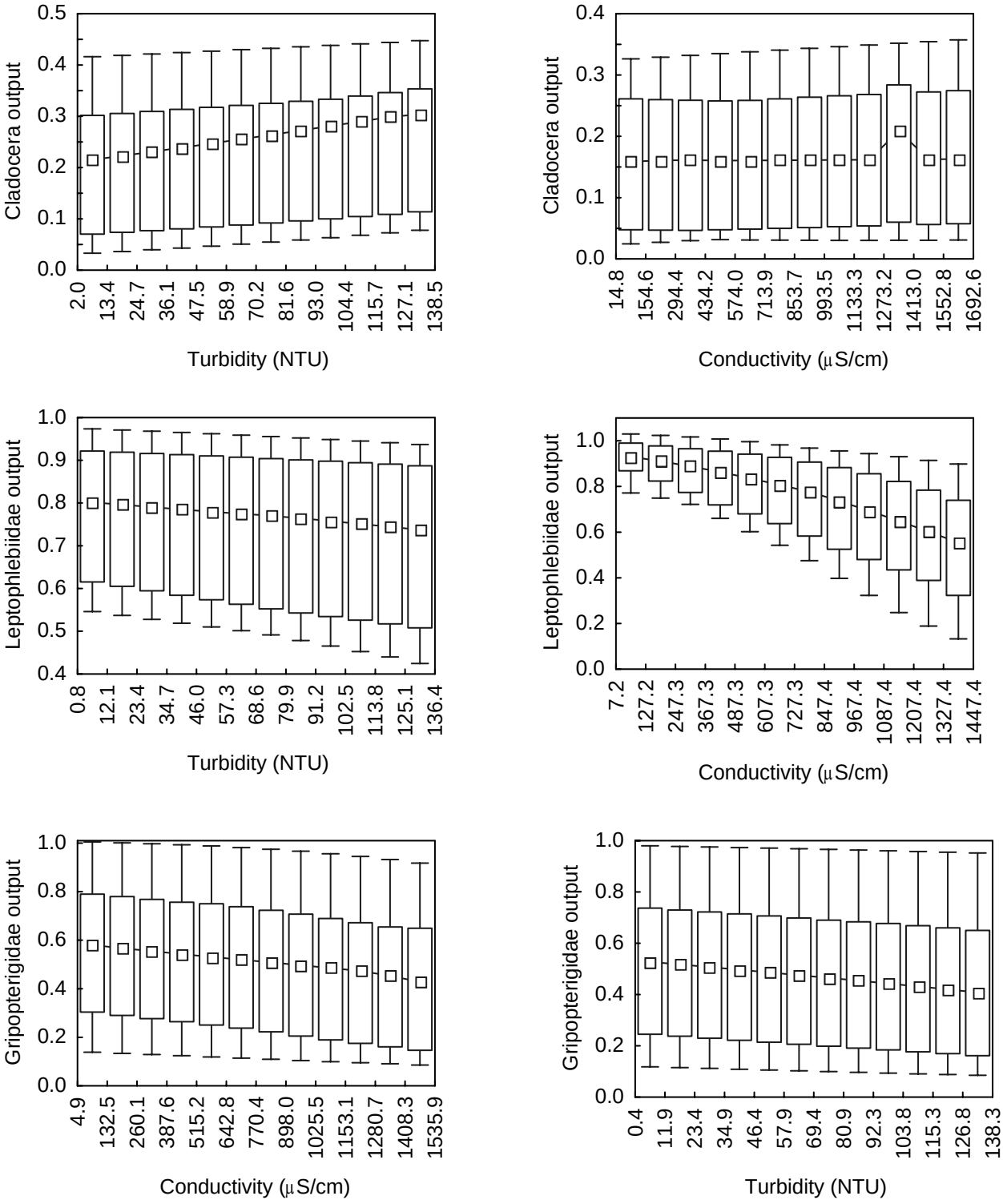


Figure 6.1. Box and whisker plots showing variability between individual models used for the sensitivity analysis of the relationship between three taxa and two water quality variables. Center – median value, box - 20-80%, whiskers – minimum and maximum values.

Variability in the importance of predictor variables

As seen from the Table 6.2, the variability associated with the random nature of weighting in the neural network model is quite significant. In many cases standard deviation was almost as big as the mean value. Clearly, it would be highly misleading to base any conclusions about predictor importance on the single run of the individual model even though the overall accuracy of the model might be good.

Table 6.2. Variability in the estimation of the predictor importance.

Variable	Leptophlebiidae		Gripopterigidae		Cladocera	
	Mean	St.dev	Mean	St.dev	Mean	St.dev
Season	0.10	0.09	0.27	0.17	0.19	0.03
Depth (m)	0.31	0.18	0.21	0.17	0.47	0.16
Max Velocity (m/s)	0.25	0.21	0.18	0.13	0.32	0.11
Mean phi	0.01	0.01	0.01	0.00	0.01	0.00
Latitude	0.34	0.20	0.49	0.19	0.34	0.11
Longitude	0.54	0.24	0.32	0.24	0.44	0.14
Altitude (m)	0.00	0.00	0.00	0.00	0.00	0.00
Stream Order	0.04	0.03	0.04	0.03	0.03	0.01
Slope (km/m)	10.38	8.83	25.29	8.20	3.68	3.06
Distance From Source (km)	0.00	0.00	0.00	0.00	0.00	0.00
Ratio of a / b	0.18	0.16	0.92	0.34	0.14	0.04
Water Temp (°C)	0.01	0.01	0.02	0.01	0.02	0.00
Conductivity (µS/cm)	0.00	0.00	0.00	0.00	0.00	0.00
DO (mg/L)	0.02	0.01	0.04	0.03	0.04	0.01
pH	0.06	0.06	0.03	0.03	0.02	0.01
Turbidity (NTU)	0.00	0.00	0.00	0.00	0.00	0.00
Total N (mg/L as N)	0.10	0.07	0.02	0.01	0.03	0.02
Total P (mg/L as P)	0.28	0.17	0.08	0.10	0.52	0.35

Discussion and conclusion

Estimation of the importance of each predictor variable has a potential to reduce the number of inputs therefore making the model simpler and more transparent. Huong et al. (2003) found an improvement in overall performance of the MLP models after the reduction of the number of inputs based on the output from the sensitivity analysis. However, it appears that the results from this approach can be highly variable and care should be taken with their interpretation. The degree of this variability might be associated with the nature of data used for training, and can differ from dataset to dataset.

If we would attempt to estimate the conductivity threshold exceeding tolerance of Leptophlebiidae and causing it to disappear, some of the runs will show value of about 700 µS cm^{-1}, however, according to the median value of 10 runs Leptophlebiidae remains present even at maximal conductivities used for the sensitivity analysis. In other words different runs might indicated the taxa present or absent at the same conductivity based on the variability of weighting used by each individual model, even though the accuracy of the all resulted models is comparable.

As indicated by the results above, the quantitative range of the sensitivity curve is affected by the particular set of weights in the trained network and can vary to a significant degree from one network to another, however, the general trend of the sensitivity curve (rising, falling, flat) essentially stays the same which indicates relative reliability of sensitivity analysis when the purpose is to detect a general qualitative trend of the taxa response to the changes in a particular condition.

When sensitivity analysis is used to estimate response in numeric terms a number of runs should be used in order to account for the variability associated with the random nature of weighting in the neural network. Estimation of the predictor importance appears to be totally unreliable, at least using the method described and this particular software package. The main conclusion from this study is that care should be taken when using sensitivity analysis as quantitative method especially based on the output from a single model, however, when used qualitatively, sensitivity curves appear to be an interesting tool for the investigation of relationships between variables.

6.2 Response of stream macroinvertebrates to the changes in salinity and the development of a Salinity Index

Introduction

In the past several decades increases in salinity due to the human disturbance to the natural hydrological cycle have caused increasing problems in Australia. The area estimated to be affected by dryland salinity in Queensland (QLD) is 48 000 ha and this figure could increase to 3.1 million hectares by the year 2050 (Gordon, 2002). A number of streams and wetlands have been affected by rising salinity leading to significant changes in flora and fauna (Hart et al., 1991). Secondary salinisation can affects aquatic systems in a multiple ways, including direct toxic effects, changed chemical processes and loss of habitat in the water, riparian zones and adjacent floodplains. Current understanding of the resilience of freshwater biota to these impacts is limited and more research is needed (James et al., 2003).

Several authors have studied the occurrence of macroinvertebrate taxa in streams and rivers of southern and western parts of Australia (Williams et al., 1991; Metzeling, 1993, Kay et al., 2001; Bunn and Davies, 1992; Kefford, 1998). Different responses have been observed, but there is a general acceptance that freshwater ecosystems experience little ecological stress at the electric conductivity (EC) levels below 1500 µS cm^{-1} (Hart et al., 1991). Marshall and Bailey (2004) conducted field experiments to examine the effect of short-term releases of saline wastewater on stream macroinvertebrate communities. Significant reduction in the abundance of some species (Ferrissia tasmanica, Baetis sp. 5) were observed at 1500 mg L^{-1} (approximately 2205 µS cm^{-1}). However, Kefford et al. (2003) showed that even though the majority of macroinvertebrates is highly tolerant under acute exposures, sublethal effects might occur at salinities as low as 480 µS cm^{-1}. Therefore, it is possible that long-term progressive salinisation might affect macroinvertebrate communities indirectly and the effective EC concentrations might be lower than currently accepted. In this respect analysis of a field data collected on wide

geographical scale and over many years can provide an important insight into the effects of salinisation including all potential mechanisms by which salinity may impact on freshwater organisms.

Statistical methods as correlations, multiple linear regression, generalised additive models, logistic regression, principle component analysis (PCA) and cluster analysis are commonly used to study species distribution along an environmental gradient. However, there are two potential problems associated with this approach. Firstly, it is often difficult to know for certain that observed changes in biota are caused by the factor in question as salinity gradient and not by a multitude of the other underlying factors. Secondly, our confidence in the results is often limited by the method's inability to meet a number of assumptions, such as statistical distribution, independence of variables and linearity of the relationships.

Use of machine learning methods as ANNs allows to overcome limitations of many traditional statistical methods such as data distributions or non-linearity. Sensitivity analysis with ANNs is a relatively new method for the estimation of complex non-linear relationships between environmental factors and the taxa distribution. It allows simulation of the changes in biota specific to a variable in question when all the other variables in the model are kept static. This means that we can have some confidence that predicted changes are caused by the factor in question (as conductivity) and not underlying natural or other anthropogenic gradients.

Aquatic systems affected by the secondary salinisation often also experience poor water quality as higher concentration of nutrients, suspended particulate matter and toxicants (Hart et al., 2003). In this respect it might be advantageous to consider the effect of EC in combination with the other water quality factors. Canonical Correspondence Analysis (CCA) is one of the newer methods, which allows to describe and visualise relationships between multiple environmental variables and biota (ter Braak and Verdonschot, 1995). Partial CCA (ter Braak, 1988) is of particular interest to us as it permits to partial out the effect of covariables (as temporal and natural variability) and use residuals to examine variables of interest (as EC and other water quality variables). The method is also relatively robust in term of statistical distribution.

The aim of this study was to investigate changes in macroinvertebrate communities associated with a conductivity gradient in streams and rivers. We are not aware of any similar study conducted in QLD. The analysis of state-wide data can provide an interesting broad scale insight into the differences between structures of macroinvertebrate communities under varied levels of stress from the secondary salinisation. This study utilises sensitivity analysis using ANNs to estimate conductivity tolerances of the specific macroinvertebrate taxa and to develop an index reflecting changes in macroinvertebrate communities as a result of changes in the conductivity level. Partial CCA is used to provide an additional insight into the effect of EC and the other water quality factors on the structure of the macroinvertebrate communities.

Data

The data used for the current study were collected each spring and autumn from 1994 to 2002 from 1008 sites (Figure 6.2) by the NR&M. The data from two different

habitats were considered separately in order to reduce the effect of natural variability as much as possible, plus there is a geographical difference between subsets as many samples from the edge habitat were collected in western inland areas when the majority of samples from the riffle habitat were collected in the coastal regions. A salinity level was measured as EC (µS cm^{-1}) adjusted for temperature. We used salinity categories given by Williams (1967), see Figure 6.2.

The purpose of this study was to identify general trends using common macroinvertebrate taxa with state-wide distributions. We used taxa occurred at more than 5% of all samples. Several taxa with limited geographic distributions were excluded from the analysis in order to avoid confounding the results with their natural geographical distribution range. This left 57 taxa (mostly family level) from the edge habitat and 60 taxa from riffle habitat for the analysis.

Methods

I implemented a variety of methods in order to estimate the sensitivity of macroinvertebrates to various levels of EC: SOM component planes, frequency distribution along the conductivity gradient, sensitivity curves produced by Multi-Layered Perceptron Neural Network and ordination by Canonical Correspondence Analysis. This approach has been adopted in order to 1) validate output of one method by the outputs from the other methods, 2) compare the potential of the methods themselves.

The first questions I asked was: what is the relationship of conductivity with other abiotic and biotic variables? Which natural or water quality gradients are likely to have a gradient similar to that of the conductivity? To answer those questions I built an SOM neural network using combined set of 28 environmental variables (Table 6.3) plus biotic variables as taxonomic richness, PET richness and SIGNAL index. Using component planes I compared distribution of the biotic and abiotic variables in relation to each other.

The next task was to detect changes within macroinvertebrate communities caused by the changes in conductivity, and ensure as much as possible that those changes are not caused by the factors other than EC. I implemented the following logical sequence to address this task:

1) Identify EC sensitivity of each taxon using SOM component planes, MLP sensitivity curves and taxon's frequency distribution along the EC gradient.

2) Assign a score to each taxon according to its EC sensitivity.

3) Calculated cumulative score (Salinity Index (SI)) as a measurement of the community sensitivity.

4) Investigate the relationship of the SI with conductivity and the other variables, identify possible confounding factors and practical applicability of SI.

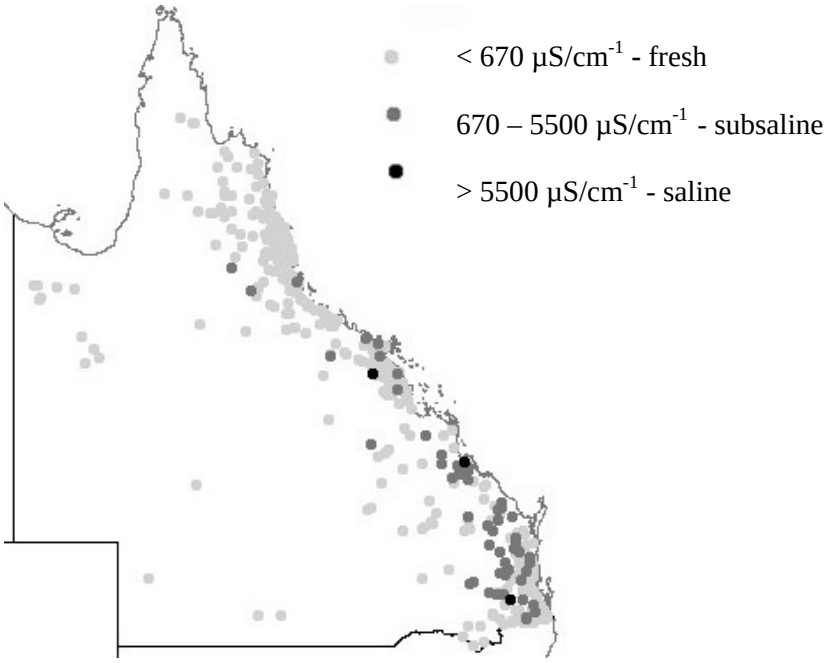


Figure 6.2. Distribution of conductivity values in Queensland dataset (ranges of conductivity taken from Williams (1967)).

In order to visualize SOM component planes I used occurrence pattern of macroinvertebrate taxa and conductivity as an input for SOM. Taxa frequently occurring in the areas with high conductivity were labeled as very tolerant (vt), taxa frequently occurring in the areas with low conductivity as sensitive (s) and taxa without any distinctive pattern in relation to conductivity as tolerant (t). Two SOMs were built for each habitat, riffle and edge.

To analyze the frequency of occurrence of macroinvertebrate families along EC gradient, I sorted the data by ascending conductivity and divided it into 25 equal sized bins without any regard to when and where the data were collected. Mean percentage of taxa occurrence in each bin was calculated and plotted as continuous frequency curves. According to the shape of the frequency curve, the trend of occurrence for each taxon was described as ‘increasing’ when the frequency of occurrence generally increased with increases in conductivity, ‘decreasing’ when the frequency of occurrence decreased with increases in conductivity or as ‘no visible trend’ for flat, irregular or unimodal shapes of the frequency curve. Even though, this simple analysis provides some insight into the salinity preferences of macroinvertebrate taxa, it may also reflect trends in underlying natural variability or trends in the other water quality parameters as well as salt tolerance. In order to ensure that there was no interference from the other variables, I used ‘Profile’ sensitivity analysis to qualitatively estimate taxa specific responses to changes in one variable while all the other variables were kept at their respective means.

In order to estimate salinity tolerance of each macroinvertebrate taxon I built 117 taxa-specific MLP models (57 for the edge habitat and 60 for riffle habitat) with the

following architecture: 25 neurons in the input layer associated with predictor variables, 6 neurons in the hidden layer and 1 neuron in the output. The variables used as predictors (only abiotic variables were used) are shown in Table 6.3. Data was range standardised between 0 and 1.

'Tahn' was used as a transfer function and networks were trained using feed forward propagation with momentum (Principe et al., 2000). Optimum number of epochs during which overtraining does not occur was determined using cross-validation method with 10% of the data. Quality of each model was estimated using Mean Square Error (MSE) between predicted and actual output.

Sensitivity analysis was performed using the following method. The mean value of conductivity was changed by 10 standard deviations and calculated 50 steps in each direction, while all other predictor variables were kept at their respective means. Each taxon specific model was simulated and continuous output for each taxon has been plotted against conductivity. Trends in probability of taxon occurrence with increase in conductivity was described as 'increasing', 'decreasing' or 'no visible trend' for flat, irregular or unimodal curves.

Table 6.3. List of abiotic variables used for the study with Product Moment correlation with conductivity.

Variable	Statistical correlation with conductivity
Total taxonomic richness	-0.06
PET Richness	-0.17
SIGNAL score	
Depth (m)	-0.14
Velocity-max (m/s)	-0.18
Bedrock (%)	0.004
Boulder (%)	-0.04
Cobble (%)	-0.14
Pebble (%)	0.02
Gravel (%)	0.16
Sand (%)	0.01
Silt/Clay (%)	0.08
Water Temperature (^{0}C)	-0.08
DO (mg L^{-1})	-0.03
pH	0.28
Turbidity (NTU)	-0.02
Alkalinity (mg L^{-1} CaCO3)	0.61
Total Hardness (mg L^{-1} CaCO3)	0.91
Total N (mg L^{-1} as N)	0.18
Total P (mg L^{-1} as P)	0.16
Latitude	-0.33
Longitude	0.29
Altitude (m)	-0.06
Stream Order	-0.03
Slope (km/m)	-0.12
Distance From Source (km)	0.02
Ratio of mean wet season monthly rainfall to mean dry season monthly rainfall	-0.25
Mean annual rainfall (mm)	-0.32

Salinity Sensitivity Score and Salinity Index

All taxa were assigned the following Salinity Sensitivity Scores (SSS): '10' for sensitive, '5' for generally tolerant, and '1' for very tolerant. I mainly used the outputs of sensitivity analysis as the method least affected by the interference of natural variability or other environmental factors. Outputs of frequency curves and SOM component planes were used for the comparison. The following rules were used to determine which score should be assigned (quantitative cut-off values are based on mean conductivities in given habitat dataset with the gap of 50 µS cm^{-1} to provide for some additional distance between sensitive and very tolerant taxa):

Sensitive: IF shape of sensitivity curve = "Decreasing" and MEAN conductivity <= 300 µS cm^{-1} (edge) 250 µS cm^{-1} (riffle)

Very tolerant: IF shape of sensitivity curve = “Increasing” and MEAN conductivity > 350 µS cm^{-1} (edge) 300 µS cm^{-1} (riffle)

Generally tolerant: All taxa which fell neither into sensitive nor very tolerant categories were assigned score “5”.

These rules were used only as a general guide, in some cases exceptions were made. For example, we had to take into consideration maximal conductivity at which taxa occur as some taxa show increasing trend but were not found in high conductivities. We labelled taxa as generally tolerant in all cases where results were inconclusive or controversial. Using the SSS of all taxa in a sample we calculated a Salinity Index similar to a chloride contamination index (CCI) suggested by Williams et al. (2000).

Salinity index (SI) $=(\Sigma X_i \times SSS_i)/N$

Where X_i =1 if taxon *i* was present, X_i=0 if absent

SSS_i = Salt Sensitivity Score of taxon *i*

N= the total number of taxa in the sample

SI can theoretically vary from a value of 1 when all the taxa in a sample are highly tolerant to a value of 10 with all taxa being sensitive. In practice we would expect opportunistic taxa present in both, unimpacted and impacted sites keeping the total score less then 10 and higher than 1. We used box and whisker plots to demonstrate changes in percentage of groups with different salinity preferences along the conductivity gradient. The data was sorted by ascending conductivity and split into 12 equal sized bins with no consideration of the site location, year, etc. The number of the bins was chosen arbitrary.

To test whether SI indeed reflects changes in macroinvertebrate communities mainly due to the changes in conductivity and not the effect of other wide spread stressors such as concentration of nutrients, we isolated a subset of data with otherwise good water quality: turbidity < 5 NTU, total nitrogen <0.375 mg L^{-1}, total phosphorus<0.05 mg L^{-1}, pH between 6.5 and 9 and dissolved oxygen > 5 mg L^{-1}. These values are taken from the water quality guidelines for the protection of aquatic ecosystems (Bloedel et al., 2000). This analysis excluded water quality parameters other than conductivity as potential cause of the observed changes in macroinvertebrate communities. However, it is still possible that these changes related to natural factors as rainfall, distance from source, flow, etc. In order to exclude this possibility we used partial CCA.

Partial CCA

Partial CCA was used to examine relationships of water quality variables to each other and macroinvertebrate communities, having excluded the influence of natural factors, flow and temporal variability. First we used CCA to determine which variables out of the set of 25 (see Table 6.3) were significant in structuring macroinvertebrate communities. Forward selection procedure (ter Braak and Verdonschot, 1995) with 999 Monte Carlo permutations was used to test the significance of each variable. Only variables with $p<0.01$ were used for the

subsequent analysis. The significant variables were divided into two sets: 1) covariables describing temporal, spatial and natural variability (as season, year, rainfall, distance from source, etc.), 2) water quality variables as EC, water temperature, pH, turbidity, etc.). Using partial CCA we excluded temporal and natural variability and built two separate biplots for each habitat reflecting only the effect of water quality variables.

Results

Analysis of relationships between salinity and environmental variables using SOM component planes.

Figure 6.3(a) shows a component plane (riffle habitat) for conductivity with values expressed as different shades of gray, the darker shade the higher the conductivity values. The second subplot (b) shows "hits" diagram, with sites

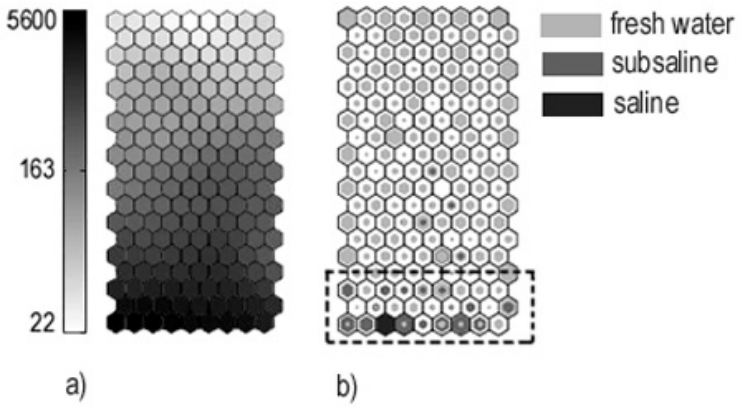


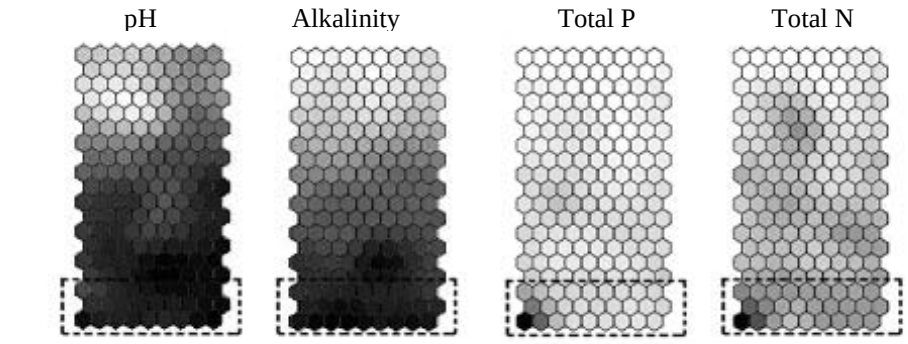
Figure 6.3. a) SOM component plane for conductivity (µS/cm), b) SOM grid with hits divided according to the classification by Williams (1967).

divided into three categories according to the classification given by Williams (1967), where waters with conductivities less than 670 µS cm^{-1} are considered as fresh, 670-5500 µS cm^{-1} as subsaline, and more than 5500 µS cm^{-1} as saline. Hits diagrams are used to show in graphic form the distribution of the best matching units for a given data set (Vesanto et al., 2000). The more best matching units for the given subset located in the cell the bigger the marker (see Chapter 3 on 'hit' diagrams). The number of 'hits' is expressed as size of the diamond inside each cell. The area of the SOM containing the highest number of 'hits' from streams falling into the subsaline and saline categories is outlined with broken line.

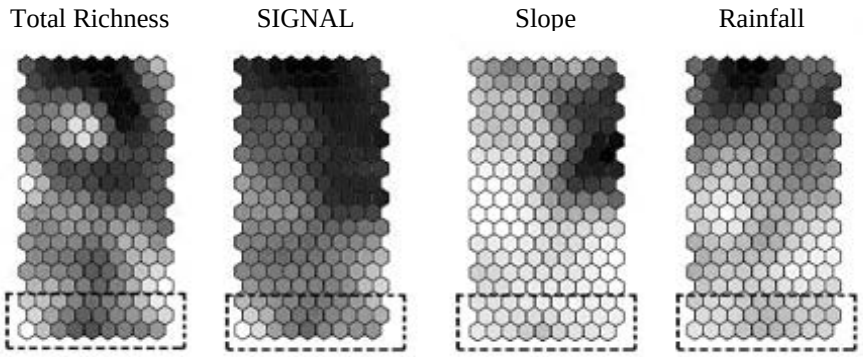
Figure 6.4 shows selected component planes with variables, which appear to have some relationship with EC (Product Moment correlation for all variables is shown in Table 6.3). Mean annual rainfall and ratio of wet season monthly rainfall to dry season monthly rainfall have some negative correlation trend in relation to conductivity according to SOM planes (Figure 6.4), with statistical correlation values –0.33 and –0.25 respectively. Out of all variables describing topography and location of the site (Distance from Source, Stream order, Altitude, Slope) only Slope appears

to have some negative correlation with conductivity (see Figure 6.4). pH and alkalinity were positively correlated with conductivity, and total Nitrogen and Phosphorus showed some positive correlation trend as well but not for all sites with high salinity.

Variables describing biological diversity and ecological state (total taxonomic richness, PET richness and SIGNAL score) all showed diminishing values towards increasing salinity, although this relationship is not straightforward (Figure 6.4). In the case of total taxonomic richness the statistical correlation value was insignificant. SOM component plane for total taxonomical richness shows that even though



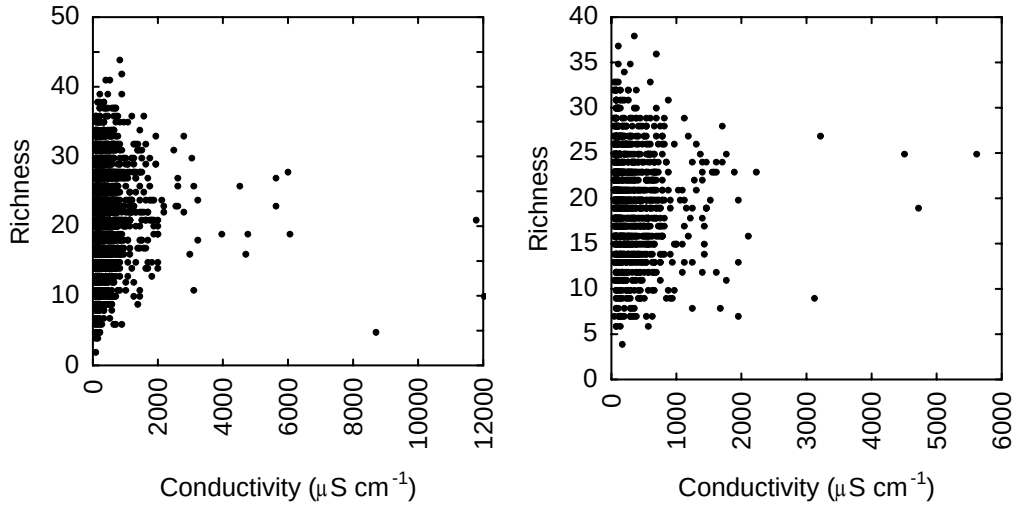
a)



b)

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)).

the taxonomical richness is the highest at the sites with low salinity it is not necessary the lowest in the areas with high salinity, in other words, macroinvertebrate communities with average taxonomic richness can be found under conditions of both, high and low salinity. Scatterplots of taxonomic richness versus EC for both habitats confirms this observation (Figure 6.5).



a) b)

Figure 6.5. Scatterplot of the taxonomic richness versus conductivity for: a) edge and, b) riffle habitats.

Exploring EC sensitivity of macroinvertebrate taxa in relation to conductivity.

Using SOM component planes

Selected component planes illustrating occurrence patterns of the different macroinvertebrates in relation to conductivity are shown on Figure 6.6 (riffle habitat). The taxa showing the most prominent negative trend towards rising conductivity were Aeshnidae, Gripopterigidae, Diphlebiidae, Gomphidae, Ptilodactilidae (Figure 6 (a)), they are marked as Sensitive (s) in Table 6.5. The other taxa showing similar trend and marked as (s) as well are: Corydalidae, Hydrobiosidae, Helocopsychidae and Tipulidae. Figure 6.6 (b) shows selected taxa with the opposite, positive occurrence pattern towards rising conductivity. Those taxa (Ostracoda, Dytiscidae, Gyrinidae, Corixidae, Veliidae, Hydrophylidae, Copepoda, Atyidae, Hibrobiidae, Chyronominae) are labelled as Very Tolerant (vt) in Tables 6.4 and 6.5. The rest of the taxa do not show any particular pattern or show mixed pattern and further labelled as ‘Generally Tolerant’ or ‘t’. Taxa from the edge habitat have been analysed the similar way and only resulting labels are shown here (Table 6.4).

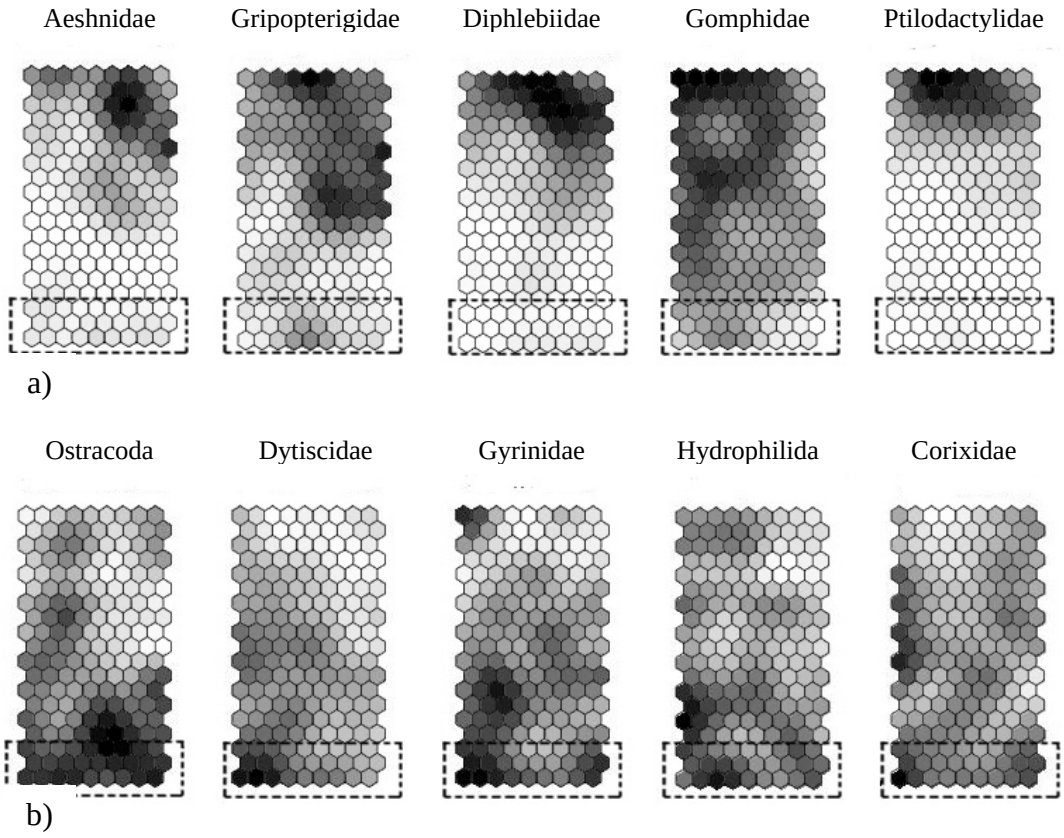
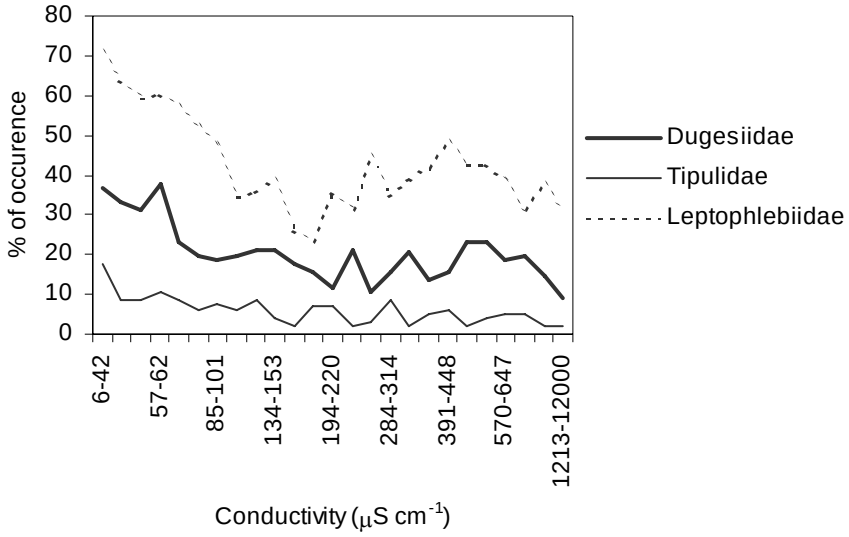


Figure 6.6. Selected SOM component planes for macroinvertebrate taxa, a) collected in mostly low salinity conditions (sensitive (s)), b) macroinvertebrate taxa collected in high salinities conditions (very tolerant (vt), darker shades indicate more frequent presence, broken outline indicates high conductivity corresponding to subsaline and saline categories by Williams (1967).

Analysis of occurrence patterns of macroinvertebrate taxa in relation to conductivity using frequency plots

Figure 6.7 shows selected plots for several taxa with typical decreasing frequency trend (potentially sensitive) and with typical increasing frequency trend (potentially very tolerant). Frequency trends for all taxa are indicated in Table 6.4 (edge habitat) and Table 6.5 (riffle habitat).

(a)



(b)

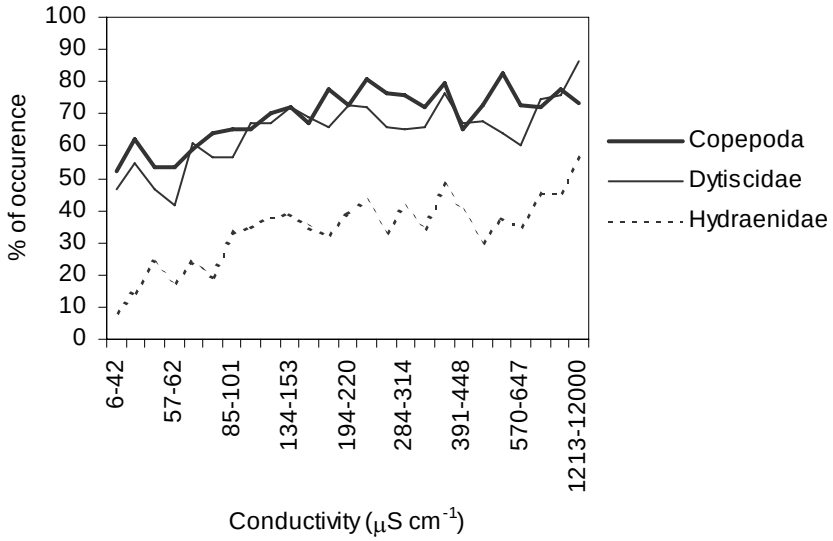


Figure 6.7. Plots of (a) 'decreasing' and (b) 'increasing' trends of selected stream macroinvertebrates along the conductivity gradient, edge habitat.

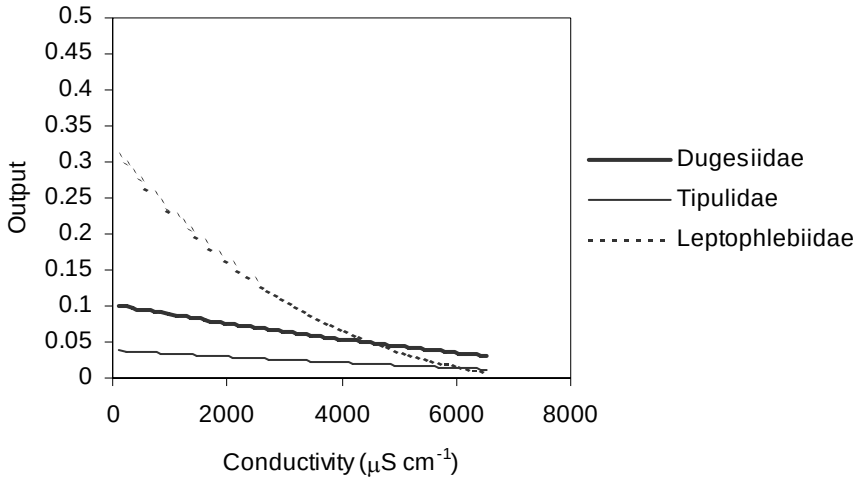
Sensitivity analysis with MLP

Average percentage of correct predictions of the 117 MLP models (tested on 10% of all data not used for training) was 72.4% ranging from 50 to 89.8%.

Figure 6.8 shows typical 'decreasing' and 'increasing' trends in the probability of taxa occurrence along the conductivity gradient. In few cases the curves were almost flat with slight tilt, those trends were described as 'slightly decreasing' or 'slightly increasing'. Trends for each taxon are documented in Table 6.4 (edge habitat) and Table 6.5 (riffle habitat). We suggest that taxa showing 'decreasing' trend can be considered as sensitive as it shows preference towards lower conductivities. Taxa with

'increasing' trends have high tolerance to salinity as probability of their occurrence increases with increase in conductivity. Taxa with flat or unimodal curves are considered as opportunistic taxa with wide tolerance range or preferences around medium conductivity range.

a)



b)

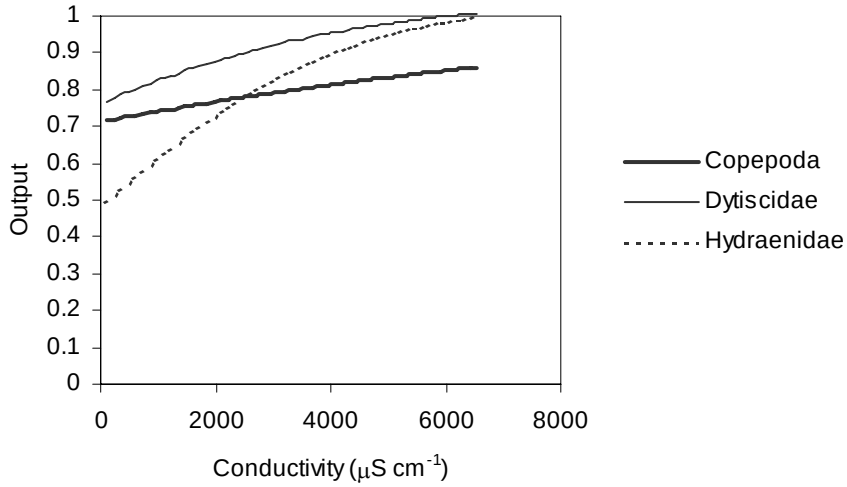


Figure 6.8. Typical sensitivity plots resulting from the sensitivity analysis of MLP for selected taxa, edge habitat. a) 'Decreasing' and, b) 'increasing' trends in the probability of occurrence of macroinvertebrate taxa along the conductivity gradient.

Salinity Sensitivity Score and Salinity Index

Because the analysis conducted was of semi-qualitative nature, in many cases it was difficult to be sure about which group the taxon should be assigned. In many cases the output of all methods clearly indicated that taxon is whether very tolerant of sensitive. However, when considering taxa with salinity preference not clearly expressed, output of the methods applied was different in many cases. In general, output of all three methods applied was highly similar in 61% of taxa, in some cases decisions taken about assigning a label were arbitrary and open to reconsideration if more data is available.

Twelve taxa were labelled as sensitive in the both datasets, 16 and 21 as very tolerant in edge and riffle respectively (see Tables 6.4 and 6.5). The generally tolerant taxa comprised the largest group in the both habitats, 29 in edge and 27 in riffle.

As conductivity increases, sensitive taxa are being replaced by very tolerant (Fig. 6.9). This trend is obvious in both habitats but appears to be more prominent in riffles with higher proportions of sensitive taxa present under low conductivities. In riffles, relatively to the low conductivity category (22-99 µS cm^{-1}) mean percent of sensitive taxa decreased from 33 to 16.7 and very tolerant increased from 9.4 to 32 in sites with EC between 800 and 1500 µS cm^{-1}.

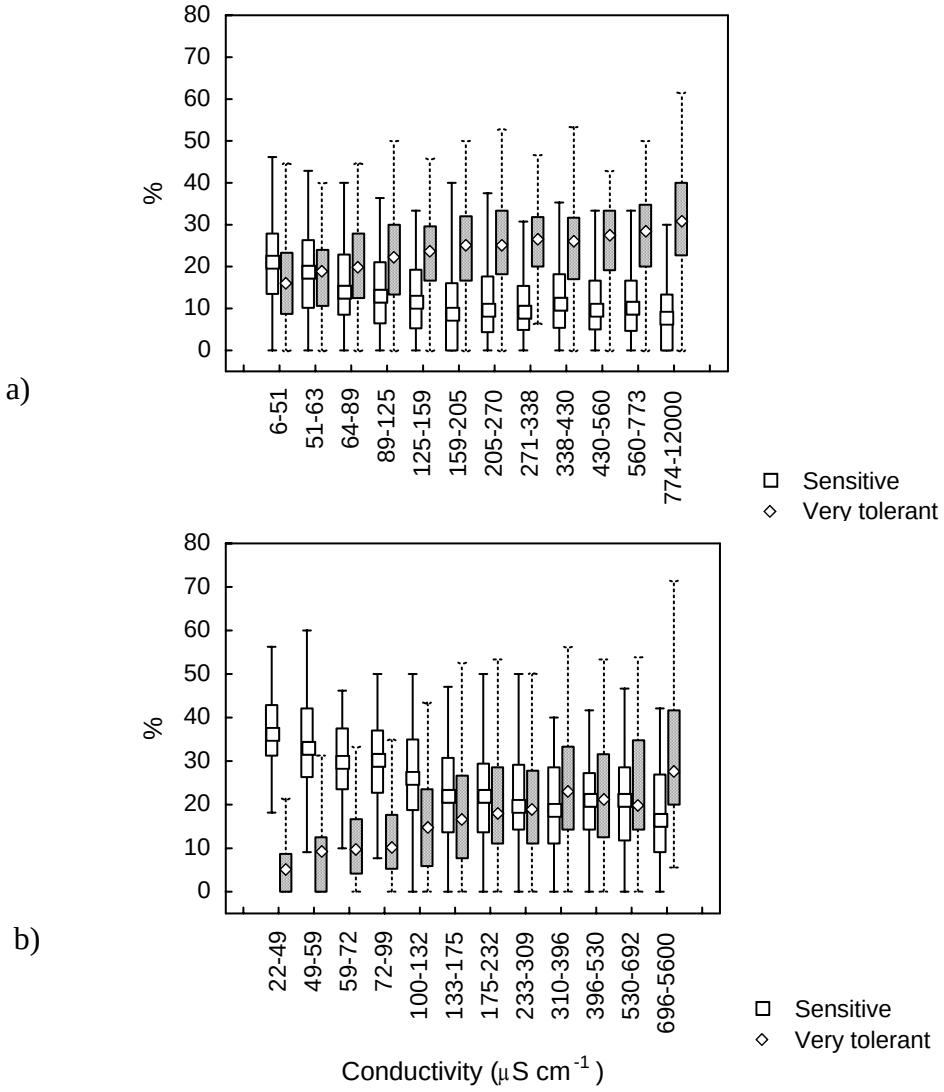
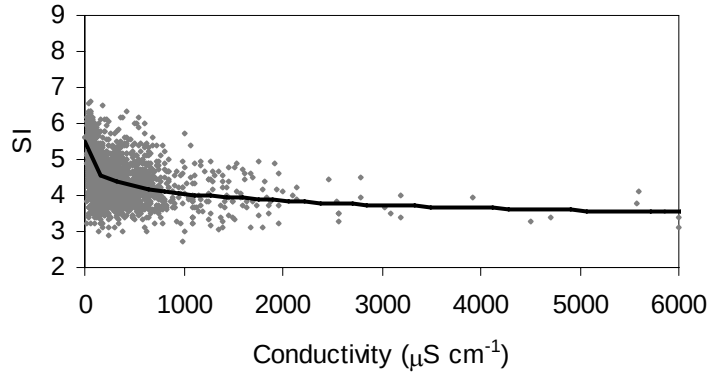


Figure 6.9. Percentage of sensitive and very tolerant taxa in 12 equal sized bins along the gradient of increasing conductivity, a) edge habitat, b) riffle habitat. Median values with boxes corresponding to 80th and 20th percentiles and horizontal bars to maximum and minimum.

SI calculated using all three tolerance groups ranged from 8 to 2.14 in riffle habitat and from 6.8 to 1 in the edge. Most of the sites with high SI had low conductivity, for example, out of 710 samples (edge habitat) with SI values between 5 and 7.3 only 9 sites (1.2%) had conductivity higher than 1000 µS cm^{-1}. However, sites with low SI not always were characterised by high conductivities. Out of 498 samples with SI between 2 and 4, 411 samples had conductivity less than 800 µS cm^{-1}. More than half of these samples (216) had elevated concentration of nutrients (total nitrogen > 0.75 mg L^{-1} or total phosphorus > 0.1 mg L^{-1}) or turbidity higher then 50 NTU, which might explain the presence of mainly opportunistic taxa in these sites. However, we could not explain low SI (between 2 and 4) at 195 samples. It also has become apparent that many samples with SI too high or too low disproportionately to the conductivity level had also relatively low number of taxa present (less than 15), which might simply be not representative enough to characterize the community.

In order to quantify the relationship between SI and EC a logarithmic trend has been fitted between these two variable. Prior to this step all samples with the number of taxa less than 15 were deleted to improve the fit of the model and reduce the amount of noise. SI calculated according to the fitted model (Fig. 6.10) was significantly ($p<0.05$) correlated with actual SI ($R = 0.45$ and 0.64, edge and riffle accordingly) and

a)



b)

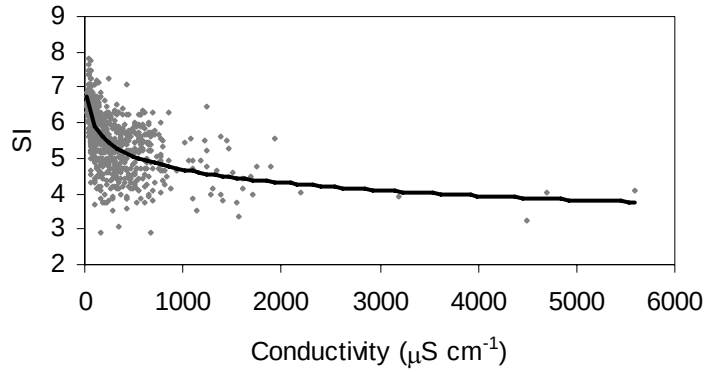
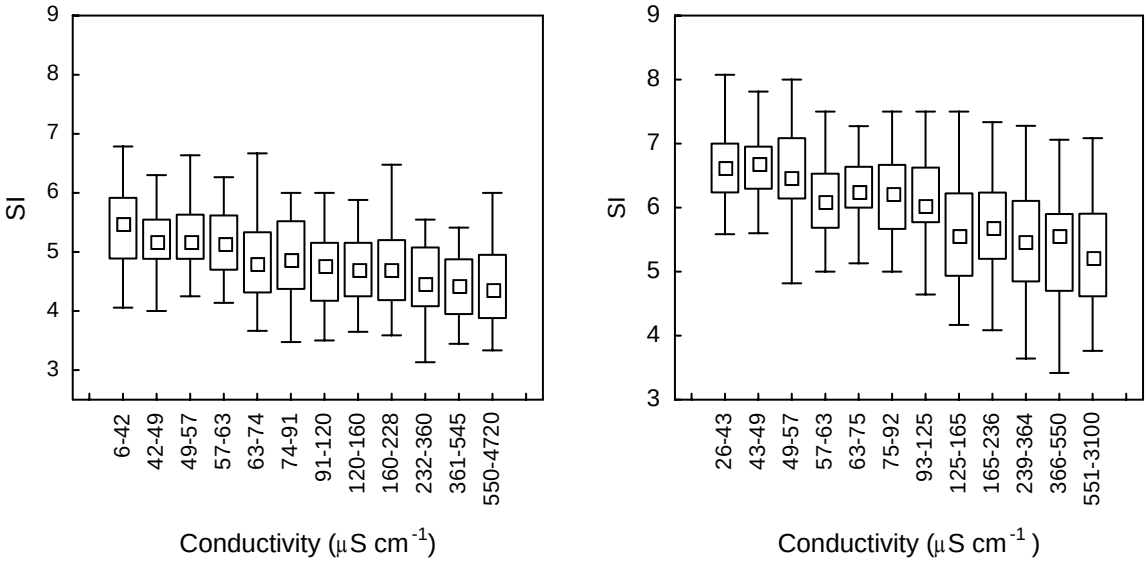


Figure 6.10. Scatterplots of SI versus conductivity with fitted logarithmic trends, a) edge, y = -0.29Ln(x) + 6.03, b) riffle, y = -0.53Ln(x) + 8.36.

with EC (R = -0.71 and -0.76 for edge and riffle). It is difficult to determine any threshold point as the change in community structure is gradual, however it is evident that the most quick and dramatic shift between groups with different salt tolerance occurs up to approximately 800-1000 µS cm^{-1}, after that communities continue changing towards the dominance of salt tolerant taxa but at slower rate. For example, in riffle habitat, predicted SI was the highest (6.71) at 22 µS cm^{-1} and decreased by two units (4.71) at 1000µS cm^{-1}, further decrease was less pronounced with the lowest value of 3.76 at 5600 µS cm^{-1}.

Subsets with good water quality sites used to validate that SI reflects changes due to EC and not other water quality factors included 745 samples from edge habitat and 576 samples from riffle habitat. It needs to be mentioned that 80% of samples from edge habitat with conductivity higher than 800 µS cm^{-1} had also elevated nutrients as total nitrogen higher than 0.375 mg L^{-1}, and 40% of those samples had total nitrogen higher than 0.75 mg L^{-1} (69% and 22% for the riffle habitat respectively). In other words we had to exclude a number of sites with high conductivities when using only sites with good water quality. It is apparent that SI still decreases with increase in EC even though that effect of other water quality factors has been ruled out, this trend is more pronounced in riffle habitat (Figure 6.11).



a) b)

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.

Figure 6.12 shows scatter plots of Salinity Index versus major possible natural gradients and flow expressed as maximum water velocity (m/s). Salinity Index was not highly correlated with any of the variables considered (the highest R = 0.3 was with maximal water velocity and mean annual rainfall).

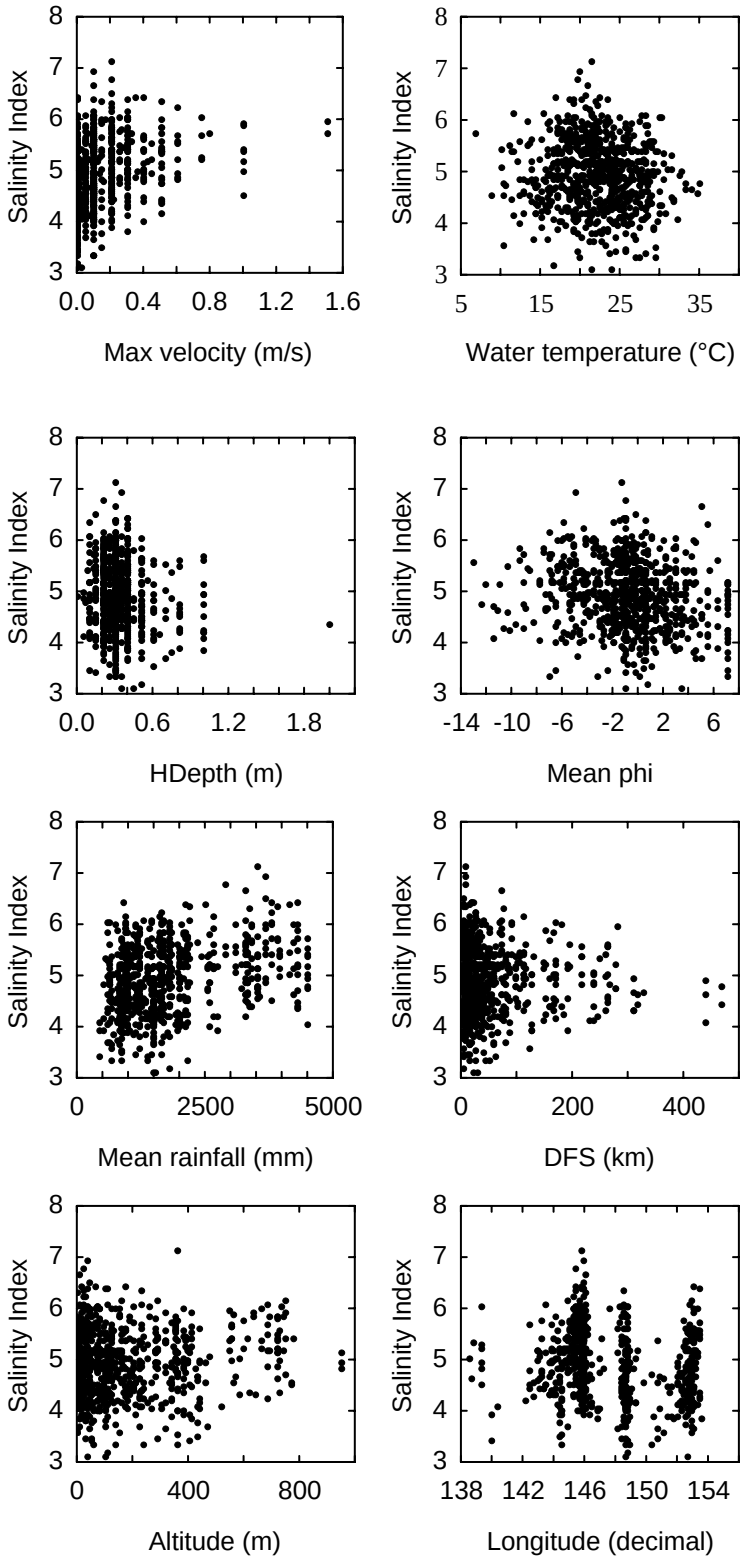


Figure 6.12. Scatterplots of SI versus flow (maximum water velocity), water temperature, habitat depth (HDepth), mean phi, mean annual rainfall, distance from source (DFS), altitude and longitude.

Partial CCA

When all significant variables were used as an input, four axes of CCA accounted for 6.7% in (riffle) and 5.6% (edge) of variability in taxa data and 62.8 % (riffle) and 66.5% (edge) of taxa-environmental relation. When temporal and natural variability was partialled out, four axes of new CCA using only water quality variables accounted for 2% (riffle) and 1.3% (edge) of variability in taxa data and 80.9% (riffle) and 77.5% (edge) of taxa-environmental relation. Figure 6.13 clearly shows that variables conductivity, water temperature, total phosphorus and pH affect macroinvertebrate communities in a similar way. It appears that total phosphorus and EC are correlated stronger in the riffle subset. Alkalinity is not shown in the riffle biplot as it was found insignificant in the previous stages of the analysis. Starting point of each arrow represents mean value of the variable. Majority of the taxa labelled as very tolerant (T) are located along the gradient of increasing, higher than average EC and the taxa labelled as sensitive (S) are located mostly on the opposite side. This confirms that most of the taxa were labelled correctly, with slightly better accuracy in the case of riffle habitat (keeping in mind that CCA is a linear method). However, it is also evident that some confounding by the other water quality variables is possible.

Discussion and conclusion

The aim of this study was to investigate changes in macroinvertebrate communities associated with the changes in the conductivity level in streams and rivers using a variety of methods as simple frequency distribution, SOM component planes and sensitivity analysis with MLP. In many cases (61% of taxa) outputs of all three methods were largely agreeable with each other, however, because of the semi-quantitative nature of these methods in some cases it was difficult to draw a line and make a decision about some taxa. For example, all three methods indicated that Copepoda is highly tolerant, it also was found at the highest conductivity in the dataset (12000 µS cm^{-1}) and has relatively high mean conductivity (383 µS cm^{-1}). In this case it was easy to label this taxon as very tolerant. Helicopsychidae was assessed as sensitive by all three methods and also was found in relatively low conductivities (232 µS cm^{-1}). It was also easy to label this taxon as sensitive. Our assessment of these two taxa also agrees with previous findings (Hart et al., 1991).

In other cases, as for Nepidae, the shape of frequency graph was unimodal, indicating preference of the medium salinity, according to component plane taxa is sensitive and according to sensitivity curve taxa is very tolerant. The decision for this taxon was taken on the basis of sensitivity curve and the fact that it was found in relatively high conductivity (5570 µS cm^{-1}) and also had high mean conductivity (359 µS cm^{-1}). According to previous findings (Hart et al., 1991; Kefford et al., 2003) Hemiptera are generally quite tolerant and this taxa was labeled as very tolerant.

In the majority of the cases our assessment of taxa salt sensitivity was in agreement with previous findings, however in some cases there were disagreements. For example, Kefford et al. (2003) suggested that Australian freshwater mollusks appear to be salinity sensitive but in our case all three methods indicated Planorbidae to be

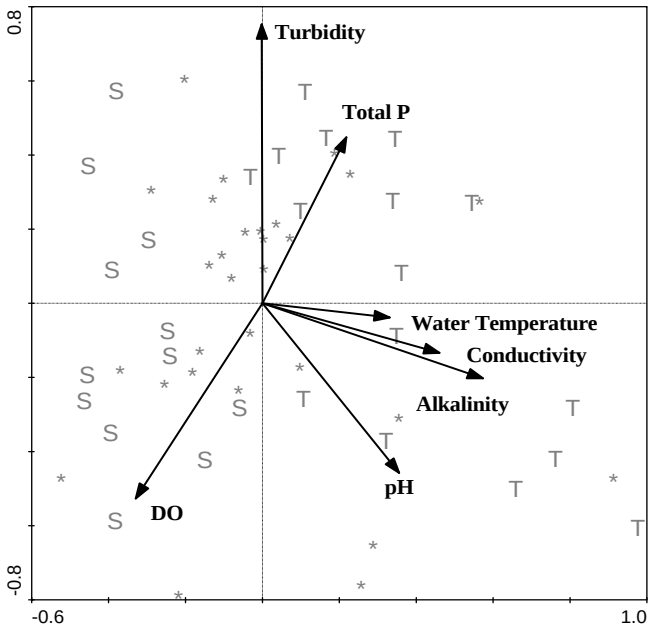
highly tolerant. It also was found in high salinities (11730 µS cm^{-1}) and had high mean conductivity as well (512 µS cm^{-1}). One possible explanation for this might be that some mollusks have coping mechanisms allowing them to deal with relatively high salinities. We also need to keep in mind that taxa indicated as tolerant here might not be so beyond the salinity range considered here.

In general, it has been shown that steady substitute of salt sensitive taxa by opportunistic and salt tolerant ones does occur even with relatively slight increase in EC. The data analysis was intended to rule out other possible causes for the observed changes as natural variability, flow or other water quality parameters. Even though we have a reasonable confidence that described changes in macroinvertebrate communities caused by changes in conductivity as primary stressor, it is impossible to rule out all the possibilities. Partial CCA showed that increase in conductivity can also be associated with increase in water temperature and nutrients level, combination of these factors might have a compounding effect on the macroinvertebrate communities and currently poorly understood. This concern has also been expressed by other authors (James et al., 2003) and more research in this direction is needed.

We have not observed any reduction in taxonomic richness within the limits of available data. Several other authors have found no change in richness along a salinity gradient although the community composition differed along that gradient (Williams et al., 1990; Metzeling, 1993; Kefford, 1998). Therefore, taxonomic richness might not be a sensitive enough indicator to detect the effect of secondary salinisation as sensitive species are simply replaced by more tolerant ones with overall number of species remaining the same.

According to our results, the most dramatic shift between groups of different salinity tolerance occur at conductivity values of 800-1000 µS cm^{-1} and this shift seems to be more pronounced in riffle habitat. This threshold value is lower than the generally accepted value of 1500 µS cm^{-1}, above which freshwater ecosystems are likely to experience salinity related stress (Hart et al., 1991). One possible explanation for this difference is that we used the state-wide dataset containing many samples from streams in very good condition and with very low conductivities (in range of 6-40 µS cm^{-1}), this is particularly relevant in the case of riffle dataset. In other words changes in stream macroinvertebrate communities affected by secondary salinisation may be more obvious in comparison with stream systems in near pristine condition, while we might not observe any drastic changes when comparing systems already impacted to some degree by variety of anthropogenic stressors and dominated by opportunistic taxa with likewise. The possible warning resulting from this study is that comparing affected ecosystems with already disturbed ecosystems might lead to erroneous conclusion that the Australian freshwater fauna is highly tolerant to changes in salinity. Although the changes in freshwater biota might be initially subtle, this could lead to further instability of the ecosystem structure and function and compounding effects of other potential stressors.

a)



b)

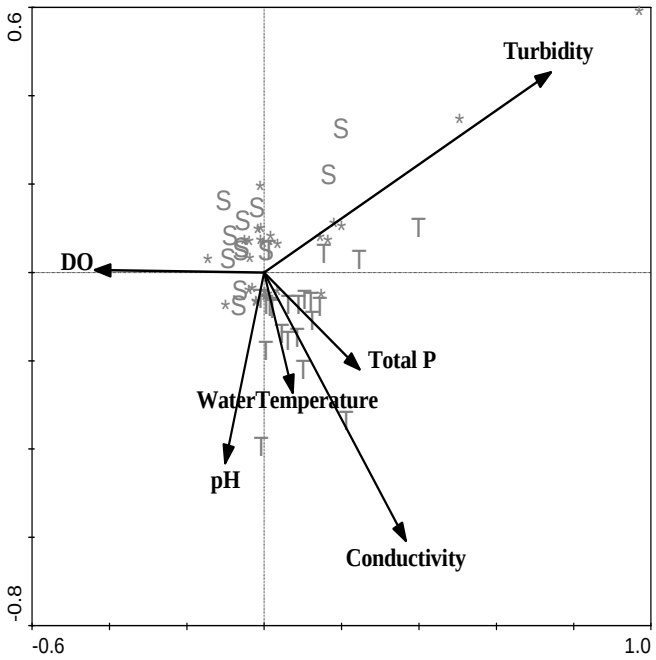


Figure 6.13. CCA biplots showing effect of water quality variables after the effect of natural and temporal variability was partialled out a) edge, b) riffle. S - sensitive taxa, T - very tolerant taxa, * - generally tolerant taxa, DO - dissolved oxygen (mg L^{-1}), Total P – total phosphorus (mg L^{-1}).

The Salinity Index suggested as a measurement of changes in community structure was decreasing along the gradient of increasing conductivity. This pattern remained the same when SI and conductivity were plotted for the data subset containing stream

sites with otherwise good water quality, which confirms that SI indeed most likely reflects changes in community structure associated with the changes in conductivity. However, there were sites with low SI and low conductivity. One possible explanation for strange SI values is generally low number of taxa in the sample (lower than 15). SI appears to be less reliable when calculated for communities with low number of taxa. Another probably explanation is that sites with unexpectedly low SI might be under stress unaccounted for in the scope of this study (impoverished habitats, presents of toxicants, etc.). There is also a possibility that some low score taxa were assigned incorrectly. This may be the case as the distribution of conductivity values in the dataset was highly skewed towards lower values, with the majority of sites falling into the fresh water category. This means that some taxa graded as tolerant on the basis of this data might not be so if more data from the higher salinity range was available. For example, some molluscs (Thiaridae, Lymnaeidae, Planorbidae) and crustacean Parasticidae generally considered salinity sensitive (Hart et al., 1991) were scored as very tolerant in this study, which might only be true for the conductivity range considered in this study. However, we do not know which particular species were in those families, which makes it difficult to compare our results with the other studies considering sensitivities of species.

The other possible factors causing discrepancies in the Salinity Index versus conductivity relationship could be: natural patterns in taxa distribution (we only used taxa with state-wide distribution, but we do not exclude that natural variability might be having an underlying effect), lag effect of previous exposure (stream had an inflow of fresh water prior the sampling event, but fauna was still recovering from a previous salinity level) and the effect of ionic composition. The last factor was not taken into consideration in this study. However, it has a potential to confound the results. McNeil & Cox. (in press) defined a series of water types in QLD based on proportions of cations and anions. The proportion of sodium chloride decreases westward from the coast. The streams from inland catchments are often high in calcium bicarbonate, sulphate and other components. Different ionic composition of water with otherwise equal conductivities might affect freshwater biota in different ways. Bayly (1969) suggested that the monovalent ions (Na^+ and K^+) are more toxic than divalent ones (Ca^{2+}). This means that higher proportions of sensitive taxa could be found in calcium bicarbonate dominated water than in sodium chloride dominated water under equal conductivities.

Even though we observed similar changes in communities from both riffle and edge habitat, these changes are more pronounced in riffles. As it was mentioned before edge habitat data included many sites from western regions, including lots of streams with intermittent flow. It is possible that macroinvertebrates inhabiting these sites are better adapted to natural changes in salinity than those from riffles, as riffle habitat by definition requires some flow all the time. This difference deserves further research, which could give us more understanding on the mechanism of adaptations and natural resilience of freshwater biota.

Our findings provide an interesting insight into broad scale salinity sensitivities of QLD stream macroivertebrates. Even though overall taxonomic richness does not change, structural changes in macroinvertebrate communities do occur with even slight increase in conductivity. However, since the analyses were based on coarse taxonomic resolution (mainly at the family level) care should be taken when applying

proposed sensitivity scores to the assessment of the risk posed by salinity. Increased taxonomic and geographical resolution combined with more data sampled in the higher conductivity streams (brackish-saline categories) and using abundance data instead of presence-absence are likely to improve the accuracy and precision of the Salinity Index.

Table 6.4. Taxa specific conductivity ranges, trends shown by frequency plots and sensitivity analysis with ANN and Salinity Sensitivity Score, edge habitat**.**

Taxa	Minimum conductivity (mS cm-1)	Maximum conductivity (mS cm-1)	Mean conductivity (mS cm-1)	Mean frequency of occurrence along salinity gradient (25 bins)	Probability of occurrence with increase in salinity (ANN models)	Sensitivity by SOM	Salinity Sensitivity Score (1-very tolerant, 5-tolerant, 10 - sensitive)
Acarina	6	11730	319.98	decr	slightly decr	s	5
Aeshnidae	31	4500	426.68	slightly incr	unimod	t	5
Ancylidae	30	2560	453.59	incr	slightly decr	t	5
Atyidae	22	12000	369.38	incr	slightly decr	t	5
Baetidae	6	11730	344.33	no vis.tr	decr	t	5
Caenidae	22	11730	332.39	no vis.tr	decr	t	5
Calamoceratidae	6	5570	331.27	no vis.tr	decr	t	5
Ceratopogonidae	20	11730	344.55	no vis.tr	unimod	t	5
Cladocera	25	12000	385.05	incr	unimod	t	5
Coenagrionidae	6	12000	383.71	incr	incr	vt	1
Copepoda	20	12000	377.84	incr	incr	vt	1
Corbiculidae	45	2150	449.46	incr	decr	t	5
Corduliidae	23	2980	291.92	decr	decr	s	10
Corixidae	20	11730	365.64	incr	slightly decr	vt	5
Culicidae	20	11730	423.85	incr	incr	t	1
Dugesiidae	6	2460	264.91	decr	decr	s	10
Dytiscidae	6	12000	396.93	incr	incr	vt	1
Ecnomidae	23	11730	331.92	no vis.tr	slightly incr	t	5
Elmidae	22	3100	236.68	decr	decr	s	10
Gerridae	25	5600	315.68	slightly decr	decr	t	5
Gomphidae	6	12000	296.9	decr	decr	s	10
Gyrinidae	6	5600	316.51	no vis.tr	incr	t	5
Helicopsychidae	22	1387	232.26	decr	decr	s	10
Hydraenidae	20	11730	455.97	incr	incr	vt	1
Hydrometridae	20	5990	443.82	incr	incr	vt	1
Hydrophilidae	6	6010	360.73	no vis.tr	slightly incr	t	5
Hydropsychidae	6	2780	252.41	decr	decr	t	10
Hydroptilidae	28	5990	292.47	decr	decr	t	10
Isostictidae	28	5600	377.17	no vis.tr	incr	t	5
Leptoceridae	6	11730	340.68	slightly decr	decr	t	5
Leptophlebiidae	6	3910	289.84	decr	decr	t	10
Libellulidae	6	11730	328.34	decr	incr	t	5

Abbreviations: ‘incr’ – increasing, ‘decr’ – decreasing, ‘no. vis tr’ – no visible trend, ‘unimod’ – unimodal, ‘s’ – sensitive, ‘t’ – tolerant, ‘vt’ – very tolerant

Continuation of Table 6.4. Taxa specific conductivity ranges, trends shown by frequency plots and sensitivity analysis with ANN and Salinity Sensitivity Score, edge habitat.

Taxa	Minimum conductivity (mS cm-1)	Maximum conductivity (mS cm-1)	Mean conductivity (mS cm-1)	Mean frequency of occurrence along salinity gradient (25 bins)	Probability of occurrence with increase in salinity (ANN models)	Sensitivity by SOM	Salinity Sensitivity Score (1-very tolerant, 5-tolerant, 10 - sensitive)
Lymnaeidae	39	6010	558.99	incr	no vis.tr	vt	1
Mesoveliidae	31	6010	402.37	incr	incr	vt	1
Naucoridae	28	5990	492.26	incr	incr	t	1
Nepidae	20	5570	359.48	unimodal	incr	s	1
Notonectidae	30	6010	413.87	incr	slightly decr	t	5
Oligochaeta	20	11730	378.67	incr	slightly decr	t	5
Orthocladiinae	6	11730	330.63	no vis.tr	decr	t	5
Ostracoda	6	6010	368.19	incr	incr	vt	1
Palaemonidae	6	12000	321.47	decr	decr	t	5
Parasticidae	33	12000	495.54	incr	incr	vt	1
Planorbidae	37	11730	512.04	incr	incr	vt	1
Pleidae	20	11730	398.44	incr	no vis.tr	t	5
Protoneuridae	6	12000	375.42	no vis.tr	no vis.tr	vt	5
Psephenidae	22	5600	393.76	no vis.tr	decr	t	5
Pyralidae	22	3200	264.54	decr	decr	s	10
Scirtidae	23	5600	367.29	no vis.tr	incr	t	1
Simuliidae	6	2460	293.57	no vis.tr	decr	t	5
Staphylinidae	29	5990	433.01	no vis.tr	incr	t	1
Stratiomyidae	52	5570	569.25	incr	decr	vt	5
Tabanidae	42	5990	420.93	incr	incr	t	1
Tanypodinae	6	11730	352.31	no vis.tr	slightly decr	t	5
Temnocephalidea	27	3040	280.85	no vis.tr	decr	s	10
Thiaridae	30	12000	449.26	incr	incr	vt	1
Tipulidae	6	2980	228.14	decr	decr	s	10
Veliidae	20	8700	354.81	no vis.tr	slightly decr	t	5

Table 6.5. Taxa specific conductivity ranges, trends shown by frequency plots and sensitivity analysis with ANN and Salinity Sensitivity Score, riffle habitat.

Taxa	Minimum conductivity (mS cm-1)	Maximum conductivity (mS cm-1)	Mean conductivity (mS cm-1)	Mean frequency of occurrence along salinity gradient (25 bins)	Probability of occurrence with increase in salinity (ANN models)	Sensitivity by SOM	Salinity Sensitivity Score (1-very tolerant, 5-tolerant, 10 - sensitive)
Acarina	26	4500	240.74	decr	slightly decr	s	5
Aeshnidae	22	1574	141.57	decr	slightly decr	s	10
Ancylidae	51	1389	357.04	incr	incr	s	1
Atyidae	31	5600	378.71	incr	slightly incr	vt	1
Baetidae	22	4500	273.2	decr	decr	t	5
Caenidae	26	5600	325.14	incr	incr	vt	5
Calamoceratidae	26	3200	292.44	decr	slightly incr	s	5
Ceratopogonidae	22	4500	260.86	decr	decr	s	5
Cladocera	45	1515	321.2	incr	incr	t	5
Coenagrionidae	30	4500	479.42	incr	incr	vt	1
Copepoda	37	4500	421.14	incr	incr	vt	1
Corbiculidae	47	3200	472.16	incr	incr	vt	1
Corduliidae	22	5600	319.29	no vis.tr	incr	t	5
Corixidae	38	3200	383.92	incr	incr	vt	1
Culicidae	60	4500	564.81	incr	incr	vt	1
Dolichopodidae	40.4	5600	358.51	incr	slightly incr	t	1
Dugesiidae	22	3200	248.96	decr	decr	s	10
Dytiscidae	44	5600	489.3	incr	incr	vt	1
Ecnomidae	27	5600	376.22	incr	slightly incr	vt	1
Elmidae	22	5600	254.91	decr	decr	s	10
Gerridae	42	4700	418.22	incr	incr	vt	1
Gomphidae	22	5600	250.76	decr	decr	s	10
Gyrinidae	44.4	5600	478.08	incr	incr	vt	1
Helicopsychidae	22	1423	270.52	decr	slightly decr	s	5
Hydraenidae	31.3	4700	503.67	incr	incr	vt	1
Hydrobiosidae	26.6	1696	215.26	decr	decr	s	10
Hydrometridae	175	1574	769	incr	incr	vt	1
Hydrophilidae	22	5600	366.29	incr	incr	vt	1
Hydropsychidae	22	5600	295.14	no vis.tr	slightly decr	t	5
Hydroptilidae	26	4500	295.87	no vis.tr	incr	t	5
Isostictidae	90	636	384.6	no vis.tr	decr	t	5
Leptoceridae	22	5600	300.02	no vis.tr	incr	t	5
Leptophlebiidae	22	5600	266.62	decr	decr	t	10
Libellulidae	27	5600	294.43	no vis.tr	incr	t	5
Lymnaeidae	70	2094	513.88	incr	incr	vt	1
Mesoveliidae	43	1750	283.5	no vis.tr	incr	t	5

Continuation of Table 6.5. Taxa specific conductivity ranges, trends shown by frequency plots and sensitivity analysis with ANN and Salinity Sensitivity Score, riffle habitat.

Taxa	Minimum conductivity (mS cm-1)	Maximum conductivity (mS cm-1)	Mean conductivity (mS cm-1)	Mean frequency of occurrence along salinity gradient (25 bins)	Probability of occurrence with increase in salinity (ANN models)	Sensitivity by SOM	Salinity Sensitivity Score (1-very tolerant, 5-tolerant, 10 - sensitive)
Naucoridae	39	2094	233.45	decr	incr	s	5
Nepidae	108	757	305.14	no vis.tr	incr	s	5
Notonectidae	40	1610	327.14	no vis.tr	incr	s	5
Oligochaeta	26.6	5600	313.54	incr	decr	t	5
Orthocladiinae	26	4500	286.9	no vis.tr	slightly decr	t	5
Ostracoda	38	5600	420.03	incr	incr	vt	1
Palaemonidae	22	4500	262.24	decr	decr	t	5
Parasticidae	39	4500	402.02	incr	decr	t	5
Philopotamidae	26	5600	238.26	decr	decr	s	10
Planorbidae	52	4500	568.18	incr	incr	vt	1
Pleidae	81	968	373.56	incr	incr	t	5
Protoneuridae	40	1450	403.35	incr	incr	t	5
Psephenidae	22	5600	283.46	decr	incr	s	5
Pyralidae	26.6	3100	236.69	decr	decr	s	10
Scirtidae	30	4700	246.34	decr	slightly decr	s	10
Simuliidae	22	4700	304.14	no vis.tr	decr	t	5
Staphylinidae	32	769	228.73	no vis.tr	decr	s	10
Stratiomyidae	56	5600	828.59	incr	incr	vt	1
Tabanidae	26	4700	298.49	incr	slightly incr	t	5
Tanypodinae	22	5600	286.43	no vis.tr	decr	s	5
Temnocephalidea	43	666	246.58	no vis.tr	decr	t	10
Thiaridae	38	5600	483.59	incr	incr	vt	1
Tipulidae	22	1951	218.03	decr	decr	s	10
Veliidae	40	5600	362.51	incr	incr	vt	1

Chapter 7

Scenario analysis based on the dirty-water approach

7.1 Predicting the effect of secondary salinisation on stream macroinvertebrate communities in Central Queensland.

Introduction

Wide spread secondary salinisation caused by the clearance of deep-rooted native vegetation is one of the major threats facing freshwater ecosystems in Australia. A number of streams and wetlands have already been affected by rising salinity leading to significant changes in flora and fauna. Macroinvertebrates in particular appear to be highly salt sensitive (Hart et al., 1991). There have been a number of studies on the effect of salinisation on macroinvertebrate taxa using both laboratory experimentations (Kefford et al., 2003; Kefford et al., 2004), field observations (Bunn and Davies, 1992; Kay et al., 2001; Kefford, 1998; Metzeling, 1993; Williams et al., 1991) and mesocosm experiments (Marshall and Bailey, 2004). The majority of these studies were conducted in southern and western states of Australia and there is not much information available for Queensland streams.

Secondary salinisation is a complex process and affects not only conductivity of stream water but other water quality parameters and stream habitats as well. It can be caused by the degradation of riparian vegetation that in turn provides less shade and increases water temperature. Increased nutrient and sediment loads in streams can also be a consequence of deteriorated riparian vegetation. Ground water flow can also contribute to changes in pH, alkalinity and ionic composition as well as to the enrichment of nutrients such as nitrate (NO3) originating from fertilisers (Brodie et al., 1984). There is currently little understanding about additive, synergistic or antagonistic effect of salinity and nutrients. In a study of the biological effects of saline lake water disposal in the Lough Calvert drainage scheme in Southwest Victoria, Kefford (2000) found that the operation of the scheme changed the community structure and abundance of macroinvertebrates. He noted that increased

salinity corresponded with increased nutrients and suspended solids having a compounding effect on macroinvertebrates communities.

This study aims at testing the predictability of the Salinity Index and percentage of sensitive taxa (PST) defined in previous study (see Chapter 6) using localised datasets from Central Queensland, and to investigate possible changes in SI and PST in response to two scenarios: scenario 1 considering an increase in conductivity and related variables, and scenario 2 considering an increase in conductivity, related variables and nutrients (total nitrogen and total phosphorus). We used two types of ANNs. While modular feedfordward ANN (modified multi-layered perceptron) were used for the prediction, SOM were applied to analyse the results.

The main hypotheses for this study are:

1) Using ANN models it is possible to predict SI and PST based on 'dirty-water' approach.

2) Simulated increase in conductivity affects the structure of macroinvertebrate communities.

3) Combined stressors like conductivity and nutrients affect macroinvertebrate communities in a greated degree than conductivity or nutrients alone.

Data and methods

Fitzroy and Burdekin are the two largest catchments in Central Queensland and were identified as priority catchments by the National Action Plan for Salinity and Water Quality (NAPSWQ).

The data for this study was collected in Central Queensland in spring and autumn from 1994 to 2001 as a part of several surveys conducted by the Department of Natural Resources and Mines (NR&M). The dataset contains 209 samples collected from riffle habit only. In order to separate data into training and simulation subset we overlayed a GIS map with sample sites and salinity hazard maps for Burdekin and Fitzroy catchments provided by NR&M. Samples collected at the sites located in the areas of moderate to high salinity hazard were selected as the simulation set (36 samples), the rest of the data was used for training (Figure 7.1). Samples from the same site collected in the different year or season were treated as separate sites.

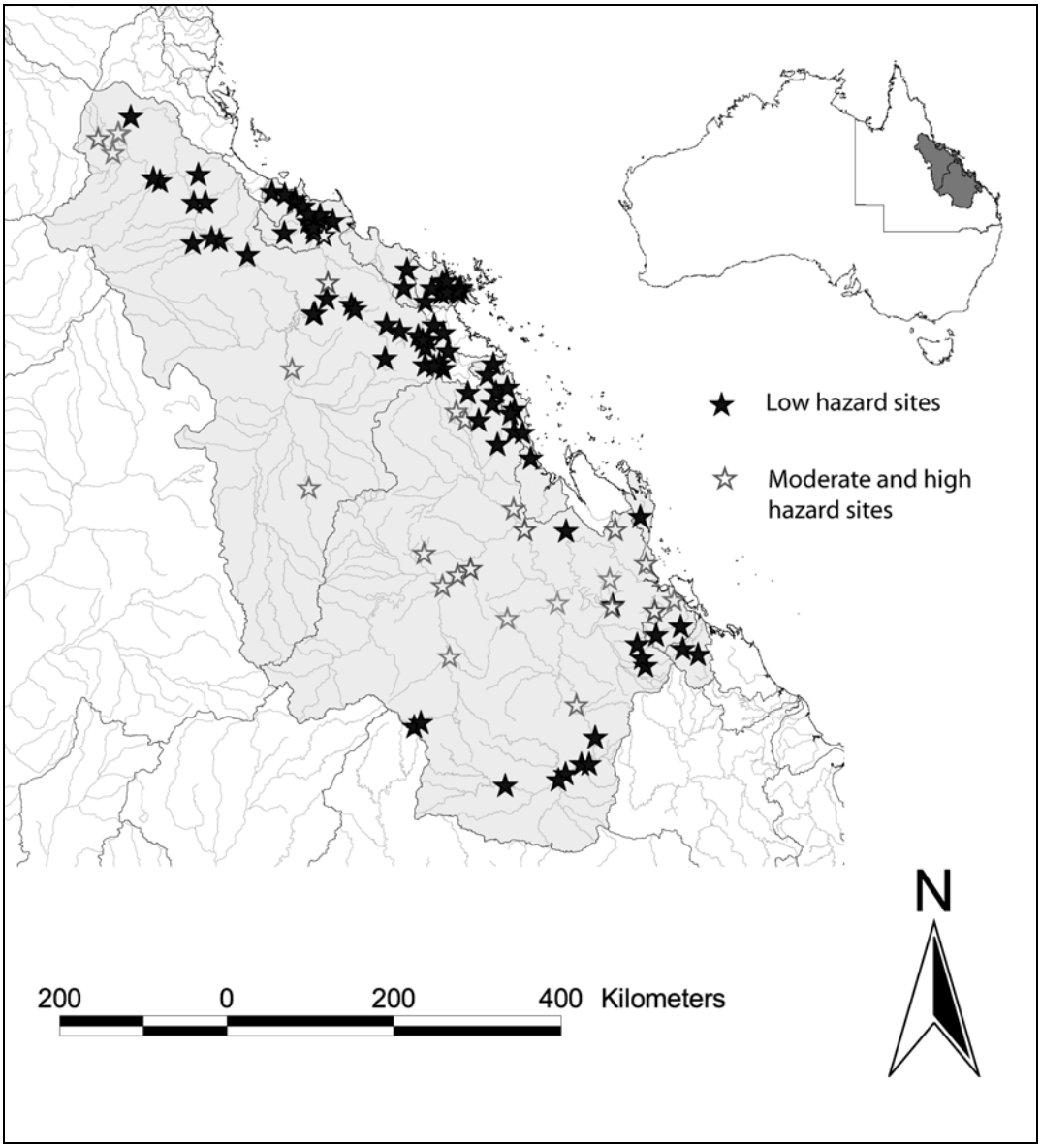


Figure 7.1. Map of the Fitzroy and Burdekin catchments with sites marked accordingly to their location in different salinity hazard zones.

The aim of the modelling process was to predict the effects of likely increases in conductivity by ANNs, which were trained with existing data. The measured conductivity ranged from 40 to 4700 μS cm^{-1}. It was shown by Hart et al. (1991) that freshwater ecosystems can be affected when conductivity reaches 1500 μS cm^{-1}. However, our previous findings (see Chapter 6) demonstrated significant changes in the SI by conductivity values between 800-1000 μS cm^{-1}. Therefore, we assume that conductivities up to 4700 μS cm^{-1} are reasonable for the ANN modelling in order to reveal possible deleterious effects.

By aiming to define a scenario for likely stream salinisation we have taken into consideration possible interactions between conductivity and other factors, such as temperature, nutrient load, turbidity, pH and dissolved oxygen. These interactions

have been analysed by using scatter plots and product moment correlation. In cases where such interactions were detected, linear or logarithmic trends have been fitted and related variables have been calculated in accordance to the conductivity level used for the simulation.

A variety of ANN models with different architecture was built, trained and tested, including Multilayered Perceptron (MLP), Generalized Feed Forward network (a generalization of MLP were connections can jump over one or more layers) and Modular Feedforward neural network. We found that for a given dataset Modular Feedforward neural network showed the best performance when tested on simulation subset, which was not used for training.

Modular feedforward networks are a modification of commonly used Multi-Laeyered Perceptron neural networks (MLPs). These networks process their input using several parallel MLPs, and then recombine the results. This tends to create some structure within the topology, which will foster specialization of functions in each sub-module. In contrast to the MLPs, modular networks do not have full interconnectivity between their layers. Therefore, a smaller number of weights are required for the same size. This tends to speed up training times and reduce the number of required training exemplars (Principe et al. 2000).

We used 29 input variables, including physico-geomorphological features and water quality variables (Table 7.1), 2 hidden layers with 6 neurons in each and tahn transfer function and 2 neurons in the output layer for the SI and the PST.
The model was trained for 1500 iterations using only the training sub-set and validated using a simulation set from the moderate and high hazard salinity areas. To use as much data as possible we did not use cross-validation, instead the model was trained for a various number of iterations and simulated on both training and simulation sets. The models showing big discrepancies between the accuracy of prediction for training and validation sets were discarded as overtrained.

Table 7.1. Input variables used for training of the predictive neural network.

Variable (units)	
Season (categorical)	Slope (km/m)
Habitat depth (m)	Distance from source (km)
Maximal current velocity (m/s)	Mean wet season monthly rainfall (mm)
Bedrock (%)	Mean dry season monthly rainfall (mm)
Boulder (%)	Mean annual rainfall (mm)
Cobble (%)	Conductivity (µs/cm)
Pebble (%)	Water temp (°C)
Gravel (%)	Dissolved oxygen (mg/L)
Sand (%)	pH
Silt/clay (%)	Alkalinity (mg/L CaCO3)
Mean phi	Turbidity (NTU)
Latitude (decimal)	Total nitrogen (mg/l as n)
Longitude (decimal)	Total phosphorus (mg/l as p)
Altitude (m)	0-8. substrate categories
Stream order (categorical)	

The accepted model was simulated five times. First the model was simulated (Simulation 1) using actual data, then four more times with conductivity values increased in increments of 1000 μS cm^{-1} and related variables calculated in relation to the new conductivity values. In other words, conductivity was increased by 1000 μS cm^{-1} for simulation 2, by 2000 μS cm^{-1} for simulation 3 and so on. At the sites where conductivity was initially high, increases of conductivity by 3000 and 4000 μS cm^{-1} resulted in the exceedance of maximal conductivity value in the model's expertise (4700 μS cm^{-1}) so these values were capped at the 4700 μS cm^{-1}. The same capping by the maximal values in the dataset was performed for some sites with the high concentration of nutrients in case of Scenario 2.

Two scenarios were defined: Scenario 1 with only increases in conductivity and directly related variables. Scenario 2 – with increases in conductivity, related variables and nutrients (total nitrogen and total phosphorus). The simulation dataset prepared for scenario 1 was used for scenario 2 plus total nitrogen was increased by 1 mg L^{-1} for each conductivity increment (1000 μS cm^{-1}), total phosphorus was calculated from an equation describing the relationship between total nitrogen and total phosphorus for the given area. In order to compare the combined effect of conductivity and nutrients and nutrients only we simulated the model once using only increase in nutrients (+ 4mg L^{-1} of total nitrogen and the calculated total phosphorus) keeping conductivity and related variables as actual values.

Results

Defining relationships between water quality variables

Product moment correlations between all water quality variables are shown in Table 7.2. The highest correlation was between conductivity and alkalinity (0.53) and total nitrogen and total phosphorus (0.46).

Table 7.2. Product-Moment correlations between water quality variables in Central Queensland.

	Conductivity	Water temperature	DO	pH	Alkalinity	Turbidity	Total N	Total P
Conductivity (μS cm^{-1})	1.00	-.01	-.03	.17*	.53*	-.15*	.09	-.02
Water temperature (°C)		1.00	-.11	.14*	.03	-.05	.17*	.03
DO (mg/L)			1.00	.24*	-.05	-.04	.05	-.01
pH				1.00	.36*	-.02	.02	-.06
Alkalinity (mg/L CaCO3)					1.00	-.14*	-.00	-.06
Turbidity (ntu)						1.00	.19*	.13
Total N (mg/l as N)							1.00	.46*
Total P (mg/l as P)								1.00

*significant ($p < 0.05$)

Given that conductivity is the sum of all the ions present in the solution, higher concentration of Ca and CO_3 ions associated with increase in alkalinity will result in increased conductivity as well. Similarly for pH, increase in either Hydroxyl or Hydrogen ions will contribute to the increase in conductivity.

A scatter plot for conductivity and alkalinity with fitted logarithmic trendline ($y = 88.602\ln(x) - 386.17$, $R^2 = 0.57$) is shown at Figure 7.2.

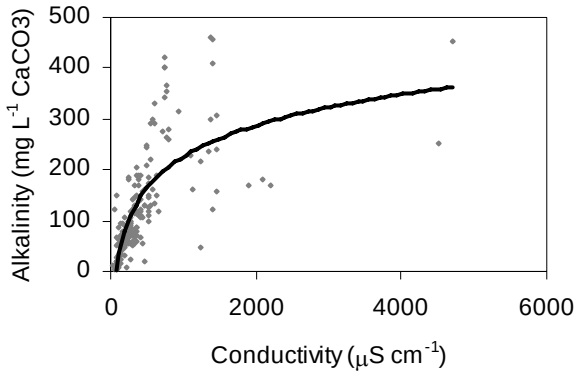


Figure 7.2. Scatterplot of alkalinity versus conductivity with fitted trendline.

We used the above-mentioned trend to calculate changes in alkalinity with simulated increases in conductivity. A similar relationship was observed between pH and conductivity, $r = 0.17$, and a scatter plot with fitted logarithmic trend ($y = 0.2964\ln(x) + 6.0412$, $R^2 = 0.15$), is shown at Figure 7.3.

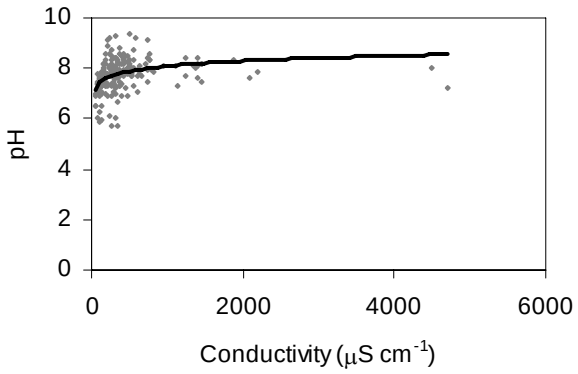


Figure 7.3. Scatter plot of pH versus conductivity with fitted trendline.

Turbidity was negatively correlated with conductivity, however, we could not fit any statistically sound trend to the scatter plot (Figure 7.4.) nor predict turbidity using ANN model. Generally, turbidity is the highest at conductivities between 100 µS cm^{-1} and 500 µS cm^{-1}, but it is almost never high when conductivity is higher than 1000 µS cm^{-1}. This might be explained by the effect of coagulation and settling of suspended particles with a consequent clarification of the water column. Oliver et al. (1999) examined the effect of saline groundwater intrusion on water quality in Darling river (NSW, Australia) showing that increases in water column conductivity under low flow conditions caused major decreases in the turbidity of surface water. A

statistically significant inverse correlation between conductivity and turbidity was also observed in the Klein Modder and Modder Rivers (South Africa), where forty-six per cent of the variation in conductivity was associated with the variation in turbidity (Koning and Roos, 1999).

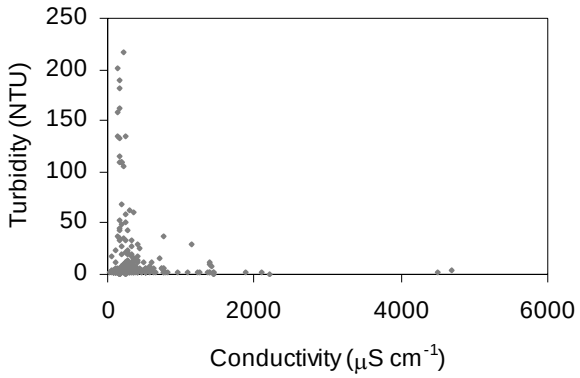


Figure 7.4. Scatter plot of turbidity versus conductivity.

The possible effects of an interaction between turbidity and conductivity are highly dependent on local conditions (such as particle size and current velocity) and may be very complex. Despite the complexities associated with this phenomenon, it was necessary to simplify it for the purpose of this study. As turbidity was observed to be below 30 NTU for all samples having a conductivity less than 1000 μS cm^{-1} (the value accepted as in increment for the simulation of increase in conductivity) 30 NTU was used as a maximum turbidity value for all samples having conductivity > 1000 μS cm^{-1}, with all the values below that being kept at their actual level.
It is possible that a rise of groundwater can cause deterioration of riparian vegetation with subsequent effect of more light coming into the stream. This could cause an increase in water temperature, algal growth and other water quality parameters. However, in the content of this study we do not attempt to model interactions of conductivity and water temperature and keep water temperature at the observed values in all simulations.

No correlation between nutrients and conductivity were found to be significant (Table 2), however, there is a significant correlation between nutrients themselves. As we later attempt to model the combined effect of increased conductivity and nutrients values this relationship needs to be taken in consideration for the simulation process. Figure 7.5 shows a scatter plot with fitted linear trend ($y = 0.1251x - 0.01$, $R^2 = 0.22$) for total nitrogen and total phosphorus. For the subsequent simulations (Scenario 2) we used increases in total nitrogen up to 5.3 mg L^{-1} (maximum occurrence in the dataset for Central Queensland) with an increment of 1 mg L^{-1} , and phosphorus values calculated using abovementioned equation.

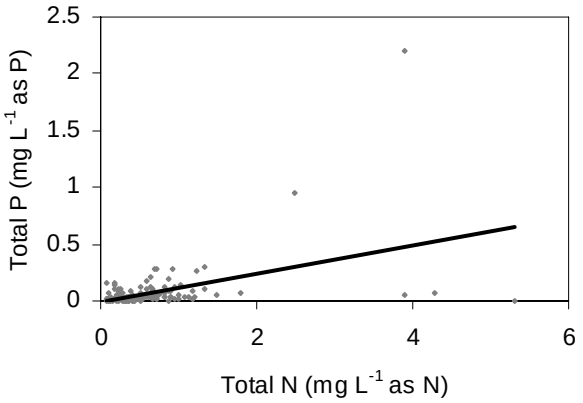
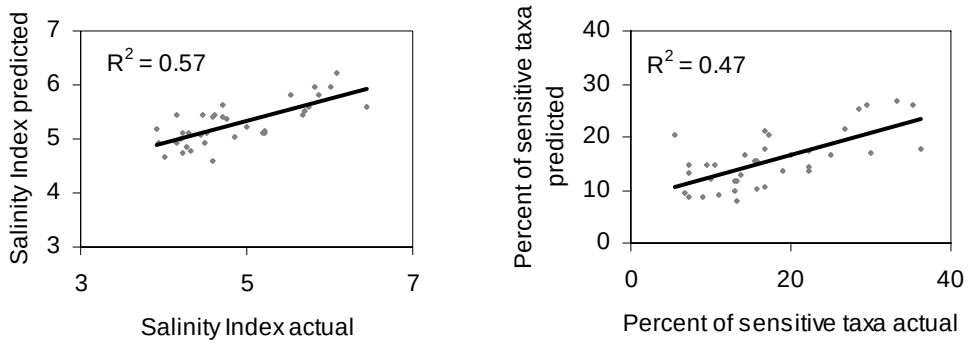


Figure 7.5. Scatter plot of total phosphorus versus total nitrogen with fitted trendline.

Simulation results

The correlation coefficient between the actual and predicted output for the simulation set was 0.75 for the Salinity Index and 0.68 for the Percent of Sensitive taxa (0.79 and 0.77 respectively for the training set). Figure 7.6 shows scatter plots with fitted linear trends and R^2 values for both variables.



a) b)

Figure 7.6. Actual versus predicted a) Salinity Index, and b) Percent of sensitive taxa, simulation dataset.

Figure 7.7 shows the range and the median of predicted outputs for SI and PST for both Scenario 1 and Scenario 2. It is obvious that the combined increases in the conductivity and nutrient concentrations had more effect on the macroinvertebrate communities than an increase in conductivity alone. For Scenario 1 the mean SI and PST decreased from 5.27 and 15.53 respectively in Simulation 1 (actual values for conductivity and related variables) to 4.68 and 9.94 in Simulation 5 (actual conductivity + 4000 $\mu S\ cm^{-1}$). When effect of nutrients has been added the Simulation 5 output for SI and PST was 4.32 and 7.14 respectively. Figure 7.8 shows

comparisons of the PST outputs for the Simulation 5 (+4000 μS cm^{-1}), under Scenario 1 (Conductivity), Scenario 2 (Combined) and only increase in nutrients (+ 4mg L^{-1} of total nitrogen, total phosphorus = 0.1251x (total nitrogen)- 0.01).

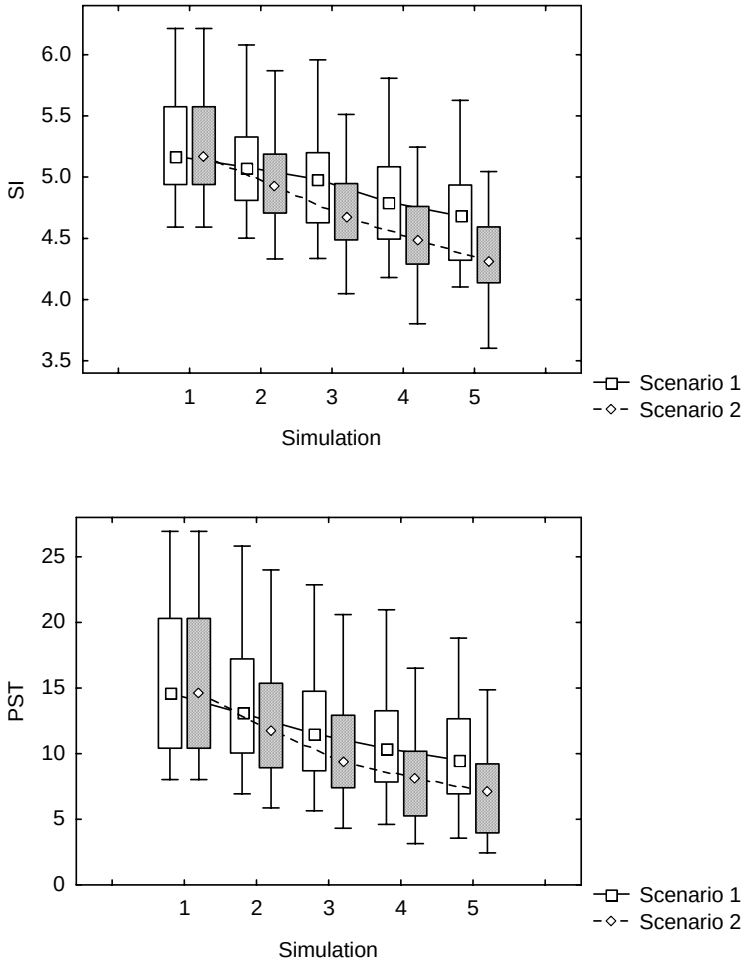


Figure 7.7. Box plots for simulation results for Scenario 1 and Scenario 2, median values, box 20-80%, whiskers minimum and maximum.

The lowest PST resulted from the combined impact of conductivity and nutrients. Nutrients only had the lowest impact on the percent of sensitive taxa.

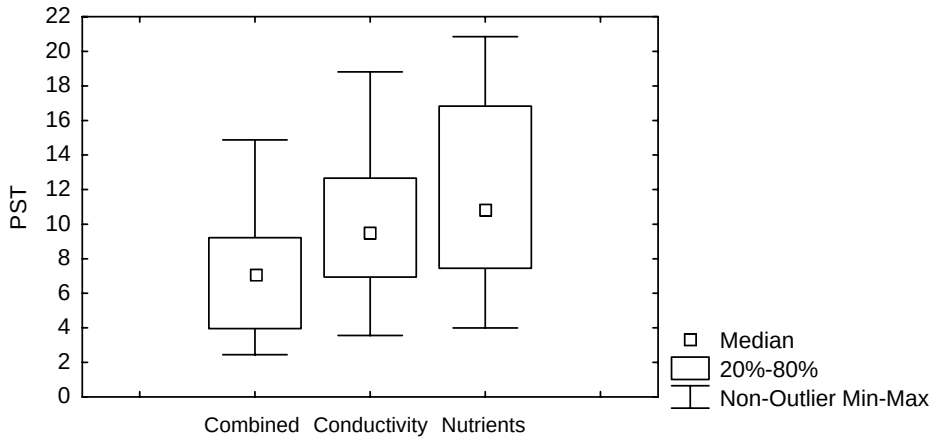


Figure 7.8. Box plot for the PST outputs for the Simulation 5 (+ 4000 μS cm^{-1}), for Scenario 1(Conductivity), Scenario 2 (Combined) and only increase in nutrients (+ 4mg L^{-1} of total nitrogen, total phosphorus = 0.1251x (total nitrogen)- 0.01).

Discussion and conclusions

It was possible to predict the SI and PST defined in the previous study (see Chapter 6) with reasonable accuracy (hypothesis 1 is true). Although the SI was defined using state-wide data it was still predictable on a smaller geographical scale. The model responded well to the increases in conductivity and changes in alkalinity, pH and nutrient concentration. According to the results of this study increase in conductivity in Central Queensland streams will result in loss of sensitive taxa and changes in the structure of macroinvertebrate communities (hypothesis 2 is true).

It has been shown that combined conductivity and nutrient concentrations may have a synergistic effect resulting in stronger impact on macroinvertebrate communities than single impacts of conductivity or nutrients alone (hypothesis 3 is true). This may also indicate that when the immediate effect of increased salinity appears as insignificant the indirect cumulative effects can still make the ecosystem vulnerable.

7.2 Using methods in combination: analysing results of the scenario analysis with Self Organising Maps (SOM).

Introduction

I have demonstrated the applicability and performance of SOM for the investigation of the natural variability in distribution of stream macroinvertebrates (Chapter 4) and elucidation of relationships between macroinvertebrate communities and

environmental variables (Chapter6). This study demonstrates applicability of SOM to the analysis of the results from the predictive modelling.

The previous study shown that the increase in conductivity results in the loss of sensitive taxa. It is clearly evident from Figure 7.7 that the loss of sensitive taxa and changes in macroinvertebrate communities occur at the different rate in different streams, ranging from 0 to 11%. What causes this difference? What are the similarities between sites were the high or low loss of sensitive taxa is predicted? In order to answer this question I designed the following study.

Methods

In order to investigate how the differences in the rate of loss of sensitive taxa can be explained by the conditions in each particular site I calculated the difference between the first simulation (actual data as predictor variables) and each subsequent simulation using Scenario 1 output. A new data matrix with four variables for the differences between simulated outputs was prepared. For each site we calculated the difference (d) as:

$d_i = s_1 - s_i$,

Where, i – simulation number, s – simulation output.

In order to group sites by similar pattern in PST change I partitioned this new matrix unto 4 clusters using a Self Organising Map (SOM) neural network. The optimum number of clusters was chosen using the Silhouette index (Rousseuw, 1987). Resulting clusters were further analysed using ANOVA.

Results

Resulted SOM (size 6x5, average SOM quantisation error = 2.307) was partitioned into 4 clusters with k-means algorithms using the Silhouette index (Rousseeuw, 1987) as an indicator of clustering quality. The Silhouette index for the resulting clustering was 0.79 indicating a strong structure.

Figure 7.9 shows a box plot with median values (whiskers for maximum and minimum) of differences (d) between simulation outputs in the clusters 1 and 2 (clusters 3 and 4 contained only few sites and not shown here). All the differences between simulations are noticeably larger in the first cluster than in the second, in other words, macroinvertebrate communities from the sites in cluster 1 show more pronounced response in all simulations than the communities from the cluster 2.

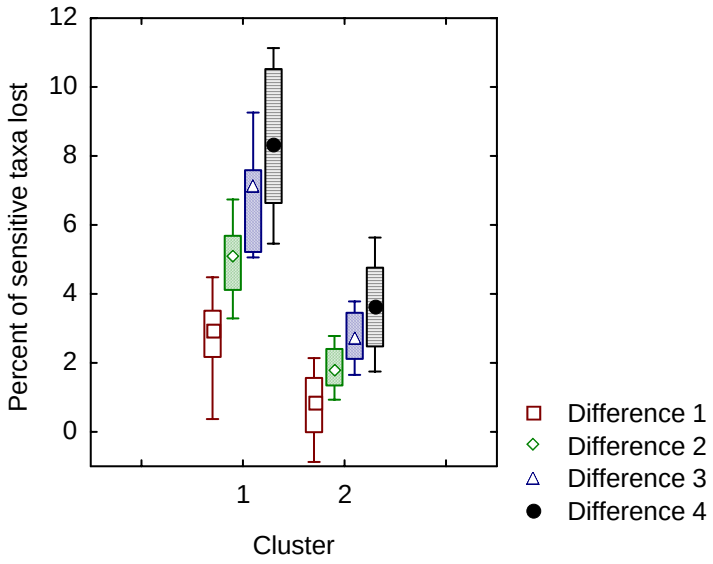


Figure 7.9. Box plot for the median values (box 80-20%, wiskers for maximum and minimum) of differences between simulation outputs in two SOM defined clusters.

On average, cluster 2 was characterized by lower percentage of sensitive taxa present initially, higher concentration of total nitrogen, higher turbidity, lower pH and lower dissolved oxygen in comparison with cluster 1 (Table 7.3). Clusters 3 contained only one site, which was characterized by the lowest initial PST – 6.6%. Even though turbidity was low at this site, the total nitrogen and conductivity were relatively high, plus the site was characterized by high water temperature. Cluster 4 contained only few sites also characterized by low PST, high water temperature and high conductivity.

Table 7.3. Mean values of the water quality variables and actual PST in four SOM defined clusters.

Number of the cluster	1	2	3	4
Number of sites in the cluster	14	18	1	3
Water temp (°C)	21.87	21.58	23.10	23.33
Dissolved oxygen (mg L^{-1})	8.40	7.61	7.60	6.06
Total N (mg L^{-1}as N)	0.46	0.73	0.83	0.41
Total P (mg L^{-1} as P)	0.04	0.06	0.03	0.01
Conductivity (µS cm^{-1})	438.14	713.4	1391.00	734.66
pH	8.18	7.46	8.10	7.75
Alkalinity (mg L^{-1} CaCO3)	172.98	141.23	408.00	189.16
Turbidity (NTU)	29.21	62.15	11.90	4.66
Actual PST	21.63	15.54	6.66	13.22

Total N, pH, and actual PST were significant in discriminating between clusters 1 and 2 (Table 7.4). Only 1 and 3 sites were in clusters 3 and 4 respectively, and these clusters are not considered in the further analysis. Although the statistical difference

in conductivity between clusters is not significant, the mean values of conductivity for cluster 1 and 2 are quite different, 438.14 μS cm^{-1} and 713.4 μS cm^{-1} respectively.

Table 7.4. Results of the univariate analysis of variance between clusters 1 and 2.

Variable	F	P
pH	9.4	0.00
Total N (mg L^{-1}as N)	5.6	0.02
Actual PST	4.0	0.05
Total P (mg L^{-1} as P)	2.0	0.17
Conductivity (μS cm^{-1})	1.9	0.18
Dissolved oxygen (mg L^{-1})	1.3	0.26
Turbidity (NTU)	0.9	0.62
Alkalinity (mg L^{-1} CaCO3)	0.6	0.53
Water temp (°C)	0.1	0.77

Discussion and conclusions

The response of the macroinvertebrate communities to increased salinity appeared to be site specific. When the changes in PST were analysed with SOM all streams were clustered into two broad groups, one that included sites where macroinvertebrate communities responded rapidly to the increase in conductivity and the other that responded much slower and experienced less significant loss of sensitive taxa.

The first group was characterised by initially larger PST, higher pH, lower nutrients and lower conductivities, even though the difference in conductivity between the groups was not statistically significant. This implies that the macroinvertebrate communities with a high number of sensitive taxa that usually occur at low conductivity and good water quality are likely to experience stronger changes when conductivity rises compared to the communities, which are already under some kind of stress and dominated by the opportunistic taxa. This may have some implications for the sustainable environmental management when decision is taken about the acceptable conductivity levels for the specific ecosystems. Further research is needed to understand the response of specific stream ecosystems to simulated impacts.

Chapter 8

Discussion and the recommendations for further research

This study was intended as an exploratory research into the applicability and the potential of two most well known types of neural network models for developing tools to assess and predict solutions for a range of ecological problems. It has been shown that extensive data collected over the years in accordance with AusRivAs protocols or for the other purposes can be utilized in a creative and flexible manner by the application of ANNs. When it comes to the assessment of biota, it was demonstrated that SDFS is capable of producing as much information as AusRivAs plus offers a potential in many more other aspects. The complimentary application of supervised and unsupervised ANNs within the SDSF proved to be a useful framework for the study of complex stream datasets.

In SDSF it is possible to implement referential approach in a more straightforward way, without having to cluster sites as in AusRivAs. In this study we demonstrated good performance of statewide models (Victoria) for the prediction of occurrence of 15 macroinvertertebrates, ranging from rare to very common. Even though these models have demonstrated high accuracy of prediction it is still possible to improve the accuracy by using models built for a smaller geographical area, as bioregion or catchments. Comparative accuracy of the models on the different spatial scale has been demonstrated using Queensland data. Similar comparisons conducted for the other states are likely to be beneficial in order to achieve the optimum performance and select the optimum spatial scale for the modeling. This can be a direction for the future research as we are not aware of many studies addressing this question in this time.

Understanding natural variability within macroinvertebrate communities is very important step for the design of habitat assessment protocols. Application of the reference approach must consider the inclusion of the population of reference sites representing the full range of conditions that are expected to occur at all sites to be assessed (Reynoldson and Wright, 2000). However, comparison of reference sites inhabited by the macroinvertebrate communities naturally different from those of the test sites can lead to erroneous assessment. Robust clustering procedures are very important in this respect. In this study we demonstrated clustering of macroinvertebrate assemblages using SOM with two main purposes: defining naturally similar spatial groups and defining similar assemblages of trophic groups.

Even though, this study demonstrated application of SOM to understanding natural variability by clustering reference sites in Victoria and Queensland, this task is too large to be thoroughly addressed in the scope of the thesis. Particularly in Queensland where only preliminary research of natural variability on statewide scale has been conducted (Jon Marshall, NR&M, personal communication). There is much space for the future research in this direction. SOM can be used to understand variability on a smaller scales, as within bioregions and within catchments. Another approach worth researching is clustering sites on the base of geo-climatic similarities, similarities in the structure of habitats, landuse, etc. However, when using SDSF the clustering step is not necessary in order to use the referential approach. Statewide or bioregional models can be implemented in a straightforward manner and clustering can be use in order to investigate particular questions or gain a general understanding of the data on various spatial scales.

This study has demonstrated the potential of MLP for the prediction of typified assemblages of macroinvertebrates instead of the separate taxa. This could significantly speed up the process of biological assessment, as fewer models would need to be built and tested. Reference sites were clustered by the similarity of their macroinvertebrate communities using SOM and these clusters were predicted and related to the environmental variables using MLP and GA. Even though, MLP outperformed GA in this case, GA offered more transparency and explanation by generating rule-sets describing the conditions that each site has to meet in order to be classified into a certain cluster. Implementation of rule-based methods as GA and decision trees can offer an additional information to that provided by SOM and MLP. The computer scientists are constantly developing new algorithms directed at providing ANN models with more transparency. Much research has been dedicated to generating explanation structures and rule refinement in ANN (Towell and Shavlik, 1993; Andrews and Geva, 1994), which, unfortunately, is not yet widely implemented in major ANNs software packages. Further research on the application of these algorithms to the ecological problems is likely to be helpful in order to provide the neural network models with more transparency and extract more information on the importance of environmental variable as structuring forces behind biotic communities.

The question of variable importance has been briefly addressed by comparing output of the sensitivity analysis using catchment scale models for three taxa. Unfortunately, variability of the results was too high, and we do have any confidence in this method at this stage. Similar variability has been observed by the other authors (Olden and Jackson, 2002) and further research in this direction is much needed.

Sensitivity curves, the other output from the sensitivity analysis also have been characterized by high variability and we do not recommend it as a quantitative tool at this stage. However, it has been shown that when analysed qualitatively, sensitivity curves can provide interesting information on the relationships between the distribution of macroinvertebrate communities and the environmental factors. The qualitative assessments of the sensitivity curves were largely in agreement with the outputs of the other methods as SOM component planes and frequency curves especially in the situation where strong relationships between taxa occurrence and particular environmental gradient have been observed.

By implementing the referential approach in AusRivAs it is possible to assess the site as degraded but there is not much information available on the type of impact and

specific consequences for the stream biota. In this study an ambitious task of creating a pressure-specific index (Salinity Index) has been attempted. SI was suggested as measurement of the changes within macroinvertebrate communities due to the effect of salinisation. Even though the use of SI on wide practical basis is not recommended at this stage, performance of SI has been verified by various methods from testing it on the subsets with otherwise very good water quality to subjecting the resulted Salinity Scores to partial CCA designed to rule out the effect of temporal and geoclimatic variables. Potential of semi-qualitative methods as SOM component planes and sensitivity curves for the assessment of taxa sensitivities to the environmental stressors (as salinity) has been demonstrated, however, more research and particularly comparison with quantitative methods as logistic models and multivariate regression can help better understanding the difference between the potential of statistical methods and the machine-learning methods.

Analysis of the changes within macroinvertebrate communities caused by the changes in salinity levels resulted in interesting and ecologically important observations. It has been shown that shift from the salt sensitive to the salt tolerant taxa occurs at the salinity levels much lower than currently accepted by the environmental authorities. Even though, no reduction in overall taxonomic richness has been observed, dramatic changes in the community structure were demonstrated at the conductivity levels around 800-1000 $\mu S\ cm^{-1}$. This is much lower than currently accepted level of 1500 $\mu S\ cm^{-1}$ (Hart et al., 1991), below which the freshwater biota considered to be unaffected.

The Salinity Sensitivity Scores assigned as a result of the analysis of sensitivity curves and mean/maximum conductivity values of each taxon were mainly in accordance with previous knowledge, however, some discrepancies were observed. For example, Kefford et al. (2003) based on the results of acute testing concluded that Australian mollusks are salt sensitive. According to our research the opposite was observed for some taxa like Planorbidae, which appears to be very tolerant to the salinisation. Recent results of the acute testing conducted by NR&M staff indicated that indeed some mollusks (as Hydrobiidae) are quite tolerant with LC50 > 20 000 $\mu S\ cm^{-1}$ in comparison with other taxa as mayflies Leptophlebiidae with LC50 around 2 000-6 000 $\mu S\ cm^{-1}$ (Jason Dunlop, NR&M, personal communication). We hope that our results might urge the scientists and environmental managers working on salinity issues reconsider the current assumptions about salinity sensitivity of Australian macroinvertebrate. Certainly, more research is needed in this direction.

Assessment and prediction using ‘dirty-water’ models including variables potentially affected by human activity is not implemented in AusRivAs and generally not often used in Australia. Huong et al. (2001) demonstrated high accuracy of the ‘dirty-water’ models for the prediction of occurrence pattern of stream macroinvertebrate using QLD data. This study demonstrated the high potential of the method for the prediction of taxonomic richness of stream macroinvertebrates and native macrophytes using relatively small dataset from the four catchments in NSW.

The ultimate purpose of the ‘dirty-water’ models was demonstrated by the prediction of the Salinity Index suggested as a measurement of changes in macroinvertebrate communities due to the effect of salinisation. Two scenarios were simulated: gradual increase in conductivity (up to + 4000 $\mu S\ cm^{-1}$) and combined increase in

conductivity and nutrients (total nitrogen). It has been shown that the increase in conductivity results in the loss of sensitive taxa (up to 11%), however this loss was even more profound when combination of stressors (conductivity and nutrients) has been simulated. James et al. (2003) indicated that little is known currently about the combined effect of conductivity and the other stressors, and indeed compounding of the effects is possible. We hope that the results of our research make an important contribution in this direction. It has been shown that scenario analysis using ANN models is an interesting and useful tool for the simulation of possible futures. For the future research, the following types of scenario analysis might be the most practically relevant in relation to the most common environmental problems faced by the Australian streams (as vegetation clearance, salinisation, sedimentation, agricultural run-off, water extraction, etc.):

1) Effect of salinisation, turbidity and nutrients on taxonomic richness, PET richness and sensitive taxa.

2) Effect of these factors in combination (as turbidity and nutrients).

3) Effect of the reduction in flow by itself and in combination with water quality variables.

4) Effect of water temperature increase (as a consequence of clearance of riparian vegetation) by itself and in combination with turbidity and nutrients.

Further experiments in order to predict the measurements of ecosystem functioning as benthic metabolism (respiration, production, etc.) using ANNs can be also very interesting and important. This could lead to the scenario analysis for the prediction of changes in ecosystem functioning in response to the various anthropogenic stresses or restoration measures.

In order to be able to conduct accurate and practically relevant scenario analysis it is important to realize the relationships between the predictor variables, so when changes in one variable is made, the variables somehow related should not be left unchanged, otherwise the simulation will be just a formal exercise and have little practical relevance. This thesis demonstrated the applicability of SOM component planes for the analysis of spatial distribution of biotic and abiotic variables and the relationships between them. It has been demonstrated using NSW data, that component planes allow comparison of many variables at ones and detection of both linear and non-linear relationships between them. We recommend wider use of SOM component planes when purpose is to get an initial understanding of data and its underlying structures. It was possible to make general assessments regarding conditions of habitats and biota in four NSW subcatchments just using SOM component planes, these assessments were largely in agreement with those conducted by the environmental authorities (Chessman, 2002).

Use of both SOM component planes and clustering SOM with k-means algorithm revealed a number of interesting relationships between water quality and trophic structure of macroinvertebrate communities. In general, it was shown that trophic structure is affected by a number of both natural geoclimatic characteristics and water quality parameters. Increase in water temperature, turbidity and elevated nutrients can affect natural succession of FFG along the stream order gradients. In particular,

elevated proportion of collectors and predators at the expense of the other trophic groups can be expected at the sites experiencing some kind of anthropogenic impact. Some interesting results from this study certainly deserve further attention, namely:

1) It appears that in a case of riffle habitat in Queensland succession of FFG along stream order gradient to some extent follows the assumptions of RCC when stream conditions are close to natural. In the case of edge habitat we could not observe any detectable changes in FFG along the river continuum neither in reference nor in test sites. One of the possible explanations is that edge habitat dataset includes many more sites from inland areas with low rainfall and intermittent flow, when riffle habitat by definition requires flowing water. The effect of flow might be masking natural gradients along the river continuum as ephemeral streams would be dominated by highly resilient fauna adapted to the surviving extended no flow periods despite of the site's stream order. The question for the further research is: What is the difference between succession of FFG along the river continuum in sites with intermittent flow and sites with permanent flow? Effect of combination in flow disturbances and changes in water quality on trophic structure can be also analysed using scenario analysis.

2) According to the results of CCA season appears to be important factor for the trophic structure of macroinvertebrate communities. As sites were sampled only in spring and autumn we assume that the difference is caused by the previous season, which in the case of Queensland is wet season (summer) or dry season (winter). However, we have not found any definitive difference in the pattern of FFG succession along the river continuum between two seasons. What is the effect of seasons on the trophic structure of macroinvertebrates in Australian streams? It has been observed (cited from Boulton and Brock, 1999) that community structure in Australia seems to be more related to small-scale, local variations in physical factors such as substrata and current velocity than to the large scale factors. Thus, patterns in feeding groups representation are more complicated than the river continuum concept suggests (Lake et al., 1986).

The investigation of the relationships between trophic groups and water quality also demonstrated the potential of untraditional combination of methods, namely SOM and CCA. This approach provided simultaneous reduction of the data dimensionality and the visualisation of relationships between different types of assemblages and environmental variables. This particular combination can be especially useful in the situations where there is a need to analyse a data containing large number of species (as diatom communities). Thus, it is possible to cluster species into typical assemblages using SOM and then relate those assemblages to the environmental gradients using CCA.

Flexibility and universality of SOM have also been demonstrated by analysing the outputs of scenario analysis in order to understand why the same increase in conductivity resulted in the different loss of sensitive taxa in different streams. It has been shown that communities already affected by some kind of anthropogenic activity and dominated by opportunistic taxa appear to be less affected by the salinisation in comparison with communities in close to the natural state. This observation is interesting in ecological terms and might have important implications in respect to environmental management.

Even though, it has been shown that macroinvertebrates are strongly and deterministically linked to the habitat features as substrate, discharge, hydraulics, riparian vegetation and water chemistry (Giller and Malmqvist, 1998), biotic interaction is another significant factor structuring macroinvertebrate communities. Bun and Davies (2000) showed that an assumption that changes in macroinvertebrate communities are always reflection of changes in environmental factors overrides importance of biological processes and can lead to erroneous conclusions about causes of these changes. For example, Power (1990) showed that in the presence of predatory fish, smaller predators were reduced, tube-weaving chironomids larvae proliferated, and the benthic substrate was reduced to a midge-infested residue. Predation effect of this kind can be a major cause of spatial and temporal variation in stream community structure. Biomonitoring models based entirely on abiotic variables would be unable to predict such marked changes in the nature of the stream. Predator-caused shift in the community structure can be mistakenly attributed to some form of anthropogenic disturbance. Modelling biotic interactions within macroinvertebrate communities and between fish and macroinvertebrate communities can be an interesting direction for the future research.

Another important question, which was not addressed in this study but certainly could be researched in the future, is the assessment and prediction of physical and chemical conditions. For example, Habitat Predictive Modelling approach uses large-scale catchment features to predict local-scale physical habitat features (Davies et al., 2000). The same logic and the statistical analysis are used as for the biological assessment in AusRivAs. Ratio of habitat features expected to occur at a test site (E) are compared against the habitat features that were actually observed at the site (O). The ratio of these values (O/E) indicates a continuum of physical habitat condition. Prediction of physical and chemical characteristics using ANNs is an interesting direction worth further researching.

This study gave some answers to the practical questions asked by the practicing ecologists, namely: How the accuracy of the model related to the number of predictor variables? Which predictive models are more accurate: generic or local? What is the effect of temporal variability and variability between different habitats on the accuracy of predictive models?

It has been shown that even though MLP performs slightly better with comprehensive set of predictors, it can cope with significant reduction of inputs. Reasonably accurate predictions were produced by the models with as few as 9 predictors. This corresponds well with previous findings, for example, Cereghino et al. (2003) developed an accurate model for the prediction of EPTC (Ephemeroptera, Plecoptera, Trichoptera, Coleoptera) richness using only 4 input variables. Local models were more accurate than generic but this accuracy was variable. Seasonal variations appear to be more important than annual, however, ANN models were still capable of achieving high rate of correct predictions when trained on one season and tested on another. These results can be an important step to the practical implementation of SDSF as a conceptual basis for the analysis, assessment and prediction of biological conditions in Australian streams.

From the Stream Decision Support Framework to the Stream Decisions Support System (SDSS).

There is enough evidence available worldwide that ANNs is a useful and flexible tool for the assessment and prediction of biological conditions. In Europe, SOM and MLP have been used as a basis for the software tool PAEQANN (Predicting Aquatic Ecosystems Quality using Artificial Neural Networks). The PAEQANN is a research project supported by the European Commission under the Fifth Framework Programme and contributing to the implementation of the Key Action "Sustainable Management and Quality of Water" within the Energy, Environment and Sustainable Development. PAEQANN project assemble European scientists in the aim to provide predictive tools to define effective policies and to improve freshwater management and to apply techniques identifying problems in ecosystem functioning for a future restoration of its integrity.

The PAEQANN's aim is to set up robust and sensitive ecosystem evaluation procedures that will work across a large range of running water ecosystems throughout Europe,

- Firstly to point out the cause and effect relationships between environmental conditions (physical, chemical, due to management actions) and certain relevant aquatic communities (diatoms, macroinvertebrates and fish).
- And then, to predict biocenosis structure in disturbed ecosystems, taking into account al the relevant ecological variables
- To test ecosystem sensitivity to disturbance.
- To explore specific actions to be taken for restoration of ecosystem integrity (Bretin et al., 2003).

PAEQANN provides a user interface where it is possible to select a country and organism for the assessment and modeling (Figure 8.1). It is possible then to conduct clustering of sites using SOM (Figure 8.2) or predict a taxa occurrence according to the referential approach and conduct sensitivity analysis using 'PaD' algorithm (Figure 8.3).

SDSF can be developed into a SDSS in a similar way. SOM can be used for the patterning, analysis of the relationships between the variables using component planes. MLP can be used for the assessment using referential approach, sensitivity analysis and scenario analysis. The models developed in this study and by Huong (2001) can be used as an initial basis for this system which can be developed further to provide a basis for analysis, assessment and prediction of freshwater biota.

Figure 8.1. PAEQANN interface.

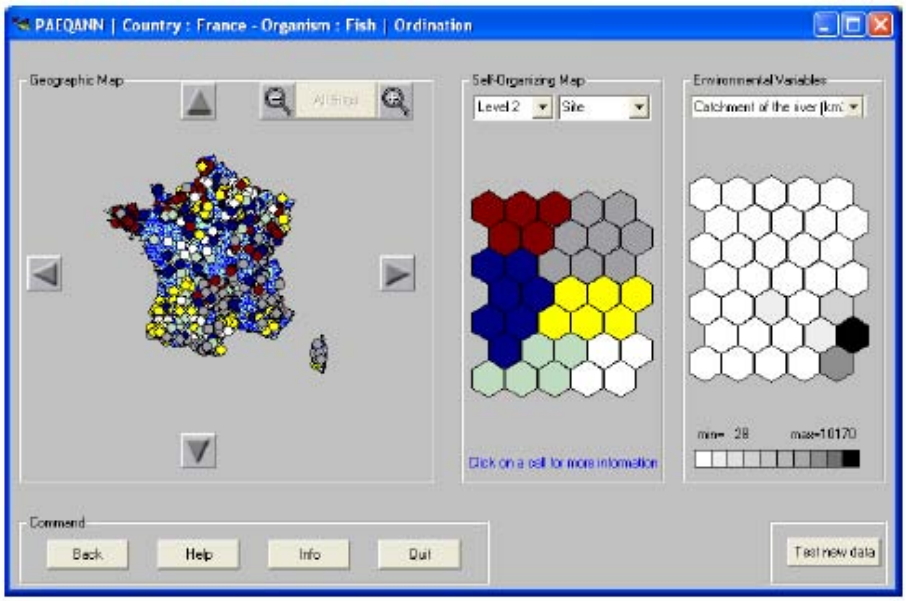


Figure 8.2. Ordination (using SOM) screenshot.

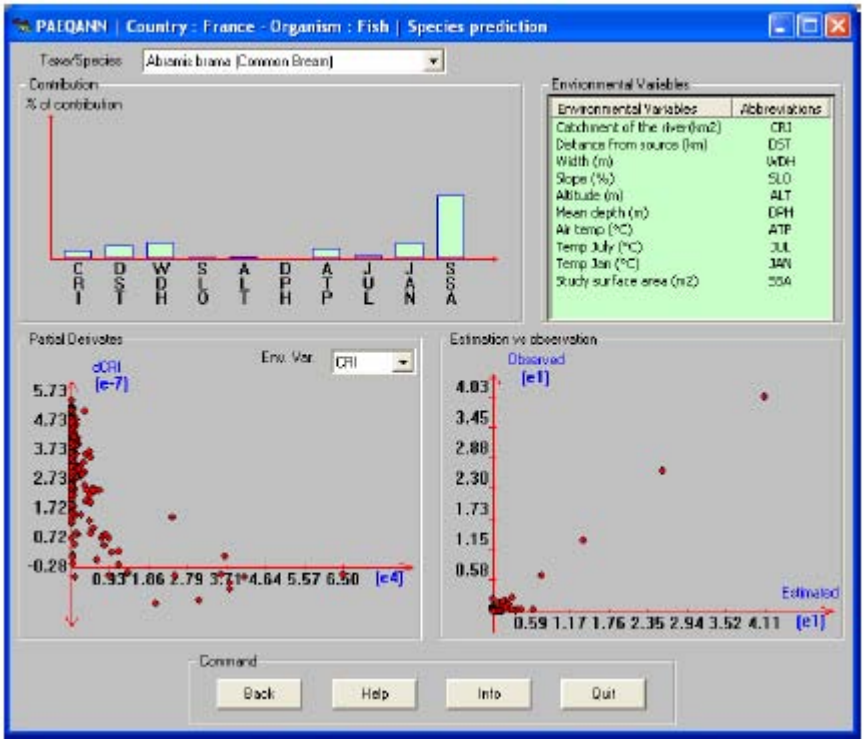


Figure 8.3. Prediction (using MLP) and sensitivity analysis screenshot.

Summary

Using SOM to explore the natural variability

Victoria data

- The clusters of sites with similar macroinvertebrate assemblages discovered by the SOM were largely in accordance with the previously defined bioregions of Victoria (Metzeling et al., 2001).

- All but one environmental variables were significant discriminating between the clusters, with alkalinity having the highest F value. The most dissimilar clusters were highland ecosystems and Murray plains.

- SOM component planes provided easy and highly visual way to assess relationship between variables.

Queensland data

- The patterns in distribution of macroinvertebrate communities discovered by the SOM vaguely resembled the biological regions defined by NR&M. Still,

SOM identified changes in community composition from south to north and from coastal areas to inland, which are the major directions in the distribution of bioregions.

- Three main environmental gradients have been identified using ANOVA and comparison of SOM component planes, namely: geographical location, rainfall pattern and the distance from source. Water quality, temporal variables and substrate composition appear to be less important but still significant in discriminating between the SOM clusters.

Recommendation for the future research:

SOM can be used to understand variability on a smaller scale, as within bioregions and within catchments. Clustering sites on the base of geo-climatic similarities, similarities in the structure of habitats, landuse, etc. can be helpful in order to identify regions with similar habitat characteristics.

Using SOM for the exploration of relationships between biotic and abiotic variables in NSW

- The SOM component planes provided quick and simple method for the evaluation of geographic distribution, possible gradients and correlations in the data.

- The distribution and the extent of anthropogenic impact in various areas of NSW have been characterised with reasonable accuracy using SOM component planes only.

- Detection of both linear (fishes and altitude) and non-linear relationships (macroinvertebrates, water temperature and turbidity) has been demonstrated using SOM component planes only.

Recommendation for the future research:

Closer look at the component planes for NSW dataset is likely to provide much more information than has been described here. SOM built for each subcatchment would provide more spatial resolution, and possibility to detect gradients on a finer scale.

Using SOM to explore the effect of water quality on the trophic structure of macroinvertebrate communities

- SOM produced meaningful clustering of macroinvertebrate communities based on the similarity of their trophic structure.

- Succession of FFG along the stream order gradient follows the assumptions of RCC to some extent when stream conditions are close to natural, but this trend was practically undetectable in the streams experiencing some kind of anthropogenic stress.

- Turbidity, water temperature and nutrients are the most important water quality variables affecting trophic structure of macroinvertebrate communities. Increase in turbidity causes reduction in the proportion of grazers and shredders and increase in the proportion of collectors and to some extent predators. The increase in nutrients is also associated with increase in the proportion of collectors and predators at the expense of the other groups.

- SOM component planes and CCA revealed largely similar relationships between FFG and water quality variables.

- The combination of SOM and CCA allowed simultaneous identification of similar trophic assemblages and the visualisation of relationships between those assemblages and water quality variables.

Recommendation for the future research:

The difference between succession of FFG along the river continuum in the sites with intermittent flow and the sites with permanent flow. The seasonal effect on the trophic structure of macroinvertebrates in Australian streams. The combination in flow disturbances and changes in water quality in relation to the trophic structure of macroinvertebrate communities can be also analysed using scenario analysis.

Predicting biotic variables by supervised ANNs

Using referential approach and Victorian dataset

- Good performance (average percentage of correct predictions 77.7%) of state-wide models (Victoria) for the prediction of occurrence of 15 macroinvertebrates, ranging from rare to very common has been demonstrated.

Using the dirty-water approach and NSW dataset

- The taxonomic diversity of macroinvertebrate and macrophytes has been predicted with a reasonable accuracy (correlation between actual and predicted output 0.7 and 0.79 respectively).

Recommendation for the future research:

Even though these models have demonstrated high accuracy of prediction it is still possible to improve the accuracy by building models for a smaller geographical area, as bioregion or catchments.

Optimisation of the modelling design

- Even though MLP performs slightly better with comprehensive set of predictors, it can cope with significant reduction of inputs. Reasonably accurate predictions were produced by the models with as few as 9 predictors.

- Local models were more accurate than generic but this accuracy was variable.

- Seasonal variations appear to be more important than annual, however, ANN models were still capable of achieving high rate of correct predictions when trained on one season and tested on the other.

- Habitat specific models are more accurate when simulated using the data from the same habitat, however, relatively accurate predictions are still possible using data from the different habitat.

Recommendation for the future research:

These results can be an important step to the practical implementation of SDSF as a conceptual basis for the analysis, assessment and prediction of biological conditions in Australian streams and developing it into a Stream Decision Support System.

Prediction of SOM defined groups with MLP and GA

- It was possible to predict SOM defined groups using referential model with good accuracy.

- MLP outperformed GA on approximately 10%, although both methods were able to meet the threshold of 70% of correct predictions.

- GA rules provided more quantitative explanation.

Recommendation for the future research:

The prediction of defined macroinvertebrate assemblages instead of separate taxa can be used as an extension of the referential approach. If the type of macroinvertebrate assemblage predicted does not match the one actually found, it might be then compared with other assemblages indicative for various anthropogenic stresses.

Stability and quantitative application of the sensitivity analysis

- The question of variable importance has been briefly addressed by comparing output of the sensitivity analysis using catchment scale models for three taxa. Unfortunately, variability of the results was too high, and we do have any confidence in this method at this stage.

- Sensitivity curves also have been characterized by high variability and we do not recommend them as a quantitative tool at this stage. However, when analysed qualitatively, they can provide interesting information on the relationships between the distribution of macroinvertebrate communities and the environmental factors.

Sensitivity of the stream macroinvertebrates to the changes in salinity and the development of Salinity Index

- The qualitative assessments of the sensitivity curves were largely in agreement with the outputs of the other methods as SOM component planes and frequency curves, especially in the situations where strong relationships between taxa occurrence and the conductivity gradient have been observed.

- The Salinity Sensitivity Scores assigned as a result of the analysis of sensitivity curves and mean/maximum conductivity values of each taxon were mainly in accordance with previous knowledge, however, some discrepancies were observed.

- SI was suggested as measurement of the changes within macroinvertebrate communities due to the effect of salinisation.

- The shift from the salt sensitive to the salt tolerant taxa occurs at the salinity levels much lower than currently accepted by the environmental authorities. Even though, no reduction in overall taxonomic richness has been observed, dramatic changes in the community structure were demonstrated at the conductivity levels around 800-1000 $\mu S\ cm^{-1}$. This is much lower than currently accepted level of 1500 $\mu S\ cm^{-1}$ (Hart et al., 1991), below which the freshwater biota considered to be unaffected.

- The performance of SI has been verified by the various methods from testing it on the subsets with otherwise very good water quality to subjecting the resulted Salinity Scores to partial CCA designed to rule out the effect of temporal and geoclimatic variables.

Recommendation for the future research:

Potential of semi-qualitative methods as SOM component planes and sensitivity curves for the assessment of taxa sensitivities to the environmental stressors (as salinity) has been demonstrated, however, more research and particularly comparison with quantitative methods as logistic models and multivariate regression can help better understanding the difference between the potential of statistical methods and the machine-learning methods. Some discrepancies between our results and previous knowledge regarding salt sensitivity of some taxa (as Planorbidae) need to be researched further.

Scenario analysis: predicting the effect of secondary salinisation in Central Queensland

- The increase in conductivity results in the decrease of SI and the loss of sensitive taxa (up to 11%), however, this loss was even more profound when combination of stressors (conductivity and nutrients) has been simulated.

- Flexibility and universality of SOM have been demonstrated by analysing the outputs of scenario analysis in order to understand why the same increase in conductivity resulted in the different loss of sensitive taxa in different streams. It has been shown that communities already affected by some kind of

anthropogenic activity and dominated by opportunistic taxa appear to be less affected by the salinisation in comparison with communities in close to the natural state.

Recommendation for the future research:

Little is known currently about the combined effect of conductivity and the other stressors. For the future research, the following types of scenario analysis might be the most practically relevant in relation to the most common environmental problems faced by the Australian streams (as vegetation clearance, salinisation, sedimentation, agricultural run-off, water extraction, etc.):

1) Effect of salinisation, turbidity and nutrients on taxonomic richness, PET richness and sensitive taxa.

2) Effect of these factors in combination (as turbidity and nutrients).

3) Effect of the reduction in flow by itself and in combination with water quality variables.

4) Effect of water temperature increase (as a consequence of clearance of riparian vegetation) by itself and in combination with turbidity and nutrients.

Further experiments in order to predict the measurements of ecosystem functioning as benthic metabolism (respiration, production, etc.) using ANNs can be also very interesting and important. This could lead to the scenario analysis for the prediction of changes in ecosystem functioning in response to the various anthropogenic stresses or restoration measures.

Bibliography

Allan J.D. (1995). *Stream ecology: Structure and Function of Running Water*. Chapman and Hall. London.

Allan, D. J. (2004). Landscapes and riverscapes: The influence of Land Use on Stream Ecosystems. *Annu. Rev. Ecol. Evol. Syst.* 35, 257-84.

Ball, G.R., Palmer-Brown, D., Mills, G.E. (2000). A comparison of Artificial Neuronal Network and Conventional Statistical Techniques for Analysing Environmental Data. *Artificial Neural Networks Application to Ecology and Evolution*. Springer, p. 262.

Baeck, T., (1996). *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York.

Bayly, I.A.E. (1969). The occurrence of calanoid copepods in athalassic saline waters in relation to salinity and ionic proportions. *Vernandlungen Internationale Vereinigung fur Theoretische and Angewandthe Limnologie* 17, 449-455.

Bishop C.M. (1995). *Neural Networks for Pattern Recognition.* (Oxford University Press Inc.: New York.)

Bishop, C. M. (1998). *Neural Network and Mashine Learning*. New York. Springer.

Blayo, F., and Demartines, P. (1991). Data analysis: how to compare Kohonen neural networks to other techniques? In A. Prieto (ed.) Artificial Neural Networks. International Workshop IWANN'91. Heidelberg Springer, Berlin.

Bloedel, L., Wilhelm, G., Clarke, R., Horn, T., and Churchill R. (2000). *Water Quality Exceedence, Trend and Status Assessment for Queensland*. Department of Natural Resources, 42 (Queensland: Australia.)

Bobbin. J. and Recknagel, F.(2001).Knowledge discovery for prediction and explanation of blue-green algaldynamics in lakes by evolutionary algorithms. *Ecological Modelling* 146, 253 – 262.

Borcard D., Legendre, P. and Drapeu, P. (1992). Patialling out the spatial component of ecological variation, *Ecology* 73, 1045-1055.

Boulton A.J. and Brock, M.A.(1999). *Australian Freshwater Ecology. Processes and Management*. Cooperative Research Centre for freshwater ecology.Gleneagles Publishing. 300 p.

Brosse, S., Giraudel, J.L., Lek, S. (2001). Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecological Modelling* 146, 159-166.

Bretin, L.P., Park, Y.S., Gevrey, M., Lek, S. (2003). PAEQANN tools. Manual for users. http://aquaeco.ups-tlse.fr/

Brodie, J.E., Hicks, W.S., Richards, C.C., and Thomas, F.G. (1984). Residues related to Agricultural chemicals in the groundwaters of the Burdekin River Delta, North Queensland. Environmental Pollution. 8, 187-215.

Brosse, S. Giraudel, J.L., Lek, S. (2001). Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecological Modelling* 146, 159-166.

Bunn, S.E and Davies, P.M. (1992). Community structure of the macroinvertebrate fauna and water quality of a saline river system in south-western Australia. *Hydrobiologia* 248, 143-160.

Bunn, S. E. and Davies, P.M. (2000) Biological processes in running waters and their implications for the assessment of ecological integrity. *Hydrobiologia* 422/423: 61-70.

Cairns, J. J. and Pratt, J. R. (1993). A history of biological monitoring using benthic macroinvertebrates. *Freshwater Biomonitoring and Benthic Macroinvertebrates. D. M. Rosenberg and V. H. Resh.* New York, Chapman & Hall.

Cereghino, R., Giraudel, J.L. and Compin, A. (2000). Spacial analysis of stream invertebrates distribution in Adour-Garonne drainage basin (France), using Kohonen Self Organizing Maps. *2nd International Conference of Applications of Machine Learning to Ecological Modelling*, Adelaide.

Cereghino, R. (in press). Spatial distribution patterns of macroinvertebrate functional groups over the riverbed in stony streams. *Oecologia.*

Chessman, B.C. (1986). Dietary Studies of Aquatic Insects from two Victorian Rivers. *Aust. J. Mar. Freshw. Res.* 37, 129-46.

Chessman, B. C. (1999). Predicting the macroinvertebrate faunas of rivers by multiple regression of biological and environmental differences. *Freshwater Biology* 41, 747-757.

Chessman, B. (2002). Assessing the conservation value and health of New South Wales rivers. The PBH (Pressure-Biota-Habitat) project. NSW Department of Land and Water Conservation. Canberra. 63p.

Chessman, B. (2003). New sensitivity grades for Australian river macroinvertebrates. *Marine and Freshwater Research* 54, 95-103.

Chon T.-S., Young Seuk Park, Kyong Hi Moon, Eui Young Cha (1996). Patternizing communities by using an artificial neural network. *Ecological Modelling* 90, Issue 1, Pages 69-78

Chon T.-S., Park Y-S., Cha E.Y. (2000a). Streams for the Short Time Prediction by Temporal Artificial Neuronal Networks. *Artificial Neural Networks Application to Ecology and Evolution*. Springer, p. 262

Chon. T-S., Park, Y.S., Kwak, I.-S., Cha, E.Y. (2000b). Non-linear Approach to Grouping, Dynamics and Organizational Informatics of Benthic Macroinvertebrate Communities in Streams by Artificial Neural Networks. Ecological Informatics. Understanding Ecology by Biologically –Inspired Computation. Springer. Pp. 127-178.

Chon T.-S., Young Seuk Park, June Ho Park (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 YooShin Kim. (2001). Patterning and short-term predictions of benthic macroinvertebrate community dynamics by using a recurrent artificial neural network. *Ecological Modelling* 146, 181 – 193.

Godinho, F.N, Ferreira, M.T., and Santos, J.M. (2000). Variation in fish community composition along an Iberian river basin from low to high discharge: relative contributions of environmental and temporal variables. *Ecology of Freshwater Fish* 9, 22-29.

Conrick, D., and Cockayne, B. (2000). Queensland Australian River Assessment System (AusRivAs) Sampling and processing manual. Queensland Department of Natural Resources. Freshwater Biological Monitoring Unit (Brisbane: Australia.)

Coysh, J., Nichols, S., Simpson, S., Norris, R., Barmuta, L., Chessman, B. and Blackman, P. (2000). *Australian River Assessment System – National River Health Program Predictive Model Manual.* Canberra, ACT., CRC Freshwater Ecology, University of Canberra.

Cranton P.S., Fairweather P. and Clarke G. (1996) Biological indicators of water quality. *Indicators of Catchment Health: A Technical Perspective*. (Eds J.Walker and D.J. Reuter). CSIRO Publishing. Melbourne.

Cummins, K.W., (1973). Trophic relations of aquatic insects. *Annu Rev. Entomol.* 8, 183-206.

D'Angelo, D, J., Howard, L, M., Meyer, J. L., Gregory, S, V., Ashkenas, L, R., (1995). Ecological uses for genetic algorithms: predicting fish distributions in complex physical habitats. *Can. J. Fish. Aquat. Sci.* 52, 1893-1908.

Davies, D.L., Bouldin, D.W. (1979). A Cluster Separation Measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224-227.

Davies, P.E. (1994). National River Processes and Management Program. Monitoring River Health Initiative. River Bioassessment Manual, Version 1.0. Department of Environment, Sport and Territories, Land and Water Resources Research and Development Corporation, Commonwealth Environment Protection Agency. (LWRRDC:Canberra).

De'ath, G.K. (2002). Multivariate regression-trees: a new technique for modelling species–environment relationships. *Ecology* 83,1105–1117.

Dedecker, A.P, Goethals, P.L., and De Pauw N. (2003). Overview and quantification of the factors affecting the upstream and downstream movements of Gammarus pulex (Amphipoda). *Commun. Agric. Appl. Bio.l Sci.* 68(1),25-31.

Dedecker, A.P., Goethalis, P.L.M., D'heygere, T., Gevrey, M., Lek, S., and De Pauw, N. (submitted) Application of Artificial Neural Network models to analyse the relationships between Gammarus pulex L. (Crustacea, Amphipoda) and river characteristics.

Delong, M.D. and Brusven, M.A. (1998). Macroinvertebrate community structure along the longitudinal gradient of an agriculturally impacted stream. *Environmental Management* 22, 445-57.

Department of Environment and Heritage and Department of Natural Resources (1999). Testing the waters. A report of the quality of Queensland waters 21 (Queensland Government: Australia.)

Digby P.G.N. and Kempton, R.A. (1987). *Multivariate analysis of ecological communities.* Chapman &Hall Ltd, London.

Foody, G.M. (1999). Applications of the self-organising feature map neural network in community data analysis. *Ecological Modelling* 120, 97-107.

Gauch, H. G. (1989). *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge.

Giraudel, J.L., Aurelle, D., Berrebi, P., Lek, S. (2000). Application of the Self-Organizing Mapping and Fuzzy Clustering to Mictosatellite Data: How to Detect Genetic Structure in Brown trout (Salmo trutta) Populations. *Artificial Neural Networks: Application to Ecology and Evolution*. Springer.

Giraudel, J. L. and Lek, S. (2001). A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling* 146, 329-339.

Giraudel, J. L. and Lek, S. (2002). The Structuring Index: a tool for analysing Self-Organising Map. http: //aquaeco.ups-tlse.fr/Results/workshop/SlideRome/Giraudel.htm

Goethals, P., Dedecker, A., Gabriels, W. and N. De Pauw (2003). Development and application of predictive river ecosystem models based on classification trees

and artificial neural networks. In: Recknagel, F. (ed.), 2003. *Ecological Informatics. Understanding Ecology by Biologically-Inspired Computation*. Springer-Verlag, New York, 91-107.

Gordon, I. (2002). Salinity in Queensland. Queensland Department of Natural Resources and Mines. Fact sheet. http://www.nrm.qld.gov.au/

Gozlan, R. E., Mastrorillo, S., Copp, G. H. and Lek, S. (1999). Predicting the structure and diversity of young-of-the-year fish assemblages in large rivers. *Freshwater Biology* 41(4), 809-820.

Hart, B.T., Bailey, P, Edwards, R., Hortle K., James, K, McMahon, A. (1991). A review of the salt sensitivity of the Australian freshwater biota. *Hydrobiologia* 210, 105 –144.

Hart, B. T., Lake, P. S., Webb, J.A., and Grace, M.R. (2003). Ecological risk to aquatic systems from salinity increases. *Australian Journal of Botany* 51, 689-702.

Haykin, S. (1999). *Neural Networks: A comprehensive Foundation.* (Prentice Hall:New Jersey.)

Hawkes, H. A. (1998). Origin and Development of the Biological Monitoring Working Party (BMWP) Score System. *Water Research* , 32 (3), 964-968.

Hawking, J.H. and Smith, F.J (1997). *Colour Guide to invertebrates of Australian inland waters.* (CRC for Freshwater Ecology, Albury.) 213pp.

d'Heygere, T., Goethals, P. and De Pauw, N. (2001). Application of evolutionary algorithms for input variables selection of classification tree models predicting benthic macroinvertebrate communities in watercourses of Flanders (Belgium). *Conf: Parameter selection in modelling aquatic community structure.*

Helawell, J. M. (1986). *Biological Indicators of Freshwater Pollution and Environmental Management*. London & New York, Elsevier.

Humphries, P., Davies, P.E. and Mulcahy M.E. (1996). Macroinvertebrate assemblages of littoral habitats in the Macquarie and Mersey Rivers, Tasmania: implications for the management of regulated rivers. *Regulated Rivers: Research and Management* 12, 99-122.

Huoung, H. (2001). Predicting freshwater habitat conditions by distribution of macroinvertebrates using Artificial Neural Network. MSc Thesis. Faculty of Agriculture and Natural Science, Department of Soil and Water, University of Adelaide. 174p.

Huong, H., Recknagel, F., Marshall, J. and Satish Choy. (2001).Predictive modelling of macroinvertebrate assemblages for stream habitat assessments in Queensland (Australia). *Ecological Modelling* 146, 195 – 206.

Huong, H., Recknagel, F., Marshall, J., and Choy, S. (2003). Elucidation of hypothetical relationships between habitat conditions and macroinvertebrate assemblages in freshwater streams by artificial neural networks. In *Ecological Informatics. Understanding Ecology by Means of Biologically-inspired Computation.* (Ed F. Reckangel.) pp. 179-192. (Springer-Verlag: New York.)

Hynes H.B.N. (1960). *The Biology of Polluted Water.* Liverpool University Press. Liverpool.

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, v.2, 813-818.

Jain, L. C., Martin N. M. (1999). *Fusion of neural networks, fuzzy sets, and genetic algorithms : industrial applications*. Boca Raton, Fl. : CRC press.354 p.

James, K.J., Cant, B., and Ryan, T. (2003). Responses of freshwater biota to rising salinity levels and implications for saline water management: a review. *Australian Journal of Botany* 51, 703-713.

Jeong, Kwang-Seuk, and Joo, G.J. (submitted). Linear and non-linear modelling of the river phytoplankton dynamics: the predictive Multi-Linear Regression and Artificial Neural Network model. *Natural Computing*.

Johnson, R.K. and Goedkoop, W. (2002). Littoral macroinvertebrate communities: spatial scale and ecological relationships. *Freshwater Ecology* 47, 1840-1854.

Joy, M.K. and Death, R.G. (2004). Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. *Freshwater Biology* 49, 1036-1052.

Karr J.R. (1998). Rivers as sentinels: using the biology of rivers to guide landscape management. *River Ecology and Management: Lessons from the Pacific Coastal Ecoregion* (Eds Naiman R.J. and Bilby R.E.) Springer-Verlag. New York.

Karr, J. R. (1981). Assessment of biotic integrity using fish communities. *Fisheries* 6(6): 21-27.

Karr, J.R. (1991). Biological integrity: a long-neglected aspect of water resource management. *Ecological Applications* 1, 66-84.

Karul, C., Soypak, S., Cileciz, A.F., Akbay, N. and Germen, E. (2000). Case studies on the use of neural networks in eutrophication modelling. *Ecological Modelling* 134, 145-152.

Kerans B.L. and Karr J.R. (1994). A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. *Ecological Applications* 4, 768-785.

Kay, W. R., Halse, S.A.,.Scanlon, M.D and Smith, M.J. (2001). Distribution and environmental tolerances of aquatic macroinvertebrate families in the agricultural zone of southwestern Australia. *Journal of North American Benthological Society* 20(2), 182-199.

Kefford, B.J. (1998). The relationship between electrical conductivity and selected macroinvertebrate communities in four river systems of south-west Victoria, Australia. *International Journal of Salt Lake Research* 7, 151-170.

Kefford, B.J., Dalton, A, Palmer, C.G. and Dayanthi Nugegoda, (2004). The salinity tolerance of eggs and hatchlings of selected aquatic macroinvertebrates in south-east Australia and South Africa. *Hydrobiologia* 517, 179-192.

Kefford, B.J. (2000). The effect of saline water disposal: implications for monitoring programs and management. *Environmental Monitoring and Assessment* 63, 313-327.

Kefford, B.J., Papas, P.J. and Dayanthi Nugegoda, (2003). Relative salinity tolerance of macroinvertebrates from the Barwon River, Victoria, Australia. *Marine and Freshwater Research* 54, 755-765.

Kefford, B.J, Paradise, T, Papas, P.J, Field, E., and Dyanthi Nugegoda (2003). Assessment of a System to Predict the loss of Aquatic Bioderversity from changes in salinity. Project No: VCE 17. Final Report to Land and Water Australia.

Kohonen, T. (1995). Self-Organising Maps. Springer-Verlag, Heidelberg.

Koning, N. and Roos, JC. (1999). The continued influence of organic pollution on the water quality of the turbid Modder River. *Water SA* Vol. 25 No.3, 285-292.

Lake, P.S., Barmuta, L.A., Boulton, A.J., Campbell, I.C. and St Clair, R.M. (1986). Australian streams and Northern Hemisphere stream ecology: comparisons and problems. *Proeeding Ecological Society of Australia* 14, 61-82.

Legendre P. and Legendre, L. (1998). *Numerical Ecology*, Elsevier Scientific, Amsterdam.

Lek, S. and Guegan, J.F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* 120, 65-73.

Lek, S. and Guegan, J.F. (2000). *Artificial Neuronal Networks. Application to Ecology and Evolution*. Springer. Pp. 4-22**.**

Lek S., Delacose M., Baran P., Dimopoulos I., Lauga J. and Aulagnier S. (1996). Application of neural networks to modelling non-linear relationship in ecology. *Ecological Modelling* 90, 39-52

Lek-Ang, S. L. D. and Lek, S. (1999). Predictive modelling of Collembola diversity and abundance in riparian habitat. *Ecological Modelling* 120, 247-260.

Lin, C-T. and Lee, C.S. (1996). *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems*. New Jersey, Prentice Hall PTP.

Luger, G. F. and Stubblefield, W. A. (1993). *Artificial Intelligence: Structure and Strategies for Complex Problem Solving.* California, The Benjamin/Cummings Publishing Company.

MacQueen J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297

Maier, H.G. and Dandy, G. C. (1996). The use of Artificial Neural Network for the prediction of water quality parameters. *Water Resources Research* 32(4), 1013-1022.

Maier, H.G., Dandy, G.C. and Burch, M.D. (1998). Use of ANN for modelling cyanobacteria Anabena species in the river Murray, SA. *Ecological Modelling* 105, 257-272.

Maier, H.R., Dandy, G.C. (1998). The effect of internal parameters and geometry on the performance of back propagation neural networks: an empirical study. *Environmental Modelling and Software* 13 (2), 193-209.

Manel, S., Dias, J-M., Ormerod, S. J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling* 120, 337-347.

Mason C.F. (1996) *Biology of Freshwater pollution.* 3rd edition. Longman. England.

Marchant, R., Hirst, A., Norris, R. and Metzeling, L. (1999). Classification of macroinvertebrate communities across drainage basins in Victoria, Australia: consequences of sampling on broad spatial scale for predictive modelling. *Freshwater Biology* 41, 253-268.

Marshall, N.A. and Bailey, P.C.E. (2004). Impact of secondary salinisation on freshwater ecosystems: effects of contrasting, experimental, short-term releases of saline waistewater on macroinvertebrates in a lowland stream. *Marine and Freshwater Research* 55, 509-523.

Marshall, J., Huong, H., Choy, S., and Rechnagel, F. (2002). Relationships between habitat properties and the occurrence of macroinvertebrates in Queensland streams (Australia) discovered by a sensitivity analysis with artificial neural networks. *Vernandlungen Internationale Vereinigung fur Theoretische and Angewandthe Limnologie* 28, 1-5.

McNeil, V.H., and Cox, M.E.(in press). Chemical types of waters and assessment of spatial variation in streams throughout Queensland, Australia. *Hydrological Sciences Journal.*

Merritt, R.W. and Cummins, K.W. (1978). An Introduction to the Aquatic insects of North America. (Kendall/Hunt, Dubuque, Iowa.)

Metzeling, L. (1993). Benthic Macroinvertebrate Community Structure in Streams of Different Salinitites. *Australian Journal of Marine and Freshwater Research* 44, 335-51.

Metzeling, L., Wells, F., Newall. P., Tiller, D., Reed, J. (2001). Biological objectives for rivers and stream – ecosystem protection. Policy background paper. Environmental Protection Authority Victoria.

Moss, D., Wright, J. F. and Armitage, P. D. (1987). The prediction of the macroinvertebrate fauna of upoluted running water sites in Great Britain using environmental data. *Freshwater Biology* 17, 41-52

Norris, R. H. (1999). What is river health? *Freshwater Biology* 41, 197-209.

Obach, M., Wagner, R., Werner, H., Schmidt, H-H. (2001). Modelling population dynamics of aquatic insects with artificial neural networks. *Ecological Modelling* 146, 207-217.

Olden, J. D., and Jackson, D. A. (2002). A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology* 47, 1976-1995.

Olden, J.D., Joy, M.K. and Death, R.G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* 178, 389-397.

Oliver, R.L., Hart, B.T., Olley, J., Grace, M, Rees, C. and Caitcheon, G. (1999). The Darling river: algal growth and the cycling and sources of nutrients. Murray-Darling basin commission project M386. CRC for Freshwater ecology, CSIRO Land and Water.

Park, Y-S., Kwak I-S., Chon T_S., Jwa-Kwan Kim, Sven Erik Jorgensen (2001). Implementation of artificial neural network in patterning and prediction of exergy in response to temporal dynamics of benthic macroinvertebrate communities in streams. *Ecological Modelling* 146, 143-157.

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

Parsons, M. and Norris, R.(1996). The effect of habitat-specific sampling on biological assessment of water quality using a predictive model. *Freshwater Biology* 41, 43-49.

Paruelo, J. M. and Tomasel, F. (1997). Prediction of functional characteristic of ecosystem: a comparison of ANN and regression models. *Ecological Modelling* 98, 173-186.

Peterson, G. D., Cumming, G. S., and Carpenter, S. R. (2003). Scenario Planning: A Tool for Conservation in an Uncertain World. *Conservation Biology* 17/2, 358-366.

Principe, J.C., Euliano, N.R., Lefebvre, W.C. (2000). *Neural and Adaptive Systems. Fundamentals Trough Simulations*. John Wiley & Son, Inc., 656.

Poff, LeRoy N. (1996). Stream hydrological and ecological responses to climate change assessed with an artificial neural network. *Limnology and Oceanography* 41(5), 857-863.

Power, M.E. (1990). Effect of fish in river food webs. *Science* 250, 811-814.

Pudmezky, A., Marshall, J. and Satish., C. (1998). *Preliminary application of Artificial Neural Network model for predicting macroinvertebrates in rivers*. Queensland, Water Monitoring Group – Queensland Department of Natural Resources.

Quinn, G.P. and Keough, M.J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press. Pp. 467-472.

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

Recknagel, F. and Wilson, H. (2000). Elucidation and prediction of aquatic ecosystem by Artificial Neural Networks. *Artificial Neural Networks: Application to Ecology and Evolution.* S.Lek and J. F. Guegan, Berlin, New York, Springer Verlag.

Recknagel, F., (2001). Application of machine learning for ecological modelling. *Ecological Modelling* 146, 1-3, 303-310.

Recknagel, F., Bobbin, J., Whigham, P. and H. Wilson (2002). Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. *Journal of Hydroinformatics*. IWA Publishing. 4, 2, 125-134.

Reynoldson, T.B. and Wright, J.F. 2000. The reference condition: Problems and Solutions, in: J.F. Wright, D.W. Sutcliffe and M.T.Furse (eds), Assessing the Biological Quality of Fresh Waters: RIVPACA and other Techniques, Freshwater Biological Association, Ambleside, UK., pp.293-250.

Rojas, R. (1996). *Neural Networks: A systematic introduction*. Springer. 503p.

Rosenberg D.M. and Resh V.H. (Eds.) (1993). *Freshwater Biomonitoring and Benthic Macroinvertebrates.* Chapman and Hall. NewYork and London.

Rousseeuw P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp App. Math* 20, 53-65.

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

Scardi, M. (1996). Artificial neural networks as empirical models for estimating phytoplankton production. *Marine Ecology Progress Series* 139, 289-299.

Schleiter, I. M., Borchardt, D., Wagner, R., Dapper, T., Schmidt, K., Schmidt, H. and Schwefel HP (1995). *Evolution and Optimum Seeking*. John Wiley and Sons, New York.

Schofield, N. J. and Davies, P. E. (1996). Measuring the health of our rivers. *Water* 23, 39-43.

Schwefel, H.P. (1995) Evolution and Optimum Seeking. John Wiley and Sons, New York.

Simpson, S., Norris, R., Barmuta, L. and Blackman, P. (1997*). Australian River Assessment System- National River Health Program Predictive Model Manual (first draft).* Canberra, ACT, CRC Freshwater Ecology, University of Canberra.

Smith, M. J., Kay, W. R., Edward, D. H. D., Papas, P. J., Richardson, K. S. J. and Simpson, J. S. (1999). AusRivAs: using macroinvertevrates to assess ecological conditions of rivers in Western Australia. *Freshwater Biology* 41, 269-282.

Ter Braak, C. J. F. (1988). Partial canonical correspondence analysis. In *Classification and related methods of data analysis*. pp. 551-558. (Amsterdam: North-Holland.)

Ter Braak, C. J. F., and Verdonschot, P. F. M. (1995). Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences* 57/3, 255-285.

Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R. and Cushing, C.E. (1980). The river continuum concept. *Canadian Journal of Aquatic Science*, 130-137.

Vesanto, J., Alhoniemi, E. (2000). Clustering of the Self-Organizing Map, *IEEE Transactions on Neural Networks*.

Vesanto, J, Himberg, J, Alhoniemi, E, Parhankangas, J. (2000). SOM Toolbox for Matlab 5. Technical Report. Helsinki University of Technology, Neural

Networks Research Centre.
http://www.cis.hut.fi/projects/somtoolbox/documentation/

Walley, W. J. and Fontama, V. N. (1998). 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.

Ward J.V. and Stanford, J. A. (1983). The intermediate disturbance hypothesis: An expolanation for biotic diversity patterns in lotic ecosystems. *Dynamics of lotic ecosystems*. T.D.Fontaine and S. M. Bartell. Michigan, Ann Arbor Science.

Wen, C. G. and Lee, C. S. (1998). A neural network approach to multi objective optimisation for water quality management in a river basin. *Water Resources Research* 34(3), 427-436.

Werner, H. (1999). Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. *Ecological Modelling* 120, 271-286.

William J. Walley and Mark A. O'Connor. (2001). Unsupervised pattern recognition for the interpretation of ecological data. *Ecological Modelling* 146, 219 – 230.

Williams, W.D., Taaffe, R.G. and Boulton, A.J. (1991). Longitudinal distribution of macroinvertebrates in two rivers subject to salinization. *Hydrobiologia* 210,151 –160.

Williams, D. D., Williams, N.E. and Yong Cao (2000). Road salt contamination of groundwater in a major metropolitan area and development of a biological index to monitor its impact. *Water Research* 34(1), 127 –138.

Williams, W.D., Boulton, A.J., Taaffe, R.G.(1990). Salinity as determinant of salt lake fauna: a question of scale. *Hydrobiologia* 197, 257-266.

Whigham, P.A. (2000). Induction of a marsupial density model using genetic programming and spacial relationships. *Ecological Modelling*, 131, 2-3, 299-317.

Wright, J.F. (1995). Development and use of a system for prediction the macroinvertebrate fauna in flowing waters. *Freshwater Biology* 20: 181-197.

Wright, J.F., Furse, M.T., Clarke, R. T. and Moss, D. (1991). Testing and further development of RIVPACS, *IFE Interrim report to the National River Authority*: 1-141.

Yabukana, K., Hosomi, M. and Mirakami, A. (1997). Novel application of artificial neural network model formulated to predict algal bloom. *Water Science and Technology* 36, 89-97.

Ysebaert, T.P, Herma, M.J., Meire, P., Craeymeersch, J., Verbeek, H, Heip, C.H.R. (2003). Large-scale spatial patterns in estuaries: estuarine macrobenthic communities in the Schelde estuary, NW Europe. *Estuarine, Coastal and Shelf Science* 57, 335-355.