# The application of artificial neural networks to the interpretation and classification of freshwater benthic invertebrate communities

Brendan Michael Ruck

Doctor of Philosophy

The University of Aston in Birmingham

December 1995

The University of Aston in Birmingham

# The application of artificial neural networks to the interpretation and classification of freshwater benthic invertebrate communities

Brendan Michael Ruck

Submitted for the degree of Doctor of Philosophy
December 1995

## Summary

This thesis presents a thorough and principled investigation into the application of artificial neural networks to the biological monitoring of freshwater. It contains original ideas on the classification and interpretation of benthic macroinvertebrates, and aims to demonstrate their superiority over the biotic systems currently used in the UK to report river water quality.

The conceptual basis of a new biological classification system is described, and a full review and analysis of a number of river data sets is presented. The biological classification is compared to the common biotic systems using data from the Upper Trent catchment. This data contained 292 expertly classified invertebrate samples identified to mixed taxonomic levels.

The neural network experimental work concentrates on the classification of the invertebrate samples into biological class, where only a subset of the sample is used to form the classification. Other experimentation is conducted into the identification of novel input samples, the classification of samples from different biotopes and the use of prior information in the neural network models. The biological classification is shown to provide an intuitive interpretation of a graphical representation, generated without reference to the class labels, of the Upper Trent data.

The selection of key indicator taxa is considered using three different approaches; one novel, one from information theory and one from classical statistical methods. Good indicators of quality class based on these analyses are found to be in good agreement with those chosen by a domain expert. The change in information associated with different levels of identification and enumeration of taxa is quantified.

The feasibility of using neural network classifiers and predictors to develop numeric criteria for the biological assessment of sediment contamination in the Great Lakes is also investigated.

Key words: `Freshwater biological monitoring, indicator taxa, multilayer perceptrons, river water quality, sediment toxicity.`

# Acknowledgements

My principal thanks must go to Bill Walley, my supervisor, for his guidance, encouragement, support and friendship throughout the duration of my studentship. It has been a pleasure to work with Bill, and I hope that some day I can come close to matching Bill's perseverance and dedication.

My other set of principal thanks go to Bert Hawkes, without whom this project would not have been possible. Throughout this dissertation it is Bert who is described, rather impersonally, as the Expert, and it is he who added credibility and extra leverage in the promotion of our ideas. Bert is a true expert and I feel particularly privileged to have had access to his unique 'knowledge-base'.

From the National Rivers Authority, I would like to thank Shelley Howard and Brian Walters, who originally provided a 'real' data set for our use, and all the other biologists who have contributed to our work. Additionally I would like to thank Mike Furse of the IFE for the provision of the current database of Maitland codes.

I was exceedingly fortunate to be invited to spend three months working in Canada at the NWRI, and for this opportunity I must thank both Trefor Reynoldson and Kristin Day. My visit to the NWRI was extremely enjoyable and productive, and thanks are also due to Sherri, Craig and the crew and scientists on board the CSS Limnos. I would also like to thank Saso Džeroski, at the Institut Jožef Stefan, for his interest in our work and the application of various machine learning paradigms to our data.

On a more personal note, I would like to express my appreciation of the Systems and Remote Sensing Group at Aston who put up me with slowing down their computers for the last three years. Special thanks to Jane, Naomi, John Elgy, Mike, Rob and John White. I also wish to thank the Neural Networks Research Group (aka Spasm?) at Aston for providing an excellent list of both 'in house' and invited speakers at their weekly meetings. Additionally, I would also like to express my gratitude to Guy Houlsby and Lionel Tarassenko, both of the University of Oxford, for their support while I finished off the final drafts of this dissertation (and also for providing me with gainful employment).

For making my time at Aston both extremely enjoyable and memorable I would like to say a big 'THANK YOU' to my friends, numerous flatmates and the members, especially Pete and Gavin, of the various bands that I have played in while at Aston. Last, and most importantly of all, I wish to thank my mother and Sheila for their constant love and support.

# Contents


**1 Introduction** **16**
- 1.1 Background . . . 16
- 1.2 The Scope of this Dissertation . . . 18
- 1.3 Organisation of the Dissertation . . . 19

**2 Freshwater Biomonitoring** **21**
- 2.1 Introduction . . . 21
  - 2.1.1 Benthic Macroinvertebrates . . . 22
- 2.2 Traditional Methods of Biomonitoring . . . 23
  - 2.2.1 Saprobic Approach . . . 24
  - 2.2.2 Diversity Indices . . . 25
  - 2.2.3 Biotic Methods . . . 27
  - 2.2.4 Other Indices . . . 33
- 2.3 Recent Developments in Biomonitoring . . . 34
  - 2.3.1 RIVPACS . . . 34
  - 2.3.2 Ecological Quality Index (EQI) . . . 36
  - 2.3.3 AI Methods . . . 38
  - 2.3.4 Other Techniques . . . 41
- 2.4 Summary of Biological Monitoring . . . 41

**3 Artificial Neural Networks** **43**
- 3.1 Introduction . . . 43
- 3.2 The Multilayer Perceptron . . . 45
- 3.3 The Back-Propagation Algorithm . . . 48
- 3.4 Generalisation . . . 50
- 3.5 Model Selection . . . 52
- 3.6 Bayesian Perspective . . . 54
- 3.7 The Future . . . 54
- 3.8 Summary . . . 55

**4 Analysis of River Data** **56**
- 4.1 Introduction . . . 56
  - 4.1.1 Chronology of Project Data . . . 57
- 4.2 Biological Classification and Knowledge Elicitation . . . 58
  - 4.2.1 Introduction . . . 58
  - 4.2.2 Biological Classification . . . 58

4.2.3 Direct Elicitation . . . . . 62
4.2.4 Indirect Elicitation . . . . . 66
4.2.5 Sources of Information Loss . . . . . 67
4.3 Severn-Trent Data Set . . . . . 71
4.3.1 Base Data . . . . . 71
4.3.2 Construction of 292 Sample Database . . . . . 73
4.3.3 Confirmation Tests . . . . . 78
4.3.4 Comparison of Biological Classification with the BMWP Score, ASPT, TBI and Number of Taxa . . . . . 80
4.3.5 Frequency of the Original BERT Taxa . . . . . 85
4.3.6 Recapitulation . . . . . 88
4.4 Other River Data . . . . . 89
4.4.1 Yorkshire Water Authority Data . . . . . 89
4.4.2 Synthetic Data . . . . . 89
4.4.2.1 Sample Generation from the Conditional Probabilities . . . . . 90
4.4.2.2 Expert Classification of Synthetic Data . . . . . 95
4.4.2.3 Discussion . . . . . 96
4.4.3 National Data . . . . . 97
4.4.3.1 Comparison of Scores and NWC Classification . 99
4.4.3.2 Distribution of Taxa by Region . . . . . 105
4.5 Summary . . . . . 108

**5 Neural Network Experiments** **111**
5.1 Introduction . . . . . 111
5.2 Preliminaries . . . . . 111
5.2.1 Implementation . . . . . 111
5.2.2 Experimental Methods . . . . . 113
5.2.2.1 Training, Validation and Testing . . . . . 113
5.2.2.2 Cross Validation . . . . . 114
5.2.3 Minimisation . . . . . 116
5.3 Direct Interpretation . . . . . 118
5.3.1 Overview . . . . . 118
5.3.2 Hidden Units . . . . . 118
5.3.3 Regularisation . . . . . 120
5.3.4 Data Encoding . . . . . 121
5.3.4.1 Procedure . . . . . 122
5.3.4.2 Results . . . . . 123
5.3.4.3 Using Additional Information . . . . . 123
5.3.4.4 Discussion . . . . . 124
5.3.5 Combination of Models . . . . . 125
5.3.5.1 Overview . . . . . 125

5.3.5.2 Balancing of Data Sets . . . 128
5.3.5.3 Procedure . . . 128
5.3.5.4 Results . . . 128
5.3.5.5 Discussion . . . 128
5.3.6 Discussion: Direct Interpretation . . . 130
5.4 Classification within Different Biotopes . . . 130
5.4.1 Introduction . . . 130
5.4.2 Procedure . . . 131
5.4.3 Results . . . 133
5.4.4 Discussion . . . 134
5.5 Detection of Novel Samples . . . 136
5.5.1 Overview . . . 136
5.5.2 Procedure . . . 137
5.5.3 Results . . . 138
5.5.4 Discussion . . . 140
5.6 Encoding Prior Information . . . 142
5.6.1 Overview . . . 142
5.6.2 Procedure . . . 142
5.6.3 Results . . . 143
5.6.4 Discussion . . . 144
5.7 Self-Organising Maps . . . 144
5.7.1 Introduction . . . 144
5.7.2 Mapping the Severn-Trent Data . . . 145
5.7.3 Discussion . . . 146
5.8 Comparison to the BERT System . . . 147
5.9 Summary . . . 148

**6 Selection of Key Indicator Taxa** **150**
6.1 Introduction . . . 150
6.2 Indicator Taxa . . . 151
6.2.1 Biological and Ecological . . . 151
6.2.2 Variables for Use in Computer Models . . . 152
6.3 Determination of Indicator Taxa . . . 153
6.3.1 The RMS-D Method . . . 154
6.3.1.1 Philosophy . . . 154
6.3.1.2 Derivation . . . 154
6.3.2 An Information Theoretic Approach . . . 159
6.3.2.1 Review . . . 159
6.3.2.2 Mutual Information of Class and Attribute . . . 159
6.3.3 Classical Methods . . . 160
6.4 Comparison of Indicator Expressions . . . 161
6.4.1 Procedure . . . 161

6.4.2 Results . . . 161
6.4.3 Discussion . . . 164
6.4.4 Absent/Present and Information Loss . . . 167
6.5 Model Performance Using Increasing Numbers of Indicators . . 171
6.5.1 Procedure . . . 171
6.5.2 Results . . . 171
6.5.3 Discussion . . . 172
6.6 Input Encoding Using RMS-D Values . . . 173
6.6.1 Procedure . . . 173
6.6.2 Results . . . 174
6.6.3 Discussion . . . 174
6.7 Summary . . . 177

**7 Great Lakes** **180**
7.1 Introduction . . . 180
7.2 The Development of Sediment Guidelines . . . 181
7.2.1 Description of Study . . . 181
7.2.2 Reference Sites . . . 183
7.2.3 Field Procedures . . . 183
7.2.4 Data Analysis . . . 183
7.2.5 Numerical Guidelines . . . 187
7.3 Reference Site Data . . . 188
7.3.1 Ordination . . . 188
7.3.2 Ordination of Environmental Variables . . . 189
7.3.3 Ordination of Community Structure . . . 194
7.3.4 Ordination of Bioassay Data . . . 196
7.4 Classification of Biological Groups . . . 201
7.4.1 Objectives . . . 201
7.4.2 Preliminary Experiments . . . 201
7.4.2.1 Procedure . . . 201
7.4.2.2 Results . . . 202
7.4.3 Analysis of Results of Preliminary Experiments . . . 202
7.4.3.1 Procedure . . . 202
7.4.3.2 Results . . . 204
7.4.4 Committees of Networks . . . 207
7.4.4.1 Procedure . . . 207
7.4.4.2 Results . . . 208
7.4.5 Discriminant Analysis . . . 208
7.4.5.1 Procedure . . . 208
7.4.5.2 Results . . . 209
7.4.6 Discussion . . . 210
7.5 Classification of Toxicity Groups . . . 210

7.5.1 Preliminary Experiments . . . 210
7.5.1.1 Procedure . . . 210
7.5.1.2 Results . . . 211
7.5.2 Committees of Networks . . . 212
7.5.2.1 Procedure . . . 212
7.5.2.2 Results . . . 212
7.5.3 Discriminant Analysis for Toxicity Tests . . . 212
7.5.3.1 Procedure . . . 212
7.5.3.2 Results . . . 214
7.5.4 Discussion . . . 214
7.5.5 Conclusion . . . 215
7.6 Prediction of Ordination Vectors . . . 216
7.6.1 Motivation . . . 216
7.6.2 Preliminary Experiments . . . 216
7.6.2.1 Procedure . . . 216
7.6.2.2 Results . . . 216
7.6.2.3 Discussion . . . 217
7.6.3 Committees of Networks . . . 221
7.6.3.1 Procedure . . . 221
7.6.3.2 Results . . . 221
7.6.4 Discussion . . . 224
7.7 Prediction of the Abundance of Taxa . . . 225
7.7.1 Procedure . . . 225
7.7.2 Results . . . 226
7.7.3 Discussion . . . 226
7.7.4 Conclusion . . . 229
7.8 Summary . . . 229

**8 Discussion and Conclusions** **231**
8.1 Introduction . . . 231
8.2 Discussion . . . 231
8.2.1 Contributions and Summary . . . 231
8.2.2 Practical Application and Implementation . . . 234
8.2.3 Future Research . . . 235
8.3 Conclusions . . . 238

**References** **240**

**Appendix** **255**
A1: Taxa Recorded in the Severn-Trent Database . . . 255
A2: Taxa Recorded in the National NRA Database . . . 258
A3: Great Lakes Benthic Community Structure Study: Species List . 260

# List of Figures


2.1 Water quality and non-water quality determinands of benthic community in rivers . . . . . 24
2.2 Schematic illustration of the EQI system . . . . . 38
2.3 Probabilistic interpretation of species response . . . . . 40

3.1 Schematic illustration of a single perceptron unit . . . . . 46
3.2 A typical multilayer perceptron (MLP) network . . . . . 46
3.3 Division of data for parameter and model selection . . . . . 53

4.1 Classification of river water quality into classes based on organic pollution . . . . . 60
4.2 Comparisons of direct and indirect elicitations of biological class with the Saprobic valencies for *Gammarus pulex* and *Asellus aquaticus* . . . . . 61
4.3 Comparisons of direct elicitation of biological class with the Saprobic valencies for *Baetis rhodani* . . . . . 61
4.4 Histograms of the probability of finding a taxon present in a given water quality class . . . . . 64
4.5 Typical elements of a monitoring programme . . . . . 68
4.6 Schematic illustration of riffle and pool diversity . . . . . 70
4.7 A typical sample from the Severn-Trent database . . . . . 74
4.8 Map showing the distribution of the 292 sample database taken from the Upper Trent Catchment . . . . . 77
4.9 Histograms showing the frequency of the biological water quality classes in the Severn-Trent database . . . . . 78
4.10 Summary of BMWP Score, ASPT, TBI and 'Number of Taxa' in terms of the five biological classes . . . . . 83
4.11 Summary of BMWP Score, ASPT, TBI and 'Number of Taxa' in terms of the thirteen biological classes . . . . . 84
4.12 A plot of BMWP Score against ASPT showing biological quality class . . . . . 85
4.13 Mean and standard deviation of union value . . . . . 88
4.14 Sampling of synthetic distributions in a B1b+ class . . . . . 91
4.15 Summary of BMWP Score and ASPT for synthetic data . . . . 94
4.16 Distribution of BMWP Score within each NRA region . . . . . 101
4.17 Distribution of ASPT within each NRA region . . . . . 103

4.18 Percentage occurrence of Gammaridae, Asellidae and Baetidae within the 10 NRA regions . . . . . 106
4.19 Percentage occurrence of Heptageniidae, Leuctridae, Nemouridae and Perlodidae within the 10 NRA regions . . . . . 107

5.1 MSE for training, validation and test data . . . . . 114
5.2 Error rate for training, validation and test data . . . . . 115
5.3 MSE plotted against time for three minimisation methods . . . 118
5.4 Number of weight updates plotted against time for three minimisation algorithms . . . . . 119
5.5 Minimum validation MSE plotted against number of hidden units for 16 PCA Severn-Trent data . . . . . 120
5.6 MSE plotted for a selection of regularisers . . . . . 121
5.7 Schematic illustration of mixtures of experts network for riffle and pool samples . . . . . 132
5.8 Schematic illustration of novelty detection . . . . . 137
5.9 MSE plotted against log likelihood for riffle and pool test data . 139
5.10 Two Self-Organising Maps from the Severn-Trent data . . . . . 145

6.1 Graphical representation of indicator value . . . . . 157
6.2 Performance of Linear, Quadratic and MLP models using different numbers of indicator taxa . . . . . 172
6.3 The effect of scaling the input encoding using indicator values . 175
6.4 Input values for the best 10 indicator taxa . . . . . 178

7.1 Map showing the reference site locations in the Great Lakes . . 184
7.2 Graphical relationship between experiments . . . . . 186
7.3 Location of environmental groups in ordination space . . . . . 192
7.4 Location of community structure groups in ordination space . . 195
7.5 Location of bioassay groups in ordination space . . . . . 200
7.6 Dendrogram of environmental ordination showing misclassified sites . . . . . 205
7.7 Regression of neural network predictions against vector scores from ordination of community structure . . . . . 223
7.8 Plot of Sphaeriidae abundance against sample depth . . . . . 228
7.9 Plot of Haustoriidae abundance against sample depth . . . . . 228

8.1 Schematic illustration of an example causal belief network for biological monitoring . . . . . 237

# List of Tables


2.1 The Extended and Trent Biotic Indices . . . . . 29
2.2 The BMWP Score System . . . . . 32
2.3 Proposed EQI bandwidths . . . . . 37

3.1 Frequency of published neural network papers since 1986 . . . . 45

4.1 The forty-one taxa used in the BERT system . . . . . 63
4.2 Sample information available in Severn-Trent database . . . . . 72
4.3 Description of abundance codes in the Severn-Trent database . . 72
4.4 Classification of 205 sample Severn-Trent database . . . . . 75
4.5 Classification of 292 sample Severn-Trent database . . . . . 76
4.6 Results of confirmation tests . . . . . 79
4.7 Comparison of the frequency of BERT taxa with a modified set of taxa . . . . . 86
4.8 Conditional probabilities for *Gammarus pulex* for generation of synthetic data . . . . . 91
4.9 Occurrence of taxon in pools and riffles . . . . . 92
4.10 Hydraulic characteristics of riffle and pool biotopes . . . . . 93
4.11 Classification of artificial samples . . . . . 95
4.12 Summary of the National NRA database . . . . . 98
4.13 River quality in 1990 by NRA region . . . . . 99
4.14 Summary of river invertebrate data . . . . . 108

5.1 Input data for pre-processing experiments . . . . . 123
5.2 Effects of different input and output encoding . . . . . 124
5.3 Confusion matrix for MLP trained with invertebrate sample and number of taxa as input . . . . . 124
5.4 Classification rates for committees of networks . . . . . 129
5.5 Interference effects for a MLP trained on the combined synthetic riffle and pool data . . . . . 133
5.6 Classification of synthetic riffle and pool data where gating network had invertebrate sample as input . . . . . 134
5.7 Classification of synthetic riffle and pool data where gating network had physical variables as input . . . . . 135
5.8 Confusion matrices for riffle and pool test data for MLP trained only with riffle data. . . . . 140

5.9 Confusion matrix for MLP training with Severn-Trent and synthetic data . . . . . 143

6.1 Conditional probabilities, $P(e_{ik}|H_j)$ for *Asellus aquaticus* and *Gammarus pulex* . . . . . 155
6.2 Marginal probabilities for *Asellus aquaticus* and *Gammarus pulex* 156
6.3 Conditional probabilities, $P(H_j|e_{ik})$, for *Asellus aquaticus* and *Gammarus pulex* . . . . . 156
6.4 Indicator values for *Asellus aquaticus* and *Gammarus pulex* . . . 157
6.5 Comparison of two probabilistic indicator indices . . . . . 158
6.6 Ranking of taxa in terms of three indicator indices: RMS-D Probabilities, Mutual Information and Stepwise Regression . . . 162
6.7 Comparison of abundance and absent/present data . . . . . 168
6.8 Classification rates of linear, quadratic and MLP models using different totals of indicator taxa . . . . . 176
6.9 Confusion matrix for network with lowest error rate trained using 15 indicators and probability values as input . . . . . 176
6.10 Error per class for the network . . . . . 176

7.1 Summary of measured environmental variables . . . . . 190
7.2 Key to data sets used for training neural net models . . . . . 191
7.3 Geographic distribution of environmental groups . . . . . 192
7.4 Mean and standard deviation of environmental variables for environmental groups . . . . . 193
7.5 Geographic distribution of community structure groups . . . . . 196
7.6 Mean and standard deviation of environmental variables for community structure groups . . . . . 197
7.7 Species which occur in at least 50% of sites . . . . . 198
7.8 Confusion matrix for community structure and environmental groups . . . . . 199
7.9 Taxa and endpoints used in toxicity tests . . . . . 199
7.10 Mean and standard deviation of environmental variables for bioassay groups . . . . . 200
7.11 Classification to community structure group . . . . . 203
7.12 Rank classification order . . . . . 204
7.13 Sites misclassified more than 5 times . . . . . 207
7.14 Classification to community structure group using committees . 208
7.15 Classification of community structure groups using discriminant analysis . . . . . 209
7.16 Classification to bioassay group . . . . . 211
7.17 Ranking and classification of committees . . . . . 213
7.18 Confusion matrix for toxicity groups . . . . . 213

7.19 Classification rate for toxicity test using discriminant analysis . 214
7.20 Prediction of first vector ordination scores . . . . . 218
7.21 Prediction of second vector ordination scores . . . . . 219
7.22 Prediction of third vector ordination scores . . . . . 220
7.23 Prediction of vector components . . . . . 222
7.24 Correlations between actual and predicted ordination vectors for each biological group . . . . . 222
7.25 Summary and composition of five most common families . . . . 225
7.26 Correlation of model's predictions against target abundances . . 227

# Glossary of Terms

As this dissertation covers work from two disparate areas, so this short glossary is provided for some of the more common biological and neural network terms.

ASPT
The Average Score Per Taxa is derived by dividing the BMWP score (ibid.) by the number of scoring taxa (ibid.). It has a more monotonic response, with respect to organic pollution, than the BMWP score.

AUTECOLOGY
Contraction of 'auto'-'ecology', refers to the specific ecology of individual taxa (ibid.).

BENTHIC
Pertaining to the bed of a river or lake, etc.

BERT SYSTEM
An expert system, based on Bayesian inference, which classifies river water quality from a set of 41 invertebrate taxa. The acronym stands for Benthic Ecology Response Translator, but was named after the expert from whom the knowledge base was elicited, namely Bert (H.A.) Hawkes.

BIOLOGICAL CLASSIFICATION
A classification based on the status of the biological communities.

BMWP SCORE
The Biological Monitoring Working Party score. A simple system of biological assessment, commonly used in the UK, based on the presence/absence of families of invertebrate taxa.

CHEMICAL CLASSIFICATION
A method of classifying water quality based on measurements of particular chemicals in the water.

DETERMINAND
A feature that can be described numerically. Usually a range of determinands are considered to characterise water quality.

GENERALISATION
The ability to correctly classify or predict previously unseen (i.e. new) data.

MACROINVERTEBRATE
Invertebrate animals which are large enough to be retained in a net. Typical benthic macroinvertebrate fauna includes worms, snails, leeches, fly nymphs and crustacea.

NWC CLASS
A system for reporting and classifying river water quality based largely on chemical determinands (ibid.). Devised by the National Water Council.

MLP
Multilayer perceptron, a neural network model, commonly used for classification and prediction problems.

NEURAL NETWORK
Networks of neurons which make up the brain (Biological). A flexible non-linear mathematical model (inspired by the structure and function of the brain). (Computational)

OVER-FITTING
Occurs when a model with too many degrees-of-freedom starts to fit the noise in the data. A model which suffers from over-fitting would be expected to have a poor level of generalisation.

RARE, ESTABLISHED and ABUNDANT
For the BERT system (ibid.) the presence of any taxa (ibid.) was described by one of three states, namely *rare*, *established* and *abundant*. The thresholds for each state were set individually for each taxon.

REGULARISTION
The act of controlling the 'smoothness' or the complexity of the neural network mapping. Ideally, regularisation should assist the network to generalise (ibid.), but there is, as always, a trade-off between too little and too much regularistion.

TAXON (*pl.* TAXA)
A given taxonomic group.

TBI
The Trent Biotic Index. A popular biotic index, which has been used as the basis for many other systems (especially in mainland Europe).

TRAINING, VALIDATION and TESTING DATA
Training data are used to estimate the model's parameters. Validation data are used to optimise generalisation and prevent the occurrence of over-fitting (ibid.), and the testing data are used to evaluate model performance.

WEIGHT DECAY
A commonly used method of regularisation (ibid.) for neural network models.

WEIGHTS
The adjustable parameters in a neural network model.

# Chapter 1

# Introduction

## 1.1 Background

Freshwater is a resource of immense importance and its quality is of considerable significance to its users. As a resource it needs to be managed, and it is in man's interest that it also be conserved [53]. Man's interference in the hydrologic system, both quantitatively and qualitatively, must be carefully managed so as to maintain the functional condition of the system, taking account of both water and environmental quality. It seems natural to use observations of the state of freshwater biotic communities as measures of such qualities, as these communities constitute a key and integral component of the riverine environment. As the issues concerning the environment become increasingly prominent, there seems to be a greater acceptance that information relating to the health of ecological systems has an important role to play in any environmental management programme. Since the biology is central to the environment it is perhaps surprising how much reporting and classifying of river water quality is based on chemical determinands, with little reference, if any, to the biology.

The system presently used in the UK to classify and report river water quality is the National Water Council (NWC) classification [114]. This classification is based primarily on three basic chemical measures, namely the biochemical oxygen demand and the concentrations of dissolved oxygen and ammonia. In addition, there is a secondary component based on the EIFAC (European Inland Fisheries Advisory Commission) standards for freshwater fish, which again are mainly chemically based. The problems associated with

the NWC and its application have been well documented [112], with many questions concerning compliance, seasonal adjustment and statistical interpretations of confidence intervals still unresolved [146, 29]. Notwithstanding these contentious issues, chemically based quality measures are useful for setting discharge consents or for identifying a specific pollutant, but for routine monitoring they leave much to be desired.

There are three main deficiencies of routine chemical methods of monitoring [122]. These are that:

*i.* there is an unknown number of chemical species to be detected,

*ii.* there are some chemicals for which the available analytical methods are not sensitive enough to detect the concentrations that cause concern, and

*iii.* chemical tests represent only a snap-shot of the long term conditions.

At present it is possible to detect 1% or 2% of the chemical species that could be found in a river [122], and the number of possible species is forever increasing. For example, it has been estimated that there may up to 50 000 different substances being poured into the Rhine [46], while other sources estimate over 100 000 chemicals in commercial use, increasing at a rate of about 1 000 new chemicals a year [30]. Severn-Trent NRA has a list of over 2 000 determinands which can be used in reporting water quality, but with due consideration of economic constraints, it would be prohibitive to test for all of these determinands. Even when testing for a specific chemical there is no guarantee that the test is sensitive enough to detect the pollutant at the level where it becomes toxic to the biota. Budgetary restraints require that there is a selective element to what is tested for, and biological indicators can provide useful information on which to base this selection. Finally, most of the chemical samples that are taken only represent the conditions at an instant in time, thus it is possible to miss sporadic or periodic discharges. Although continuous chemical monitoring stations are available, they only monitor for a limited range of determinands, and are comparatively costly to run and maintain, although the relative cost of the technology is continually decreasing.

All too frequently the chemical and biological information is treated separately and rarely integrated together. Ideally a fusion between chemical and

biological data should be considered, but this is either lacking in present systems or they do not utilise available technology as well as they could. Even when chemical and biological information have been integrated the results have been mixed, with them occasionally even appearing contradictory in nature. This was the case in the National Pollution Survey of 1970, but the problem was partly due to the shortcomings in the biologically-based index that was used. More recent proposals, for example NRA [113], have suggested using a combined system where the biological information can over-ride that provided by the chemistry. The problems of using and interpreting the biology, however, are very apparent. There is a large amount of subjectivity associated with the interpretation of a biological sample, personal bias on the part of experts which reflect their experience and training. There may even be some antagonism between zoologists and botanists, each defending their own patch.

The present biologically based systems used for monitoring can best be described as inadequate. Too much effort is being placed on the interpretation of scores and indices than on interpreting the source data, the invertebrate community (or any of the floral or faunal groups) living in the river. The focus of interpretation should be the biotic community, not the biotic or diversity indices. The score systems are of use, but more efficient methods of interpretation and classification must be developed, and this dissertation aims to improve upon these presently used systems.

## 1.2 The Scope of this Dissertation

This project is concerned with the development of new efficient and effective methods of interpreting and classifying biological data, so it is first necessary to consider which tools are most suitable for this task. Since the task, when performed by humans, requires considerable knowledge and expertise, it appears that the best tools presently available are ones which have recently been developed within the field of Artificial Intelligence (AI). Within AI there are several different paradigms, of which neural computing (or connectionism) is an expanding field of interest.

This dissertation explores possible applications of artificial neural networks to the field of the freshwater biomonitoring, and aims to partly redress the

perceived technological imbalance between chemical and biological monitoring. Two principal applications of neural networks are considered. The first is the classification of invertebrate samples into biological-based classes. The second is the classification of community structure and toxicity test group type in a sediment toxicity problem.

The project is just not restricted to the application of neural networks. For example, a key element that does not involve neural networks is the identification of indicator taxa. Here, three 'non-neural' selection methods for identifying indicator taxa are considered, with the results being of possible use for the future design of monitoring programmes.

## 1.3 Organisation of the Dissertation

This dissertation is divided into four parts.

The first part consists of two review sections; Chapter 2 is a discussion of the present status of biological monitoring of freshwater systems, with the emphasis on rivers, while Chapter 3 is a review of artificial neural networks. These chapters differ slightly in their emphasis, with the biological monitoring review being more critical, and highlighting the need for this research, while the neural network review is more of an overview of the technology, drawing out the pertinent work and areas of concern which have direct relevance to this study.

The second part consists of three chapters. Chapter 4 contains the details of the benthic invertebrate data sets, taken from rivers, that have been used. The difficulties in the present monitoring programs are highlighted, with the result that unnecessary uncertainty is being added to the data. The various classification systems presently used in the UK are compared and discussed, and a national database is used to present distributions of taxa which underlie the presently used score systems. The bulk of the neural network experimental work that has been completed is described in Chapter 5, and mainly uses supervised learning techniques in conjunction with the multilayer perceptron model as the standard tool of analysis. The main concern is the direct interpretation of invertebrate samples into a quality class based on organic pollution. The detection of novel input patterns is investigated, as well as a model which

handles the direct interpretation of both riffle and pool biotopes. Unsupervised learning models, Kohonen maps, are investigated as a graphical means of representing the biological classification. Chapter 6 is an investigation into the methods of selecting good indicator taxa for use in computer models. Three methods to identify good indicators of quality class are considered, and the changes in information with respect to different levels of identification and enumeration are quantified. The relationship between the number of taxa used to form the classification and model performance is investigated, and a novel data encoding format is presented.

The third part of the dissertation, Chapter 7, is an investigation into the use of neural networks within a system for the classification and prediction of sediment toxicity in the Laurentian Great Lakes. This work was carried out during a three month period spent at the National Water Research Institute, Burlington. The neural network models are used within the framework of the community structure group prediction, based on a similar procedure to that of Wright et al. [181].

The final part, Chapter 8, critically appraises the work presented in the dissertation and reiterates the dissertation's principal contributions. Possible directions for future are also discussed. Finally, the dissertation is concluded.

# Chapter 2

# Freshwater Biomonitoring

## 2.1 Introduction

The fundamental principle that forms the conceptual basis of freshwater biomonitoring is that a biological community structure undergoes changes induced by changes in its environment. Freshwater biomonitoring (i.e. using the information provided by biological communities to assess the functional condition of their environment) has a long history, but has played a somewhat secondary role to chemically based methods of monitoring. The difficulty in using biologically based information stems from the inherent complexity of ecosystems. As any biologist's interpretation is to some extent subjective, their conclusions have not been seen to provide either an effective monitoring system or the basis for quality standards at a national level.

This dissertation is particularly concerned with the methods of surveillance and monitoring of freshwater systems. Hawkes [53] and NERC [115] provide definitions of these terms:

> *surveillance* is the repeated and standardised measurement of a variable so that a temporal trend may be detected,
>
> *monitoring* is surveillance carried out to determine trends in relation to predetermined standards.

These definitions are sufficient for a single site over a period of time, but fail to emphasise the need for uniform spatial application when dealing with more than one site. In order to permit comparisons between different sites

the reporting procedure needs to be consistent, either being independent of biotope or else to have taken the biotope into consideration.

The original definitions apply to the work which is routinely undertaken by the NRA. This is where a particular stretch of river is intensively sampled because of an apparent pollutional problem. For this, it is enough to report qualitative differences between samples over a period of time (e.g. an improvement in the diversity of the fauna over the duration of the sampling programme, or an increase in the number of sensitive species present). The concern is not for absolute quality, but rather the change, hopefully improvement, over time. However, when absolute quality is required, for example in national monitoring programmes, the importance of spatial consistency becomes paramount. The need for a uniform system is to allow for the dissemination of the biological information in the form of classification systems, which provide valuable information for managers, and for the foundation of statutory objectives. This lack of spatial consistency is the major problem with the present biotic systems, since a particular value of a score always requires additional information before the score can be properly benchmarked.

Considering all the flora and fauna available for monitoring purposes, the benthic macroinvertebrates are the ones most frequently studied in the UK [55], and consequently they have been used for this study. This is not to say that other biological groups do not provide valuable evidence, some may positively strengthen a biologist's conclusion, while others may provide insights to new interpretations. Many, if not all, of the methods presented in this dissertation are applicable across the whole range of biological groups. It would be a good exercise to fuse data from a range of flora and fauna, but constraints on data and available expertise restricted this study to the benthic macroinvertebrates.

### 2.1.1 Benthic Macroinvertebrates

As stated in the previous section, benthic macroinvertebrates are the most common element of a river's biota to be used for monitoring. The main reasons for the popularity of macroinvertebrates are:

*i.* the comparative ease and relative low cost of sampling,

*ii.* their well defined taxonomy and the availability of good keys for identification,

*iii.* their relatively sedentary nature, unlike fish for example, making them representative of the site from where they are sampled,

*iv.* they cover a wide range of trophic and pollution tolerance levels,

*v.* they are found in almost all water courses of any quality (except for extremely toxic conditions).

Even so, a skilled biologist will take notice of biota other than the benthic invertebrates and use information from a wide variety of sources to identify any problems existing at a site.

The methodological and ecological considerations associated with biological surveillance have been well documented by Hawkes [49, 51, 53], Hellawell [55] and Metcalfe [103]. Figure 2.1 summarises the environmental effects acting upon the benthic community and also the interactions between these these effects. As the diagrams depicts, water quality determinands cover a number of dimensions, some of which are unnatural (i.e. definitely caused by human interference). Different current velocities, substratum types, hardness, etc., all have direct effects on the community structure, thus it is expected that different community structures are to be found in different environments. Thus the information on the habitat (biotope) should be considered in any interpretation or classification of the communities.

## 2.2 Traditional Methods of Biomonitoring

For the purpose of this dissertation the traditional methods of biomonitoring are considered to be those which are not computer intensive. This allows for a fairly clear delineation between the diversity and biotic indices, and the newer methods, such as RIVPACS and the knowledge-based systems approach of Walley et al. [170]. The most important methods for the monitoring of river water quality, mainly in relation to organic pollution, are discussed below.

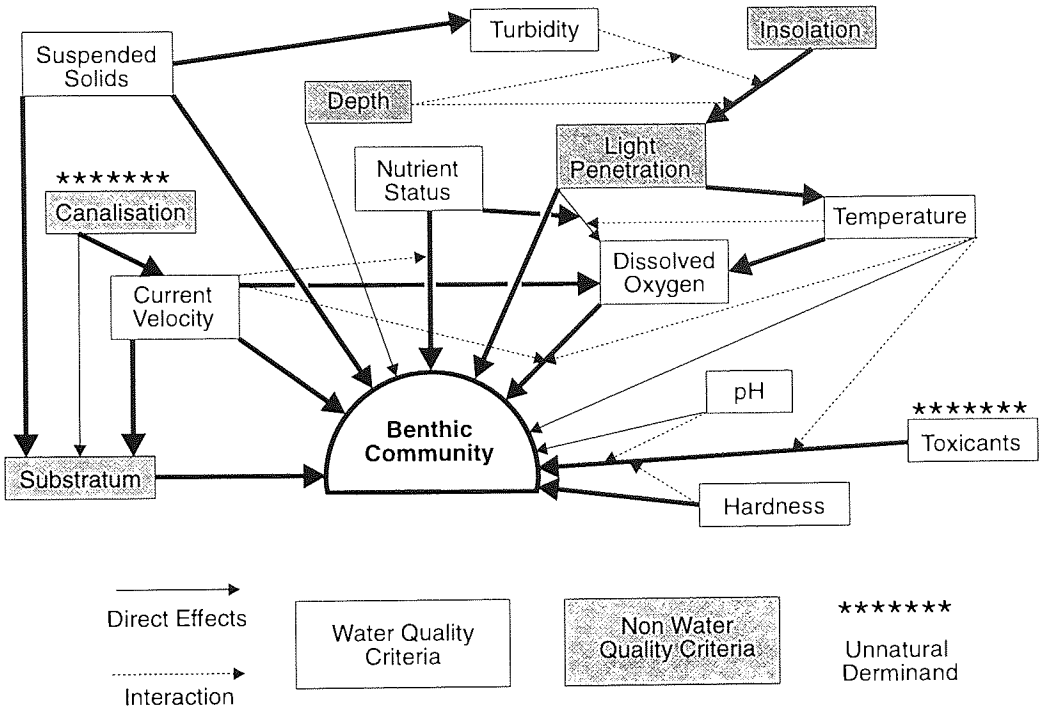



Figure 2.1: Water quality and non-water quality determinands of benthic community in rivers (after Hawkes [53]).

### 2.2.1 Saprobic Approach

The early Saprobic system, (saprobia is the dependence of an organism on decomposing organic substances as a nutrient source [127]), was developed by Kolkwitz and Marsson [85, 86]. They initially suggested four zones based on different saprobia, these being polysaprobic, $\alpha$-mesosaprobic, $\beta$-mesosaprobic and oligosaprobic. Polysaprobic waters are generally characterised by high levels of pollution, few community groups and active bacteriological decomposition; $\alpha$-mesosaprobic waters are distinctly polluted and have an oxygen content less than 50% saturated; $\beta$-mesosaprobic have mild to moderate pollution; and oligosaprobic are clean, healthy with a wide variety of plants and animals. These zones have since been increased to ten [103], but most ambient waters will normally be in one of the above four groups. The Saprobic Index is calculated as follows [123]:

$$S = \frac{\sum_i (s_i h_i)}{\sum_i h_i} \tag{2.1}$$

where:

$S$ is the Saprobic index for the site,

$s_i$ is the Saprobic valency for each indicator species ($i$),

$h_i$ is the abundance level of each species ($i$); rare(1), frequent(2), abundant(3).

The major criticisms of the Saprobic system are that [127]:

*i.* the taxonomy is controversial, especially for micro-organisms,

*ii.* the pollution tolerances of organisms are very subjective, generally based on ecological studies and not experimental work,

*iii.* the sampling regime is intensive, and

*iv.* the taxonomic list is not be applicable to many geographic locations.

The use of the Saprobic system in Britain is very limited with no NRA region using it regularly [71], although it is commonly used in Europe [53]. Unlike the systems discussed below the Saprobic system is readily applicable to all the flora and fauna of the river system, and thus indicator species include bacteria, algae, protozoans, benthic macroinvertebrates and fish. This must be considered an advantage over the following biotic systems.

### 2.2.2 Diversity Indices

These are mathematical expressions used to describe the response of a community to the quality of its environment. Generally communities under stress undergo a reduction in diversity and this results in changes in the diversity index. However, low diversity may not be indicative of polluted conditions because these may be caused by other factors, for example the physical condition of head streams. Three components of community structure are used in diversity indices; richness (number of species present), evenness (uniformity in the distribution of individuals among the species) and abundance (total number of organisms present).

Commonly used diversity indices include:

- Menhinick's Diversity Index [102]

$$D = S/\sqrt{N} \tag{2.2}$$

- Shannon-Weiner or Shannon-Weaver Index [155, 171],

$$D = \sum_{i=1}^{S} \frac{N_i}{N} \log_2 \frac{N_i}{N} \tag{2.3}$$

- Simpson's Index [156],

$$D = \sum_{i=1}^{S} \frac{N_i(N_i - 1)}{N(N-1)} \tag{2.4}$$

- Margelef's Community Diversity Index [97]

$$D = \frac{S-1}{\ln N} \tag{2.5}$$

where:

$S$ is the number of different species in the sample,

$N$ is the total number of individuals in the sample, and

$N_i$ is the number of individuals belonging to $i$th species.

Pinder et al. [129] used the above diversity indices in their study of a chalk stream. In all the above, the higher the value the greater the diversity and hence the better environmental quality. The main advantages of the diversity indices are [103]:

*i.* they are strictly quantitative, dimensionless and lend themselves to statistical analyses [22],

*ii.* they are generally independent of sample size [175],

*iii.* no assumptions are made about the relative tolerances of the individual species, unlike the subjective Saprobic system [129],

*iv.* they can be applied to measures of biomass which are less labour intensive than individual species counts [99].

Criticisms of diversity indices have been many, some of which are:

*i.* there is considerable variation in the index values depending of the equation used, sampling regime, biotope and the level of taxonomic identification,

*ii.* wide variations of values for unpolluted conditions have been cited [22]. Standards have been set for interpretation [175] but these scales are not universally applicable,

*iii.* they do not make use of species or autecological information, as the species are regarded as anonymous numbers [53],

*iv.* the response of the community to increasing pollution is not necessarily linear. For example, moderate organic pollution may lead to an increase in community diversity.

Other indices, sometimes referred to as coefficients of similarity, are occasionally used. These include Kothes' Species Deficit [50], Hellawell's Index [55], Jaccard's Coefficient [68], Sørenson's Quotient of Similarity [158] and Fager's Index of Affinity [33]. Mason [98] details applications of the above coefficients in experimental work, Hawkes [50] contains some example calculations, and Washington provides a comprehensive review [171].

### 2.2.3 Biotic Methods

Metcalfe [103] describes, using Tolkamps [164] definition, the biotic approach to biological assessment as:

> "one which combines diversity on the basis of certain taxonomic groups with the pollution indication of individual species of higher taxa or groups into a single index or score."

Essentially, biotic scores and indices are based on subjective assessments of the effects of organic pollution on the invertebrate fauna. The major difference between scores and indices are that scores are additive (i.e. they are calculated by summing individual indices for each taxon), while this is not the case for indices. The main driving force behind the development of biotic indices and

scores is the need for experienced biologists to present their findings in a form readily understandable to non-biologists. Although this is desirable, especially to managers within the water industry, a single index represents a substantial reduction in the total information available. Thus, many of the reported statistics require an explanatory note to make them meaningful.

### Trent Biotic Index

The Trent Biotic Index (TBI) was developed by Woodiwiss [179] for use by the Trent River Board and now forms a basis of many modern biotic indices and scores [127]. It originally had eleven quality classes (0-10) but was extended to cover a wider range of water qualities (0-15), the latter version being called the Extended Biotic Index (EBI), Table 2.1. The most sensitive taxon of a preselected set (which have been ordered in terms of recognised tolerance to organic pollution) is used as a benchmark for the index, with the final TBI value being determined from the total number of Trent groups present [24]. Clean streams were given the higher scores, thus with increasing levels of pollution the TBI would become smaller. The TBI and EBI are popular due to their practicality. The taxonomic requirement in determining the index is not prohibitive, only the key Trent groups are identified to species level, and the sample is only qualitatively assessed, the counting of individuals is not required. Cook [22] cites this lack of abundance measure as a drawback to the TBI as the accidental presence of an organism (e.g. due to drift) could lead to misclassification. It can also be criticised because of its brittleness, which stems from its reliance on only a few key indicators.

### Chandler's Score System

Originally designed for the Scottish upland rivers, Chandler's system [19] uses a cumulative points system as opposed to a look-up table, as in the TBI. Also, the method incorporates an abundance measure and an increased list of macroinvertebrates compared to that of the TBI. To calculate a sample score all the organisms present are identified to the appropriate level (species for some, genus for others) along with an abundance level (present, few, common, abundant or very abundant). The sample's score is calculated by summing

| Organisms in order of tendency to disappear as degree of pollution increases | Extended biotic index | Total number of groups present | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0-1 | 2-5 | 6-10 | 11-15 | 16-20 | 21-25 | 26-30 | 31-35 | 36-40 | 41-45 | |
| | Trent biotic index | Total number of groups present | | | | | | | | | | |
| | | 0-1 | 2-5 | 6-10 | 11-15 | 16+ | | | | | | |
| | | Biotic indices | | | | | | | | | | |
| Clean | Plecoptera nymphs present | More than one species | — | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| | | One species only | — | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| | Ephemeroptera nymphs present (excluding *Baetis rhodani*) | More than one species | — | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| | | One species only | — | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| | Trichoptera larvae or *Baetis rhodani* present | More than one species | — | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| | | One species only | 4 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | *Gammarus* present | All the above species | | | | | | | | | | |
| | | absent | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | *Asellus* present | All the above species | | | | | | | | | | |
| | | absent | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| | Tubificid worms and/or red Chironomid larvae present | All the above species | | | | | | | | | | |
| | | absent | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Polluted | All above species absent | Some organisms such as *Eristalis tenax* not requiring dissolved oxygen may be present | 0 | 1 | 2 | — | — | — | — | — | — | — |

Table 2.1: Extended and Trent biotic indices (after Woodiwiss [179]; Personne & De Pauw [127]).

the appropriate points, which are read off from a table, for all of the samples's taxa. An interesting feature of the system is that for increasing abundance the points allotted to sensitive groups increase, while for intolerant groups the points decrease with greater abundance. Metcalfe [103] refers to some possible criticisms of the method, these include its complexity, its rigorous taxonomy plus enumeration, its inconsistent level of taxonomic identification, that it can only be applied to upland streams, that it is geographically restricted due to the indicator species identified to genus and that the scores are subjective.

### Biological Monitoring Working Party (BMWP) Score

In the 1970 National River Pollution Survey a biological assessment was implemented to supplement the chemical classification. The biological system used was based on four quality classes A, B, C and D, each of which was associated with a group of organisms characteristic of that particular water quality. The results of the survey were disappointing in that the degree of correlation between biological and chemical classifications was smaller than expected. The main reason for the low correlation was that the biological system used was designed for fast flowing riffle sites. This resulted in the slow flowing lowland rivers being placed at a lower biological class than the comparative chemical class. This highlighted the importance of biotope in relation to the expected benthic communities at that site.

In the light of the 1970 Survey the Department of the Environment, Standing Technical Advisory Committee on Water Quality (STACWQ) set up a Biological Monitoring Working Party with the aim to recommend a national biological classification of river water quality. A 'score' system was recommended, which was based around a simplified Chandler System. The major differences being that all organisms would be identified to family level and the abundance level for each group would be disregarded. Each family is only counted once when calculating the index, no matter how many species it represents or how many of its species are present. The comparatively simple BMWP score resulted [20], Table 2.2. In effect it is not a biological classification of river water quality but a biologically-based score "of the biological condition of rivers" [53]. The different invertebrate groups are allotted points according to their intolerance of organic pollution. The score is calculated in the following

manner: list all the different families present in the sample, then find each family's particular score and sum them to arrive at the BMWP score for the sample. An important point affecting the application of the final score is that the system was designed to be applied temporally (surveillance of a particular site through time) but not spatially (for the comparison of rivers in different geographic locations). Strictly this means that rivers, and even adjacent sites on a water course, may yield different scores due to differences in biotope, even when there is no change in the water quality. Also the scores for each family are conservative, in that the most tolerant species of each family is used as the benchmark.

Prior to the computerisation of the invertebrate sample records, the BMWP score was the most commonly used index, in fact all 10 of the then water authorities used it routinely [71]. The main reason the BMWP was used in all regions was that it was nationally recognised and easy to use. Other scores seemed to be included in the monitoring programs just as an extra measure.

The key points to note about the BMWP score are that it implicitly assumes independence between taxa, all identification is to family level, all the families are assumed to be equally reliable as indicators and no measure of abundance is included [168]. Perhaps the most problematic aspect of the BMWP is its misuse. It was designed to allow for monitoring of one site over time, and the comparison of scores between rivers was to be discouraged [53]. However, this proviso is generally disregarded.

### ASPT

The 'Average Score Per Taxon' is a frequently used index and simply represents the number derived when the total score for a sample is divided by the number of scoring taxa [4]. It is applied mainly to the BMWP score, but was first applied to the Chandler score [5, 111]. Pinder & Farr [128] have studied the comparative performance of four diversity indices and three biotic indices (TBI, Chandler, BMWP and ASPT versions of the Chandler and BMWP). In the paper, the authors recommended the ASPT based on the BMWP score, because of its simple calculation, the limited degree of taxonomic expertise needed and that it is little affected by sample size. Armitage et al. [4] describe the performance of the ASPT and the BMWP for different locations, sampling

| Families | Score |
|---|---|
| Siphlonuridae Heptageniidae Leptophlebiidae Ephemerellidae Potamanthidae Ephemeridae<br>Taeniopterygidae Leuctridae Capniidae Perlodidae Perlidae Chloroperlidae<br>Aphelocheiridae<br>Phryganeidae Molannidae Beraeidae Odontoceridae<br>Leptoceridae Goeridae Lepidostomatidae Brachycentridae<br>Sericostomatidae | 10 |
| Astacidae<br>Lestidae Agriidae Gomphidae Cordulegasteridae Aeshnidae<br>Corduliidae Libellulidae<br>Psychomyiidae Philopotamidae | 8 |
| Caenidae<br>Nemouridae<br>Rhyacophilidae Polycentropidae Limnephilidae | 7 |
| Neritidae Viviparidae Ancylidae<br>Hydroptilidae<br>Unionidae<br>Corophiidae Gammaridae<br>Platycnemididae Coenagriidae | 6 |
| Mesoveliidae Hydrometridae Gerridae Nepidae Naucoridae<br>Notonectidae Pleidae Corixidae<br>Haliplidae Hygrobiidae Dytiscidae Gyrinidae<br>Hydrophilidae Clambidae Helodidae Dryopidae Elminthidae<br>Chrysomelidae Curculionidae<br>Hydropsychidae<br>Tipulidae Simuliidae<br>Planariidae Dendrocoelidae | 5 |
| Baetidae<br>Sialidae<br>Piscicolidae | 4 |
| Valvatidae Hydrobiidae Lymnaeidae Physidae Planorbidae<br>Sphaeriidae<br>Glossiphoniidae Hirudidae Erpobdellidae<br>Asellidae | 3 |
| Chironomidae | 2 |
| Oligochaeta (whole class) | 1 |

Table 2.2: The BMWP Score System

efforts and seasons. They found the ASPT to be less variable (which would be expected because it is an 'average') and could be more reliably predicted from physical and chemical data. The ASPT can be criticised because of the inherited characteristics of the BMWP score, and it should be noted that it also excludes 'absent' evidence [168].

#### LQI

Extence et al. [32] proposed a new index, called the Lincoln Quality Index (LQI), which used both the BMWP score and the ASPT. They argue that the reporting of the BMWP scores and ASPT figures are "somewhat cumbersome", and that the figures often need an interpretive comment. The LQI is designed as a single readily understood index. To derive the index all sampling sites must be classified as either habitat-rich riffles, or habitat-poor riffles/pools. After calculating the BMWP score and the ASPT a rating for each sample is obtained from the appropriate habitat table. The average of these ratings, the Overall Quality Rating, can then be converted into an equivalent LQI value. The LQI can also be related to NWC class or river use (River Quality Objectives), but the index has not become nationally recognised within the water industry. Again it can be criticised because of its reliance on the BMWP score.

### 2.2.4 Other Indices

Metcalfe [103] reviews the biological systems that are used in Europe and North America. In Europe the Saprobic Index is still popular, but other indices (e.g. the Indice Biologique de Qualité Générale, Indice Biologique Global and the Belgium Biotic Index) have been developed [25]. De Pauw et al. [23] list over a 100 different indices which have been developed for monitoring purposes, with the majority of these being based on one of the previously discussed indices, but with modified sampling strategies and adjusted taxonomic lists to suit the region of interest. The proliferation of indices indicates the increasing role of biological monitoring, but also highlights the subjectivity as workers feel the need to develop systems with which they are comfortable using.

## 2.3 Recent Developments in Biomonitoring

### 2.3.1 River InVertebrate Prediction and Classification System (RIVPACS)

An important new approach to biological surveillance using multivariate statistical techniques was developed by a team from the Freshwater Biological Association.[1] The major objectives of the programme, as outlined in Wright et al. [181], were:

*i.* the development of a biological classification of unpolluted running-water sites in Great Britain based on the macroinvertebrate fauna; and

*ii.* the prediction of the macroinvertebrate community at a site from its physical and chemical features.

The work was conducted in three phases, initially examining the possibility of predicting 'expected' communities from physical and chemical variables using data collected from 268 sites on 41 rivers. These sites were regarded as either being of 'good' or 'fairly good' quality [4]. Multiple linear regression equations were computed using BMWP score and ASPT as the dependent (target) variables and various physical and chemical parameters being the independent (predictor) ones. A standardised sampling regime was implemented, three minute samples taken three times a year, in Spring, Summer and Autumn. One conclusion from this work was that the predictive equations, based on the physical and chemical data, would enable theoretical ASPT's to be calculated. The BMWP score equations were less reliable due to the score being more sensitive to sampling effort and seasonal change. In Wright et al. [181] the same data was used to develop a classification of running-water sites based on all macroinvertebrate taxa, and a method of predicting community structure of a site from environmental data. The sites were ordinated using detrended correspondence analysis (DCA), and then classified into 16 groupings using a two-way indicator species analysis TWINSPAN [59].

[1]The RIVPACS project is now associated with the Institute of Freshwater Ecology (IFE), which was recently 'demerged' from the FBA.

TWINSPAN constructs a ordered two-way table between both samples and species, and also produces a key which enables further sites to be classified. Multiple discriminant analysis (MDA) was used to classify the site groupings from the environmental data. Using the full data set of 268 samples 76.1% of the dependent sites were correctly predicted, with a further 15.3% of sites being the second most probable. When the data was split into 228 training and 40 testing samples, the training data was predicted with an accuracy of 67.5%, while the testing data were predicted to 50% accuracy, with the second most probable group being correct in 25% of cases. In Furse et al. [41] the influence of season and taxonomic factors on the performance of the system are assessed. The predictive accuracy of the model was found to be higher when identification was taken to species level. Also, the model was found to perform better when the species lists from all three seasons were combined.

Field trials were conducted using 21 new unpolluted sites [109] using Wright et al.'s [181] model. The main impetus being to predict the species occurrence at the new sites, as opposed to trying to classify the sites. The fauna were predicted using sets of 28, 11 and 5 physical and/or chemical determinands. The authors felt that a major use for the prediction system would be the provision of a 'target' macroinvertebrate community to act as a standard for a given site when it is unpolluted. Where a site is currently polluted, the difference between the actual fauna observed and the target fauna gives a measure of the loss of biological quality as a result of the pollution. For phase 2 of the work the database was enlarged to cover 370 sites on 61 different river systems, including more small streams and lowland river sites. Phase 3 of the work saw the database expanded further to encompass 438 sites, with the RIVPACS software package being developed, which allowed for the prediction of fauna at one or all of the following taxonomic levels: species (qualitative); all families (log. categories of abundance); all families (qualitative) and the BMWP families (qualitative). Also, four sets of environmental variables were available, these being 11 or 5 physical and chemical variables and 11 or 5 physical variables only. Quoting the authors [180]:

> "RIVPACS offers site specific prediction based on environmental features and can therefore set a target from which any loss of biological quality due to environmental stress can be measured by the 'observed/expected' ratio."

At present, RIVPACS is being extensively tested nationally and undergoing continued development by the NRA [113].

To digress slightly, it seems that a potentially valuable piece of information from the RIVPACS models is not being utilised. The actual difference in the composition of the expected and observed family list could be exploited, and would provide real scope for interpretation and possibly provide a significant supplement to any classification system. For classification purposes, however, the best information source is the benthic community itself, thus the direct interpretation of this into water quality terms provides the best basis for a robust classification system.

The methods of Wright et al. represented a major change in direction of freshwater biomonitoring, increasing the reliance on numerical models. Their methodology has being frequently applied in other studies and is adopted in Chapter 7 of this dissertation for the analysis of benthic community structure in the Great Lakes.

### 2.3.2 Ecological Quality Index (EQI)

Within the 1989 Water Act powers were provided to allow the Secretaries of State for the Environment and for Wales to introduce new classification systems and to use them as the basis for new legislation. The NRA proposals for the new classification system are set out in their document Proposals for Statutory Water Quality Objectives [113]. The aim of the proposals is to expand the existing NWC classification schemes to all types of water on a consistent basis. The main elements of the new proposals are the introduction of Use Categories, the relevant EC Directives and a new General Classification Scheme. The Use Classes will form the main element of the regulatory framework.

An ecosystem use-related class is recommended, and would have an assessment of the biological component of the watercourse. The intended system for implementation is based upon RIVPACS, utilising the predictive aspect of the system to estimate the expected BMWP, ASPT and Number of Taxa (NOT) scores of the unpolluted site according to physiographic properties.

Comparing the estimated unpolluted scores with the actual (observed) score determined from sampling, a ratio can be developed to describe the

| General Ecosystem Class | Description | EQI (ASPT) | EQI (NOT) | EQI (BMWP) |
|---|---|---|---|---|
| 1 | Good | $\geq 0.89$ | $\geq 0.79$ | $\geq 0.75$ |
| 2 | Fair | 0.77-0.88 | 0.58-0.78 | 0.50-0.74 |
| 3 | Poor | $<0.77$ | $<0.58$ | $<0.50$ |

Table 2.3: Proposed EQI bandwidths [113].

status of the site. The ratio of observed to predicted scores is expressed as an Ecological Quality Index (EQI). The EQI would be used to define classes in a hierarchical style, and objectives and compliance assessed in terms of EQI banding. In Appendix 2 of the document possible values of the bands are given, and these are shown in Table 2.3. The General Ecosystem Class would be the median value of the three class assessments. The major theme running through the document is that the use of biological information should be used more objectively in making decisions on class assignment. The biological quality of a river gives a very different assessment to that which can be derived from chemical monitoring. The biological information indicates the living state of the river while the chemical criteria are more suited to discharges and pollution control requirements. Both kinds of information are important for the purposes of water quality management, but the new classification system will have strict rules governing the application of chemical criteria and it was suggested that a biological over-ride feature may be incorporated into the system, but this has been subsequently dropped.

The shortcomings of the EQI system can be demonstrated using Figure 2.2, which is based on the BMWP score for a typical riffle site. The $x$-axis represents increasing pollution while the $y$-axis gives the likely score at a typical riffle site. An EQI is calculated by taking the value of the observed BMWP score and dividing it by the site's expected BMWP score, which is generated by RIVPACS. The main problem with this is that the non-monotonic and non-linear nature of the relationship between the BMWP and quality. From Figure 2.2 only one sample out of the three, sample C, would have an EQI of less than unity. Thus, the EQI system may result in classifying some moder-

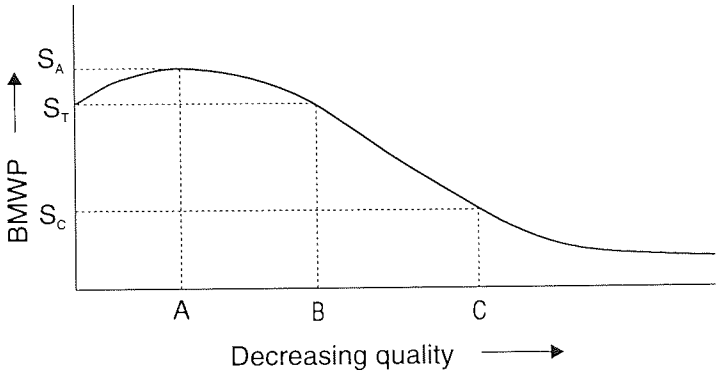



Figure 2.2: Schematic illustration of the EQI system. The response of BMWP score for a typical riffle site is shown, and the EQI for an observed sample A is the ratio $S_A/S_T$, where the expected 'clean' score is given by $S_T$.

ately polluted sites as being unpolluted. This behaviour would also be expected to occur for EQIs based on the BMWP or the number of taxa, but would be less of a problem for the ASPT based EQIs. Additionally, there is an implicit assumption that the EQI system applies across all biocoenoses, thus an EQI of, say, 0.8 is equivalent in both riffle and pool biotopes. A further criticism of the proposed EQI system is that it inherits the family level implementation of the BMWP system, and this implies the much of the available biological information will be lost or be unused in the classification system.

### 2.3.3 AI Methods

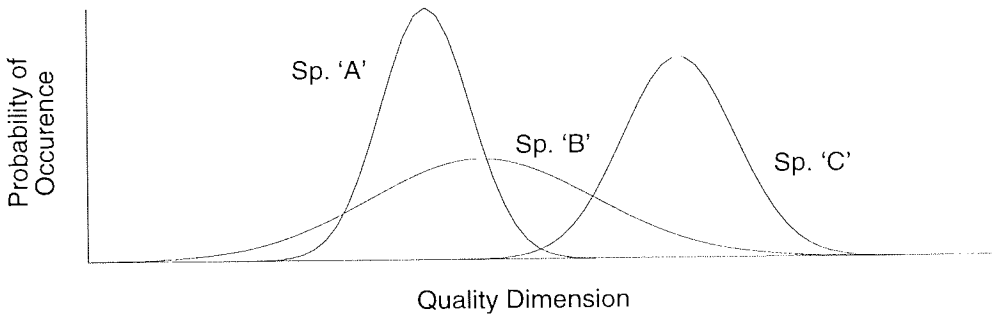
A novel approach to the interpretation and classification of invertebrates, based on probability theory, has been developed by Walley and a small team of researchers at the University of Aston [170, 167, 13, 168]. The relationship between the abundance and frequency of occurrence of a taxa and water quality lends itself to a Bayesian interpretation. For example, Hynes [65] depicts the changes in macroinvertebrates downstream of a discharge of organic effluent, showing the response of Tubificidae, *Chironomus* and *Asellus* as virtually Gaussian in nature. Treating these distributions as probabilities, the combination of the probabilities for either classification or interpretation is easily handled in a mathematically sound way within a Bayesian framework [26].

Figure 2.3 gives the fundamental principles to the probabilistic interpretation of river water quality. In Fig. 2.3a the response of each species is different for a given quality dimension, that response may by particularly specific (Sp. A or Sp. C) or relatively vague (Sp. B). In Fig. 2.3b the difference in the abundance of a taxon can be represented, again, by using different distributions. For the BERT system three levels of presence were used for each taxon, namely *rare*, *established* and *abundant* (see Table 4.1 for the exact definitions of these terms). If a taxon is present in abundance then this is more informative than if the taxon is established or rare. For taxa which occur with a very high absolute abundance (e.g. *Gammarus pulex* or *Asellus aquaticus*), it was postulated that a bimodal distribution for *rare* and *established* states may be the natural representation, indicating the taxon had been found on either side of its preferred quality (see Tables 6.1 and 6.3 for some justification of this). Finally, in Fig. 2.3c the evidence from different taxa can be combined to give a probability distribution of quality class given the taxa.

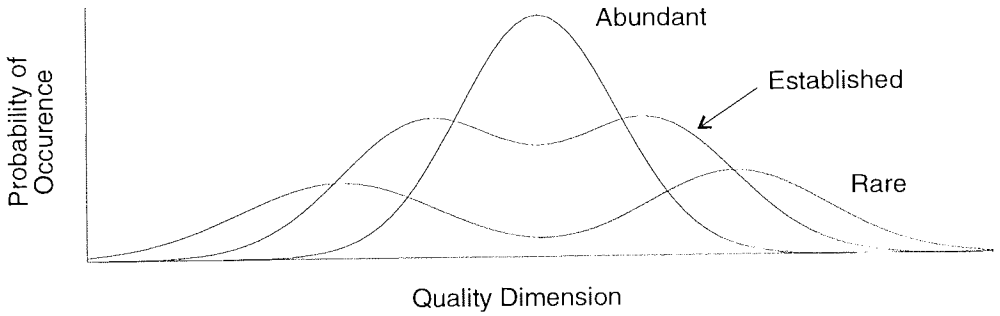
Using the conceptual basis described above, Walley et al. [170] have developed a knowledge-base system that uses a Bayesian inference model to classify water quality from the direct interpretation of benthic macroinvertebrate data. The model used forty-one indicator taxa selected by the domain expert, these same species being used in this project. Knowledge was elicited from the Expert in the form of histograms. From the histograms the conditional probabilities of four states of abundance (absent, rare, established and abundant) could be calculated. The model was tested using fifty-three samples and achieved a good degree of correlation between the predicted classes and the Expert's classification. Boyd et al. [13] reports a similar methodology, using Dempster-Shafer's Theory of Evidence in place of Bayesian inferencing. This work has been drawn upon in this study, and is discussed further in Chapter 4. Recently, Walley [168] has applied the Bayesian approach to the Severn-Trent data (Section 4.3) and shown that it significantly outperforms TBI and ASPT. Several paradigms from machine learning have also been applied, with moderate success, to the problem of classifying river water quality [27].

In a broader context, the Bayesian approach has been demonstrated as an alternative to classical frequentist statistical methods. The reliance, perhaps over reliance, on classical statistics based on hypothesis testing and assump-

(a) Response of different species




(b) Effect of abundance




(c) Combined species response

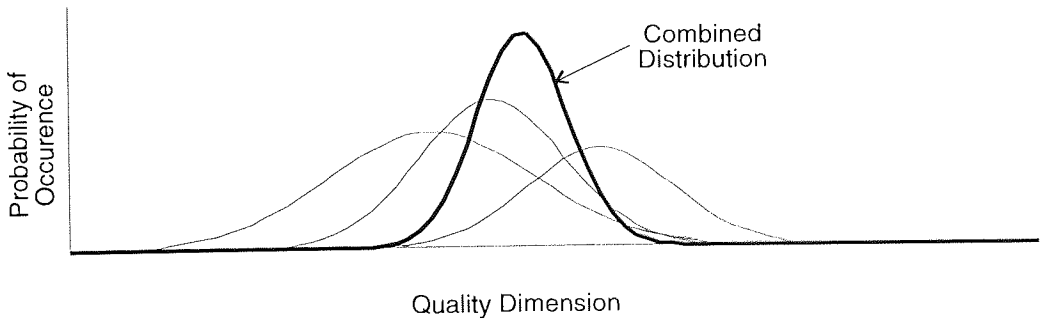



Figure 2.3: Probabilistic interpretation of species response.

tions of data normality has also been commented on within an ecological context [133, 21]. Both Reckhow [133] and Conquest [21] describe the frequent misunderstanding of the 'P-value' in hypothesis testing. The 'P-value' gives the probability of the extreme values (tails of the distribution) given that the null hypothesis is true, but this is often mistaken for the probability that the particular hypothesis is true given the observed data. Within a Bayesian framework the probability of the hypothesis being true given the data is a natural result of the calculation (see also Section 3.6).

Other knowledge-based system work is reported by Wishart et al. [177], who also consider expert systems for interpreting river quality data. One system is used for the automation of compliance testing and classification, while a second application is described for interpreting ammonia levels in fish.

### 2.3.4 Other Techniques

All of the methods so far described fall into the category of structural approaches. Contrasting methods, often termed functional approaches [138], are also gaining in popularity. These include toxicity tests and studies based on the functional feeding groups [132]. Chapter 7 describes and uses toxicity test responses of four taxa for the development of sediment toxicity guidelines.

The use of invertebrates for the rapid assessment of water and habitat quality is becoming more widespread. For example, in the United States the Environmental Protection Agency (EPA) has produced a document which gives a series of protocols (three macroinvertebrate and two fish) which are technical references for conducting cost-effective rapid biological assessments of lotic systems.

## 2.4 Summary of Biological Monitoring

Papers by Walley et al. [170] and Hawkes [53] contain sections discussing the limitations of biological surveillance. These can be summarised as follows.

*i.* Biological surveillance provides an indirect measure of the water quality, thus chemical analysis is still required to pinpoint the cause of any pollution.

*ii.* Biological surveillance should ideally incorporate all aspects of the ecosystem, but such appraisal demands considerable sampling effort and processing.

*iii.* At present the data generated by the biological surveillance can only be fully understood by an experienced river ecologist, thus score systems and indices have been developed to simplify the data to an easily understood form, but hence result in a substantial loss of information.

*iv.* Poor quality water, when determined by biological methods, can be considered to be poor quality, but biologically good quality water may not be good from a human point of view, since in may contain many pathogens.

*v.* Water quality is not the only factor affecting the composition of benthic communities: others include the substratum composition, current velocity and the nature of the surrounding catchment. This makes spatial comparison between sites difficult using the traditional score systems.

*vi.* It is difficult to use biological measures to set environmental quality objectives, as it is very hard to predict what remedial action is necessary to restore a degraded environment to achieve a biological objective.

Despite these weaknesses biological surveillance provides the following benefits.

*i.* The benthic communities act as continuous monitors, in contrast to the periodic sampling regimes used for chemical analysis,

*ii.* The response of the benthic communities covers a wide range of determinants and pollutants, whereas chemical tests are only carried out for chemicals which are likely to be found. Thus, the less common or complex chemicals will not be tested for and hence not detected, even if present in significant concentrations, other than by biological surveillance.

The potential of artificial intelligence in overcoming some of the above limitations, especially in the classification and interpretation, is very real. However, its acceptance into mainstream use is likely to meet resistance because of its novel approach and the relative complexity of its mathematical foundations.

# Chapter 3

# Artificial Neural Networks

## 3.1 Introduction

This chapter reviews the areas of neural network research that are pertinent to the applications considered in this dissertation. There is only a small body of literature on the specific application of neural networks to the interpretation or classification of benthic invertebrates for river water quality classification, including Ruck et al. [148] and Walley [168] which is discussed in Chapters 4 and 5.

The term 'neural network' envelops a wide range of static and dynamic mathematical models that are characterised by a high level of interconnectivity of simple processing units. Even this broad definition, however, fails to cover all of the models that fall within neural network research. The term 'neural' is a little misleading, for although the original motivation behind early research was to model the structure and function of the brain, present-day models are often biologically implausible, and are frequently referred to as connectionist systems to remove the biological overtones. What neural nets do represent are generalised non-linear models which are well suited for use in prediction and classification tasks, for they are universal approximators capable of performing complex function approximations and mappings [145, 40]. The models, while not being strictly non-parametric, incorporate no prior assumptions of the relationship between the input/output mapping [141], and are occasionally referred to as semi-parametric models [142]. There have been numerous review articles [91, 92, 64, 60], books [172, 124, 7, 58], and collections of papers [3, 2] published which are good introductory texts to the discipline.

All the networks considered within this dissertation are static deterministic models, in that the equations used have no temporal component; the output of each network is a function of the current input only [64]. In contrast, dynamic systems, for example recurrent neural nets, have a time component, usually via feedback connections [173]. Of all the possible static models available the Multilayer Perceptron (MLP) has received the most attention in the scientific literature, and is the basic model considered in this dissertation. Another popular model is the Radial Basis Function (RBF) network [15, 134, 135]. The performance of the RBF networks are typically of the same order as that of the MLPs, and have not been considered in detail in this dissertation. However, recently their hidden layer units have been used to provide an additional level of interpretation that is not possible with MLPs[143]. This is discussed in the Section 5.5 which looks at novelty detection in the input data.

The field of neural network related research has grown rapidly since the landmark publication by Rumelhart et al. [151]. This is demonstrated by a quick search on the terms 'neural network' and 'multilayer perceptron (MLP)' of the Science Citation Index, which shows a huge increase in the frequency of the references since that publication. As shown in Table 3.1, there was a geometric increase in the number of publications over the first few years, which has only recently began to level off. This makes a comprehensive literature review difficult on two counts; firstly the diversity and sheer quantity of the references means that it is infeasible to cover (or even find) all the work relevant to this study; and secondly the expansion is happening so quickly that anything published may be out of date or superseded by the time it is published in a journal or has been presented at a conference.[1] The range of applications that have seen neural networks applied to them has been vast, with two large areas of interest being speech processing [11, 135, 108] and handwritten character recognition [89, 61].

The majority of the models in this report have been trained using supervised learning procedures. Supervised learning uses data sets that have class labels that specify the correct classifications for a particular input vector; whereas in unsupervised learning the data have no labels, the classes are iden-

[1] Even keeping track of the relevant journals and conferences is no easy task.

| Year | Frequency |
|---|---|
| 1986 | 64 |
| 1987 | 109 |
| 1988 | 550 |
| 1989 | 941 |
| 1990 | 1281 |
| 1991 | 2055 |
| 1992 | 2650 |
| 1993 | 2608 |
| 1994 | 2538 |

Table 3.1: Frequency of papers published with key-words of 'neural network' or 'multi-layer perceptron (MLP)' since 1986 from the Science Citation Index.

tified after the learning phase. Thus supervised systems are presented with a set of example input-output pairs, and are modelled (trained) to implement a mapping from input to output that matches the underlying generator of the data as closely as possible, to an appropriate level of precision. Whereas unsupervised systems form their own classifications, where the class membership is based on common features in the input data [7]. The words training and learning, in neural network parlance, refer to the selection of the model's parameters, in statistics this is referred to as estimation. For simple networks the parameters are the interconnecting links between the nodes of the network, the links can take on different strengths (weights) which are adjusted during the learning phase. The selection of the most suitable (probable) parameters is dependent upon one set of data, this data being referred to as the training data.

## 3.2 The Multilayer Perceptron

The key building block of a MLP network is the single perceptron unit, Figure 3.1 [100, 147]. A single perceptron unit takes an $n$-dimensional input vector and produces a single scalar output. A weighted sum of the inputs is computed and to this a bias value is added. The result is passed through a non-linear transform to produce the output. The non-linear function is referred to as

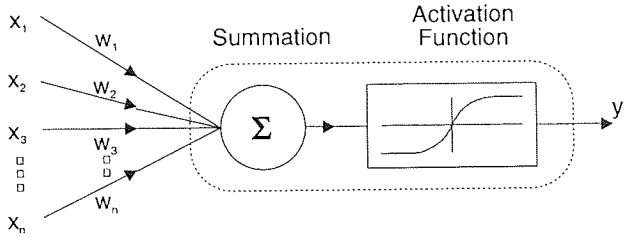



Figure 3.1: Schematic illustration of a single perceptron unit.

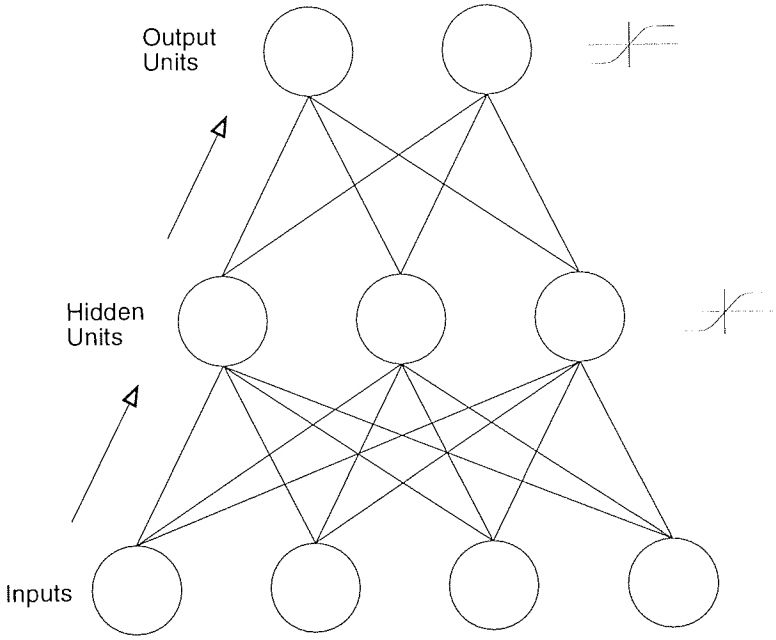



Figure 3.2: A typical multilayer perceptron (MLP) network.

the activation or transfer function, and in early networks it was commonly the step (Heaviside) function.

A cascade of single perceptron units into layers creates the topology known as the multilayer perceptron (Figure 3.2). The action of a single unit or node in the network is equivalent to a single perceptron or processing unit, although it is more usual to use a sigmoidal activation function (similar to the one shown in Figure 3.1) rather than the step function. A wide range of activation functions can be used, including sigmoids, linear and hyperbolic (e.g. *tanh*). All these functions are continuous, vary monotonically and are differentiable, conditions which are necessary for gradient evaluations (derivative information) in the minimisation algorithm.

It has been demonstrated that a three layer MLP (i.e. having one hidden layer) is capable of forming an arbitrarily close approximation to any singled-valued continuous mapping [92, 64]. These results require that the non-linear transform is a continuous, smooth, monotonically increasing function that is bounded above and below, but there is no requirement for the non-linear function to be present on the output layer. For this reason, it is quite common to see linear transfer functions on the output units as this makes the learning process less costly computationally. However, the number of hidden units required to achieve these mappings may be exceedingly large, and problems relating to training time and function minimisation become important constraints.

The first two-layer MLPs were shown by Minsky and Papert [105] to be computationally limited. Extensions to three layers (i.e. more computationally capable models) were unable to be implemented as no algorithm had been developed which could solve the so-called credit assignment problem. The error back-propagation algorithm, which allowed a three layer MLP to overcome the credit assignment problem (and hence to learn), was introduced by Rumelhart et al. [150].[2] Similar work had also been independently described by Werbos [174] and Parker [125].

The most commonly used arrangement of perceptron units is a three layer network, composed of input, hidden and output layers.[3] The nodes on the input layer are not processing units, for there is no computation carried out on these nodes; the values placed on the input nodes are simply the input part of the input-output pairs in the training data. The nodes on the hidden and output layers are processing units, with their inputs being generated by preceding layers. The interconnection scheme is usually of the form; all input units are connected to all hidden units, and all hidden units are connected to all output units, as in Figure 3.2. This is described as a fully connected network. Alternatives are possible; these include additional hidden layers, skip layer connections (e.g. input to output) or a sparser connection scheme between layers.

[2]MLP describes the architecture of the network, while back-propagation is the learning algorithm that is used to adjust the strength of the links between the processing units.

[3]Notationally, this kind of network is also frequently referred to as a two layer network. In this dissertation the network shown in Figure 3.2 is a three layer network, and is the one which is used for the majority of the experimental work in Chapter 5.

## 3.3 The Back-Propagation Algorithm

The back-propagation algorithm [150] consists of two distinct phases. The first stage is the forward propagation of the input pattern from the input units, through the hidden units, to the output units. Firstly, the hidden units' values are calculated by:

$$y_j = f^H\left(\sum_i w_{ji}x_i + \theta_j\right) \tag{3.1}$$

where $y_j$ is the output of the $j$th hidden unit, $w_{ji}$ is the weight connecting the $i$th input unit to the $j$th hidden unit, $x_i$ is the value of the $i$th component of the input vector, $\theta_j$ is the bias to the $j$th hidden unit, and $f^H(\cdot)$ is the hidden layer (H) activation function. Secondly, the network's outputs are given by:

$$y_k = f^T\left(\sum_j w_{kj}y_j + \theta_k\right) \tag{3.2}$$

where $y_k$ is the output of the $k$th output unit, $w_{kj}$ is the weight connecting the $j$th hidden unit to the $k$th output, $y_j$ is the output from the $j$th hidden unit (as defined in Equ. 3.2 above) and $\theta_k$ is the bias for the $k$th output unit, and $f^T(\cdot)$ is the output (or target, $T$) layer activation function.

In the second stage the output signals from the network are compared to the desired or target outputs, and an error signal is back-propagated through the network to adjust the strengths of the links. The usual error function used is the mean square error (MSE), thus:

$$E^{net} = \frac{1}{2P}\sum_{pk} \parallel y_k^p - t_k^p \parallel^2 \tag{3.3}$$

where $E^{net}$ is the MSE, $P$ is the total number of patterns, $t_k^p$ is the target for the $k$th output unit and $y_k^p$ is the network output of the $k$th unit for the $p$th pattern. Other error terms are commonly encountered, for example when the output units are considered to represent probabilities a cross-entropy error term is used in conjunction with soft-max constraints [14]. The soft-max algorithm constrains the outputs to sum to unity and to take on values between 0.0 and 1.0.

A steepest (gradient) descent algorithm can be used for minimising the function $E^{net}$, with the change in link strength being given by:

$$\Delta w_{ij} = -\eta \frac{\partial E_p^{net}}{\partial w_{ij}} \tag{3.4}$$

where $w_{ij}$ is the strength of the link between units $i$ and $j$, the parameter $\eta$ is the learning rate (usually lying in the range of 0.0 to 1.0) and $E_p^{net}$ is the error for the $p$th pattern. A commonly used extension to this equation is the addition of a momentum term, thus:

$$\Delta w_{ij_{new}} = -\eta \frac{\partial E_p^{net}}{\partial w_{ij}} + \alpha(\Delta w_{ij})_{old} \tag{3.5}$$

The first derivative of the error ($E_p^{net}$) with respect to the hidden-output weights is given by:

$$\partial E_p^{net} / \partial w_{kj} = \delta_k f_k'^T y_j \tag{3.6}$$

where:

$$\delta_k = (y_k - t_k) \tag{3.7}$$

and

$$f_k'^T = \frac{\partial}{\partial x} f_k^T(x) \Big|_{x=\sum_j w_{kj} y_j + \theta_k} . \tag{3.8}$$

Likewise, the derivatives of the error with respect to the input-hidden layer weights is thus:

$$\partial E_p^{net} / \partial w_{ji} = \delta_j f_j'^H x_i \tag{3.9}$$

where the delta term is given by:

$$\delta_j = \sum_k \delta_k f_k'^T w_{kj} \tag{3.10}$$

and

$$f_j'^H = \frac{\partial}{\partial x} f_j^H(x) \Big|_{x=\sum_i w_{ji} x_i + \theta_j} . \tag{3.11}$$

The weights can then be adjusted either after all the patterns in the training set have been presented (batch or off-line learning) or after each individual pattern has been presented (on-line learning). Other processes are also encountered. The weights are adjusted in an iterative fashion until a suitable

stopping criterion is met.

The problem of minimising the error function can be considered as an unconstrained optimisation problem, as the weight values are not constrained and can take any real scalar value. The two most commonly unconstrained optimisation procedures used for neural net minimisation, apart from steepest descent, are the quasi-Newton style algorithms and conjugate gradients methods [44, 38, 131]. These methods tend to be very robust when compared to simple steepest descent (with or without a momentum term), and this removes the need to choose suitable values for the learning rate, $\eta$, and the momentum, $\alpha$, which has been discussed and endlessly refined in the literature [62, 130].

All descent based algorithms suffer from problems of encountering local minima during the training phase, which can halt learning before a reasonable solution has been reached. Even so, many of the solutions obtained are satisfactory and provide good results in practice. Global optimisation procedures, such as simulated annealing, are available but are costly computationally and are rarely used in practical problems at the present time.

## 3.4 Generalisation

The performance of classifiers should be judged on new data, not just on the set of training examples used to determine the model's parameter. The ability of the model to predict or classify new data, independent of the training data, is referred to as the generalisation ability of the model. A measure of the generalisation capabilities of each model should be included in the model selection process to safeguard against a highly biased, and possibly unreliable, model being selected for use in the final system. Learning can be considered a two-level process: firstly the classifier learns the coarse features that distinguish between sets of objects in different classes; then further learning produces a more exact mapping by fitting the model to the noise. The process is not discrete, but continuous, and it is essential to stop the learning procedure before the fitting of the noise starts to degrade the model's ability to generalise.

The level of fitting allowed is related to the smoothness of the error function. The more complex the network the less smooth the fitted function is likely to be, and hence the greater the possibility of over-fitting the data. In feed-

forward networks the smoothness can be controlled by a regularisation term, which is added to the error term $E^{net}$ to penalise the complexity of the network. The total error $E$ is then given by:

$$E = E^{net} + \lambda C \tag{3.12}$$

where $E^{net}$ is an error measure based on the misfit of the mapping (e.g. Equ. 3.3), and $\lambda C$ is the regularisation term with $\lambda$ being a scaling parameter and $C$ the complexity measure.

In neural network literature the regulariser most commonly encountered is weight decay regulariser, which is defined by:

$$C = \frac{1}{2}\sum_{n} w_i^2 \tag{3.13}$$

where $w_i$ is the $i$th weight in the network, and $n$ is the total number of weights in the network. Here the weights are represented as a single vector, as topological considerations of the connection scheme are not required. The weight decay of Equ. 3.13 can also be interpreted from a Bayesian perspective [94, 95]. In linear statistical models weight decay is akin to ridge regression methods, reducing the effective number of parameters by removing the independence between the individual weights. Nowlan [120] describes other, more sophisticated, regularisers that can be implemented.

Another factor affecting the complexity of the mapping, and hence the generalisation, is the number of parameters used to achieve the mapping. The greater the number of 'effective' parameters the more likely it is that overfitting will occur, and thus the more likely that the generalisation ability of the network will be poor [107, 110]. As the number of input and output units are determined by the size of the data sets, the number of weights within the network is dictated by the number of hidden units and the interconnection scheme used. The larger the number of hidden units the greater the number of parameters within the network. The most suitable number of hidden units to use is problem dependent, with some simple mappings needing only a few units, while others require hundreds or even thousands. There are ad hoc methods for the determination of the most suitable number of hidden

units, but the best method is experimentation using either cross-validation or Bayesian based analyses.

There are various formulae for assessing the ability of a network to generalise, but these only provide upper and lower bounds limits. The difficulty in obtaining a reliable generalisation measure is due to the non-linear transfer functions and the intractability of the mathematics involved with this. Most of the formulae (e.g. the Vapnik-Chervonenkis Dimension (VC dim, [166, 64]), or the Generalised Prediction Error (GPE, [107])), suggest that the number of effective parameters (weights) should be much smaller than the number of training cases.

## 3.5 Model Selection

Final model selection should be based on data which are independent of the training data. This requires that the full set of data, all the available samples, be split into three parts as illustrated in Figure 3.3:

*i.* one for the determination of the model's parameters (training data),

*ii.* one for the selection of the most suitable model (validation data), and

*iii.* one for the testing the model's performance on the independent data (testing data).

This is considered further in Section 5.2.2.

Henery [57] discusses comparative methods of model selection for classification problems, and some of his ideas have been used in this dissertation to aid comparison between different neural net models. These include the estimation of error-rates using train and test methods, cross-validation, and the organisation of comparative trials. Hypothesis testing can be carried out between individual experiments, based on the $t$-test, and also confidence limits can be placed around a regression network's predictions [72], but these must be treated with caution as assumptions made may not be strictly valid. Alternative methods can be used, including boot-strapping, cross validation and Bayesian analyses [94, 95, 28, 162].

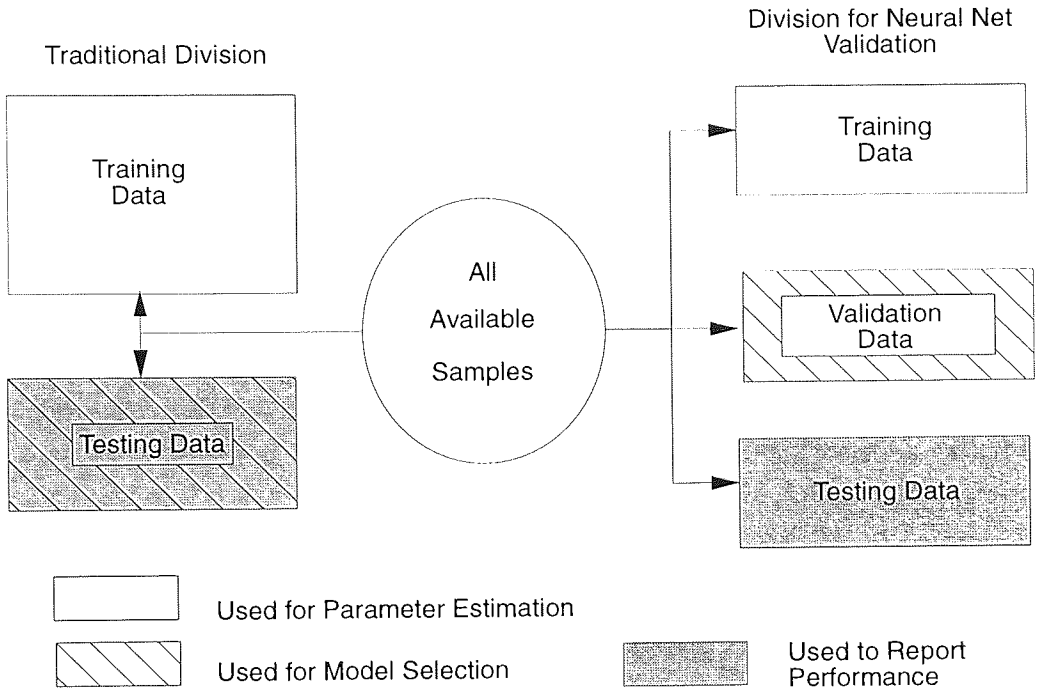



Figure 3.3: Division of data for parameter and model selection.

Comparisons of neural networks with other classification systems have been made by Thrun et al. [163], Ripley [141] and Michie et al. [104]. In Michie et al. [104] a number of machine learning, neural and statistical algorithms are compared over 22 large data sets. From these studies it can be seen that the relative performance of classification methods is very problem dependent, and although there may be some prior knowledge of when particular algorithms are unsuitable, very little can be inferred about relative performances on specific data until experiments on them have been conducted. Although some work was carried out within this project which compared the performance of standard statistical techniques with that of neural networks, it was not the major concern of this dissertation.

Another idea which is explored in this dissertation is that of using a combination of models to form a consensus classification, akin to a panel of experts which is called upon to make a decision [120, 70, 79, 88, 142]. There are several possible methods of combining a series of models and these are discussed in more detail in Section 5.3.5.

## 3.6 Bayesian Perspective

More recently a Bayesian perspective has been applied to the theory of neural network models [17, 116, 94, 95, 117]. Using Bayes' theorem the probability of a model $\mathcal{N}_i$ given an observed data set $\mathcal{D}$ can be written as:

$$P(\mathcal{N}_i|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{N}_i)P(\mathcal{N}_i)}{P(\mathcal{D})} \tag{3.14}$$

To compare different models only the $P(\mathcal{D}|\mathcal{N}_i)$ term needs to be evaluated for each model (assuming each model is equally likely), and this term is called the *evidence* for model $\mathcal{N}_i$. Bayes' theorem can be applied to different levels of the modelling process, and extends naturally to cover regularisation expressions.

The Bayesian framework provides many important features, which include for prediction (interpolation) problems being able to place error bars or confidence limits on the network predictions. This is especially important in engineering applications where a measure of confidence is essential if a prediction is to be of any value. The error bars reflect the density of the training data, in that where the training data is sparse the error bars are larger than in other regions with a higher density of training data. A second important aspect is that the regularisation coefficients can be optimized using only the training data, thus removing the need for a cross-validation data set. It also allows for comparison of models of different type (e.g. MLPs, linear discriminants or RBFs) as well as different topologies (e.g. different numbers of hidden units).

The drawbacks of the Bayesian methods are the increased complexity of the programming required to implement the theory and that numerical instabilities occasionally occur when evaluating the evidence if the parameters of the network are poorly determined.

## 3.7 The Future

At the present time, the mathematical and statistical theories of the structure, function and behaviour of neural networks are being consolidated. Much of this work is particularly mathematical in nature, especially the development of Bayesian methods, and this will take some time to filter down into mainstream

use. Perhaps, the major reason for the popularity of the back-propagation algorithm is its intuitive representation and that it is comparatively easy to write in computer code or in statistical packages. It is a fairly easy exercise to implement a working three layer MLP model using the back-propagation algorithm, and in some respects the benefit of the more sophisticated neural network models does not outweigh the costs of writing and debugging additional, and frequently more complex, code. However, the availability of good quality software packages (both commercial and public domain freeware) which implement neural network models is steadily increasing and this should help to promote the use of more advanced networks.

Areas that will receive much greater attention will include the reliability, confidence (which is related to novelty), model-order selection of the models and the use of committees of models. With the increasing use of neural networks, and other non-linear statistical models, in critical tasks and control problems the need for confidence in the output of the neural network will become more important. Some of these issues are considered in Chapter 5.

## 3.8 Summary

Multilayer perceptrons are popular mathematical models for classification and prediction tasks. They are more successful than what the critics suggest, but less remarkable than the most ardent supporters claim. However, the technology is becoming more common place and will be soon start filtering through to domains other than, for example, speech recognition, hand written character recognition and image processing, which are traditionally the interest of mathematicians and computer scientists. This dissertation aims to demonstrate that the application of these new tools of AI in the domain of biological monitoring is relatively easy and has the potential to produce many benefits, although a radical shift of ideas may be necessary for them to be realised.

# Chapter 4

# Analysis of River Data

## 4.1 Introduction

This chapter describes, analyses and discusses both the biological river data and the underpinning conceptual arguments that form the core of this dissertation. The chapter comprises three main sections:

**Section 4.2** describes the elicitation of knowledge from the Expert, and explains the basic assumptions underlying both the elicitation and the classification system. The latter is used as the conceptual basis for many of the following arguments. The probabilistic knowledge was elicited in two ways. The first was by direct elicitation of probabilistic information from the Expert, which was conducted by Boyd and Walley as part of an associated PhD project [12]. These probabilities formed the knowledge base of the evidential reasoning system christened BERT (Benthic Ecology Response Translator) by Walley et al. [170], and this can be considered as the starting point for the work in this dissertation. The second, indirect, approach to elicitation, which served the dual purpose of providing the required probabilities for BERT and a good quality data set for the training and testing of the neural network models, simply required the Expert to classify invertebrate samples into a biologically based quality class.

**Section 4.3** describes the Severn-Trent data which was the project's key data set. This data is used extensively in the neural network experiments (Chapter 5) and for the identification of indicator taxa (Chapter 6). The data set consisted of benthic invertebrate records together with some

chemical and physical information on the sample sites. The format of the data is reviewed and many of the problems associated with benthic data are discussed. The data were classified by the Expert and the consistency of the Expert's classifications is investigated and discussed. The popular biotic indices are compared with the biological classification mentioned above, by use of the expertly classified data.

**Section 4.4** describes the additional river data sets that have been studied. These include a set from the old Yorkshire Water Authority, the NRA National Survey database and synthetic data based on the conditional probabilities of the direct elicitation sessions. The National database is used to present distributions of some of the key taxa families, as well as the variation in the BMWP score and ASPT between the different NRA regions. The synthetic data set was used extensively during the neural network experimental work.

Finally, the chapter is summarised and the main findings are highlighted.

### 4.1.1 Chronology of Project Data

A feature of the project is that a wide variety of data have been studied because changes in the availability of data influenced the direction which the project took. At the start of this project the BERT system's knowledge base had just been extended from 20 to 41 taxa, and a small set of Yorkshire River Authority data, comprising some 50 samples, was available for some preliminary small-scale neural network experimental work. However, the quality of this data was poor and no further work was undertaken using this data once more suitable data became available. The next data made available was that from the Severn-Trent Region of the NRA, and this was the main data studied in the project. During the final stages of the project it was decided to extend the neural network work to provide results from a larger data set, thus a synthetic data set was created using the probability distributions derived from the direct elicitation exercise. While this work was being completed the NRA's 1990 National Database was made available, but at this late stage no more than a summary analysis was possible.

## 4.2 Biological Classification and Knowledge Elicitation

### 4.2.1 Introduction

The elicited knowledge from the Expert can be categorised into two types: qualitative and quantitative. The qualitative knowledge encapsulated the basis of biological monitoring using benthic invertebrates as indicators, while the quantitative information specified conditional probabilities of finding particular taxa in particular levels of abundance in a watercourse of a given water quality [170]. It was the latter probabilistic information that was encoded into the BERT system, while the former, qualitative information was unstructured and quite diverse, and would be difficult to apply in a plausible reasoning system like that of BERT. A rule based system could have been developed from a structured set of qualitative rules, but difficulties in handling uncertainties would have arisen.

Two methods of elicitation, namely direct and indirect, were used to generate the conditional probabilities. For the direct elicitation the Expert graphically depicted the conditional probabilities, while the indirect elicitation took the form of the Expert classifying complete samples of benthic data. The resulting classified data allowed supervised learning procedures to be used for the training of the neural networks. The classification system which was adopted was an important part of the knowledge elicitation exercise, and is explained in the following section.

### 4.2.2 Biological Classification

Classification can be viewed as a method of introducing convenience into the description of complex data. The biological classification (i.e. a system of classes based on measures taken from the biology) used in this dissertation is a simplification of complex naturally occurring phenomena, but is conceptually valid despite its simplicity. The biological classification is based upon the level of organic pollution in a river, and as such is equivalent to many of the biotic indices that are encountered in the literature. Although this may seem

somewhat narrow, as many of the more serious pollutional episodes do not fall under the category of organic pollution, it provides a means of comparing the proposed method with the existing ones, which is important for the introduction of any new system into an existing field. Additionally, if another source of pollution were to be affecting the community structure then this could be reflected by the classification. For example, a river with some organic and some toxic pollution may be classified as one grade worse on the organic scale, resulting in equivalent organic gradings in biological terms.

Another possibility could be to view water quality as a multi-dimensional space, with organic pollution considered as a single dimension within this space. Other dimensions could be considered in isolation, such as heavy metal pollution or acidification, and for these a similar conceptual classification system could be adopted. This implies that the methods used in this dissertation for classification are extendible to cover other types of pollution apart from organic.

The continuum of organic pollution was split into five main classes, with these classes being chosen to mirror the present NWC classification. The classes were labelled as B1a, B1b, B2, B3 and B4, with B1a being the best quality and B4 being the poorest.[1] Figure 4.1 depicts this, as well as a finer 13 class scale that was also used in a later elicitation exercise. The first thing to note is that there is a distinct ordering to the classes (B1a to B4), but there is no explicit idea of distance between the classes. This can be described as an ordinal set of classes [31]. The second important aspect to the classification is that it is based on the Expert's subjective assessment. This is important as it implies that the classification will differ between experts, and it is unlikely that any two experts will agree precisely.

The subjectivity of the classification appears to be a weakness, as most of the time the goal is for objectivity and the elimination of subjectivity. But, it is much better to make the subjective aspect explicit and use objective methods for any inferences from the starting assumptions than to assume the modelling process is objective (when typically it is not) and cloud the results with implicit (and often unstated) assumptions. Any biomonitoring system

[1] For the remainder of the dissertation this classification system will be referred to as the 'biological classes'.

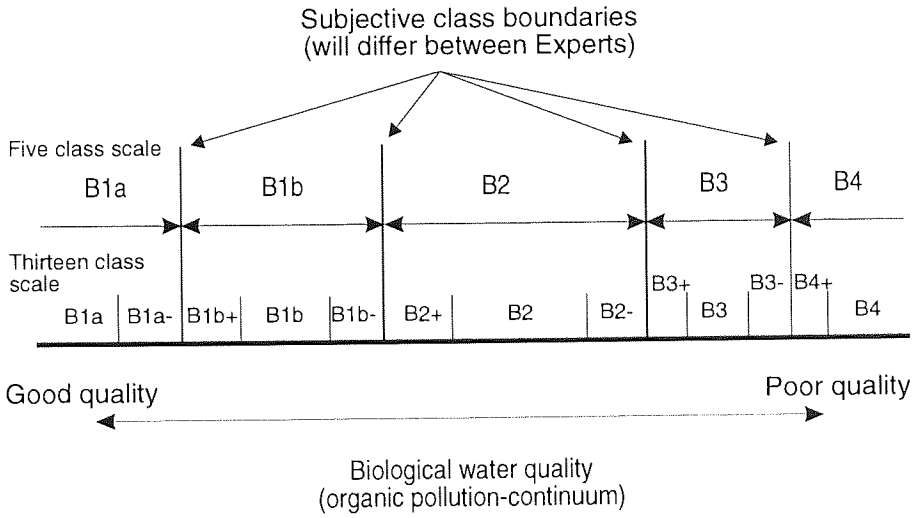



Figure 4.1: Classification of river water quality into classes based on organic pollution.

that uses autecological information will be to some extent subjective, just as water quality is [52], and, for example, the BMWP system has a subjective basis underpinning its implementation.

The similarity between the biological classification system in this dissertation and the Saprobic system is particularly apparent. The core Saprobic system uses five classes (although additional classes are occasionally drawn upon) and is based on organic pollution. Figure 4.2 shows the likelihood of finding *Gammarus pulex* and *Asellus aquaticus* in a given biological class. The Saprobic valencies (which sum to 10 not 1 like the probabilities) for *G. pulex* compare quite well with the directly and the indirectly elicited probabilities, but relatively poorly in the case of *A. aquaticus*. Figure 4.3 compares the Saprobic valencies with the directly elicited probabilities for *Baetis rhodani*, showing excellent agreement between the two. Also, during a separate elicitation exercise, the Expert referred to the NWC classes in terms of the Saprobic classes,[2] so there is some subjective relationship, in his mind, between our classification based on the NWC and that of the Saprobic system.

The main utility in using a classification system based on the NWC (and resembling the Saprobic) is that there is a clear monotonic relationship between

[2]H.A. Hawkes, personal communication.

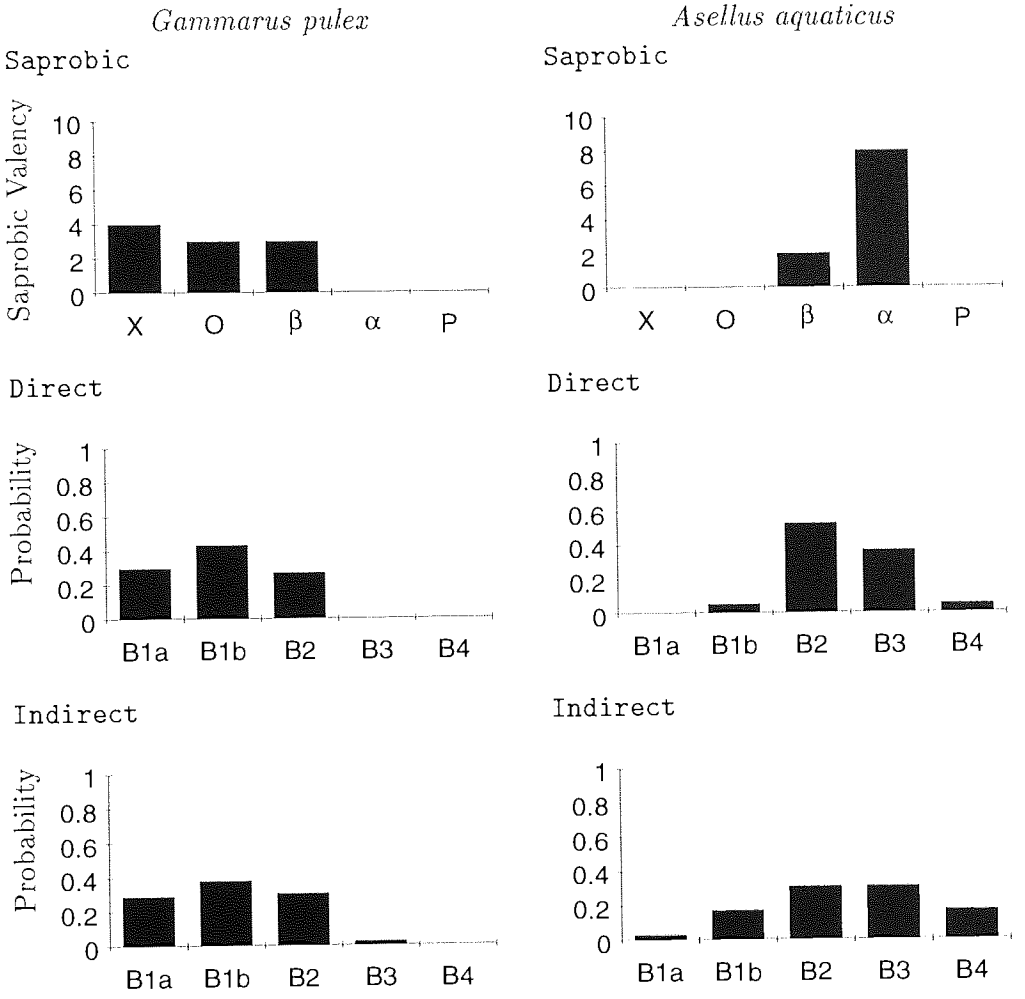



Figure 4.2: Comparisons of direct and indirect elicitations of biological class with the Saprobic valencies for *Gammarus pulex* and *Asellus aquaticus* [157].

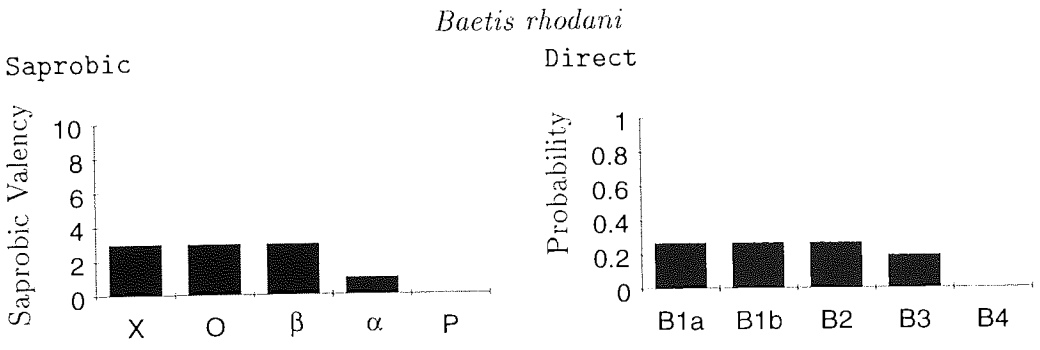



Figure 4.3: Comparisons of direct elicitation of biological class with the Saprobic valencies for *Baetis rhodani*.

the quality class and the level of organic pollution. This is not the case with, for example, the BMWP score, where it is possible to have a non-monotonic relationship between the score and the level of organic pollution.

### 4.2.3 Direct Elicitation

During the development of the BERT knowledge-based system a series of knowledge acquisition sessions was conducted by Walley and Boyd [12]. A brief summary is given here as their work was extensively drawn upon during this project. There were two stages to the direct elicitation process: initial sessions aimed at extracting the core philosophies and conceptual bases of biological monitoring, and later sessions directed toward the acquisition of 'hard' knowledge, which formed the core of the BERT knowledge base. The early interviews gathered information on the principles of biological monitoring, and without doubt helped to shape the eventual system. The information garnered included what the Expert looked for when interpreting a sample, how the information was recorded and disseminated within the water industry, the utility of using biological data, the interpretation of the data, its benefits and disadvantages when compared to chemical methods, and the existing systems used to assess and classify rivers.

Building upon the first stage, 'hard' probabilistic knowledge was elicited from the Expert using a graphical approach, which required the Expert to draw histograms representing the likelihoods of finding particular taxa given different quality classes. The list of key indicator taxa used in the BERT system was elicited in three stages, Table 4.1 lists the taxa and the stages at which they were selected. The criteria on which the list was drawn up were that the taxa had to be readily identifiable and, as a set, cover the whole range of quality classes, and that each taxon should occur fairly frequently in its preferred class [170]. Four states of abundance were defined: absent, rare, established or abundant. The precise definition of these states varied between taxa, the details of which are given in Table 4.1.

Figure 4.4 summarises the probabilities of finding the taxa in each quality class arranged in ascending order of tolerance to organic pollution. The ordering is based on a weighted average of the probability mass, with weightings

Table 4.1: The forty-one taxa used in the BERT system.

Definitions: Rare = 1 to $n_1 - 1$; Established = $n_1$ to $n_2 - 1$; Abundant >= $n_2$.

| First Set | $n_1$ | $n_2$ | Third Set | $n_1$ | $n_2$ |
|---|---|---|---|---|---|
| Lymnaea peregra | 3 | 50 | Polycelis nigra | 2 | 10 |
| Tubifex tubifex[1] | 5 | 200 | Dendrocoelum lacteum | 2 | 10 |
| Erpobdella octoculata | 3 | 20 | Potamopyrgus jenkinsi | 3 | 50 |
| Asellus aquaticus | 3 | 50 | Bithynia tentaculata | 3 | 20 |
| Gammarus pulex | 3 | 50 | Planorbis spp. | 2 | 10 |
| Leuctra fusca[2] | 3 | 20 | Ancylus fluviatilis | 3 | 20 |
| Rhyacophila dorsalis | 2 | 20 | Sphaerium spp. | 3 | 20 |
| Hydropsyche angustipennis | 3 | 50 | Pisidium spp. | 3 | 20 |
| Simulium ornatum | 3 | 50 | Hydracarina | 3 | 20 |
| Chironomus riparius | 5 | 100 | Heptagenia spp. | 2 | 10 |
| | | | Ephemerella ignita | 3 | 20 |
| Second Set | | | Amphinemura sulcicollis | 2 | 10 |
| Lumbriculidae | 5 | 100 | Isoperla grammatica | 2 | 10 |
| Glossiphonia spp. | 2 | 10 | Dytiscidae | 2 | 10 |
| Helobdella stagnalis | 2 | 10 | Elminthidae | 2 | 10 |
| Baetis rhodani | 3 | 50 | Glossosoma spp. | 3 | 50 |
| Rhithrogena spp. | 3 | 20 | Agapetus spp. | 3 | 50 |
| Ecdyonurus spp. | 3 | 20 | Polycentropidae | 3 | 20 |
| Caenis spp. | 3 | 20 | Hydroptilidae | 5 | 50 |
| Haliplidae | 3 | 20 | Limnephilidae | 3 | 20 |
| Sialis lutaria | 2 | 10 | Ceratopogonidae | 2 | 10 |
| Other Hydropsychidae[3] | 3 | 20 | | | |

[1]Later accepted as Tubificidae
[2]Later accepted as Leuctra spp.
[3]Other than *H. angustipennis*

of B1a=1, B1b=2, ..., B4=5 being used. The figure also shows the BMWP score for each taxon. There is a fairly good correlation between BMWP score and the order of taxa in the table. As can be observed from the figure, there is a good coverage of all the quality classes, with a good mixture of taxa from all the major benthic invertebrate groups.

There were two basic assumptions underlying the elicitation of the conditional probabilities, these being that the sites in question were:

*i.* riffle sites (i.e. quick flowing river, eroding substratum), and

*ii.* were affected by only organic pollution.

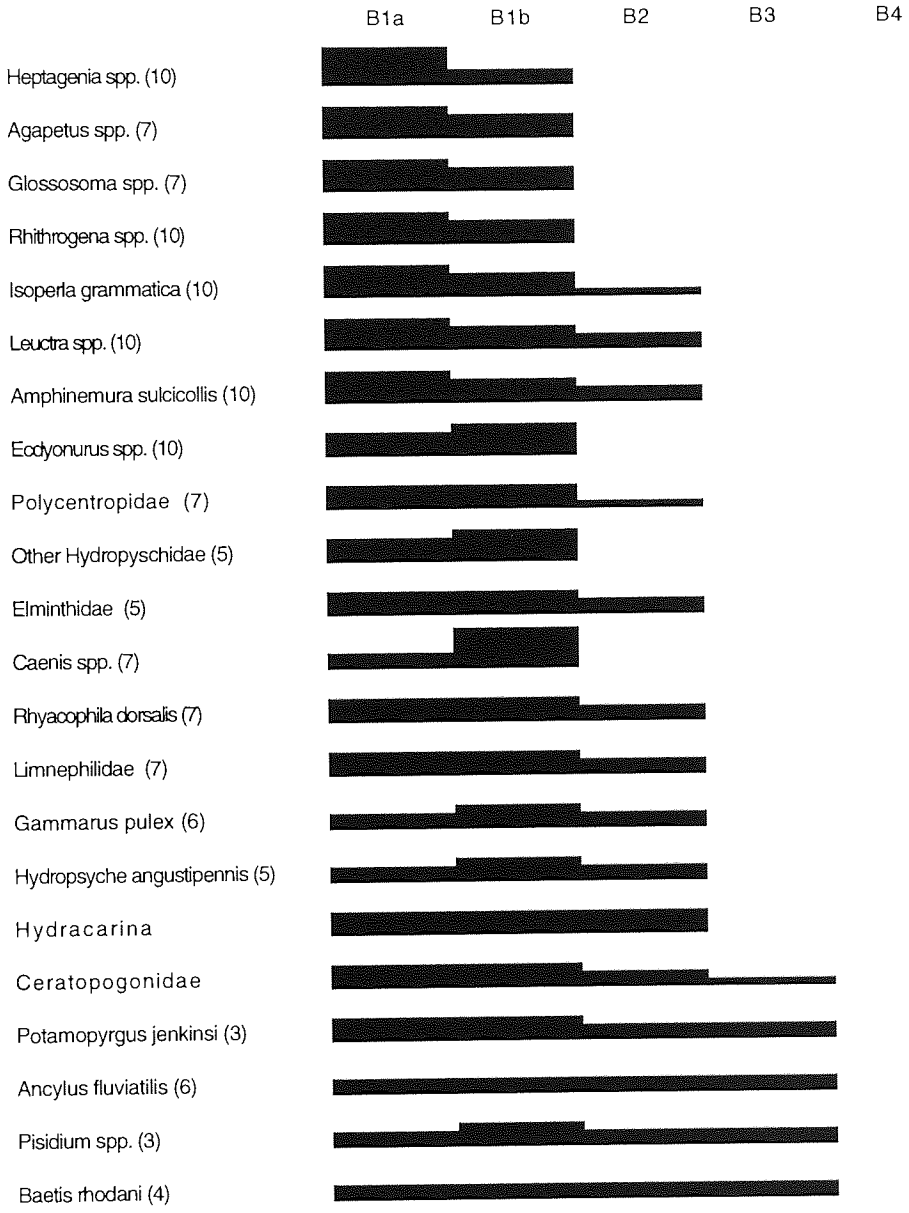



Figure 4.4: Histograms of the probability of finding a taxon present in a given water quality class (derived from the direct elicitation), with the appropriate BMWP score given in brackets. Five discrete levels are shown, these are in descending order of size: $p > 0.7$, $p > 0.5$, $p > 0.3$, $p > 0.1$ and $p > 0.0$.

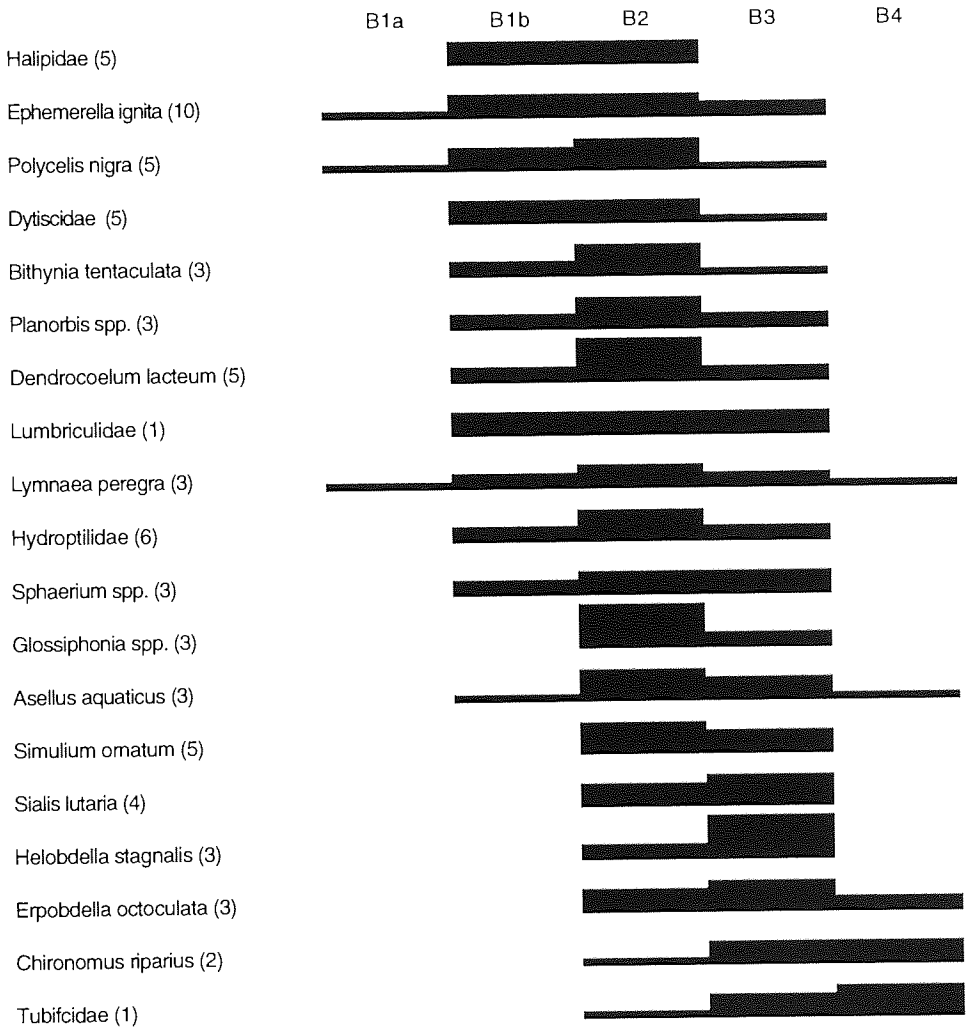



Figure 4.4: Histograms of the probability of finding a taxon present in a given water quality class (cont'd).

The probabilities derived by direct elicitation for use in BERT, also proved useful in this (neural network) study, as they allowed synthetic data sets to be created which reflected the elicited information. This is fully described in Section 4.4.2.

### 4.2.4 Indirect Elicitation

The direct estimation of frequencies and probabilities was somewhat unnatural for the Expert. It was then realised that an alternative and perhaps more effective approach would be to elicit the Expert's knowledge indirectly by getting him to classify a representative set of benthic samples, and then using this information to extract the conditional probabilities. This process came more naturally to the Expert than the estimation of probabilities. The classification of samples was a particularly attractive approach since it clearly served two purposes, these being that it provided:

*i.* the data necessary for the supervised training and testing of the neural networks developed in this study,

*ii.* a means of deriving additional conditional probabilities which were representative of real-world data.

The main data to be classified in this way was the Severn-Trent data. The synthetic data was also classified by the Expert, but this was done to gain feedback rather than to solely classify samples or to generate conditional probabilities. In addition, the indirectly elicited probabilities later provided the basis of the study into the indicator value of taxa, which led to improved coding of the input data to the networks and a corresponding increase in performance (Chapter 6).

Problems were encountered with the indirect elicitation in two areas. The first was that the Expert occasionally had difficulty classifying some samples due to conflict between the nature of the sample and the assumptions underlying the classification system. The most common problem was that the samples clearly contained several species commonly found in pools, thus indicating that the site was probably not strictly a riffle.

The second area of difficulty was how much information should the Expert be given to classify the samples. Since the computer-based systems, both neural and knowledge-based, were originally going to use the 41 indicator taxa, the question arose as to whether the Expert should base his classifications on the full sample, possibly along with any physical or chemical information, or only on the same data which would be presented to the computer systems. The decision was made to allow the Expert to see the full species list when classifying the samples since the object was to develop systems which correctly classified river quality, albeit on a subset of the available data, not to test the ability of the Expert to reason with incomplete information. This also meant that no further classification would be necessary if the list of 41 key taxa were increased, modified or even reduced to a few key taxa. However, it also meant that it was possible to have identical data input to the system having different classifications. The reason for this was that taxa outside of the subset of the 41 key groups may have swayed the Expert's classification one way or another, and this would not be reflected in the data presented to the models. This is the price of data reduction, but it is offset by gains in other areas, such as the need to record information on fewer taxa.

### 4.2.5 Sources of Information Loss

Throughout the elicitation sessions, as well as in the experimental work, an issue that kept emerging was that of information loss. This section briefly expands on this in qualitative terms, whereas Chapter 6 takes a more restricted quantitative approach. The problem of information loss arises in many forms throughout the monitoring process. From the actual population inhabiting the river bed to the final classification placed in a report there are numerous places where information is lost (or equivalently, uncertainty is added). Figure 4.5 shows this pictorially, and it is possible to conjecture about how information can be lost at any particular stage. As the main focus of this dissertation is the interpretation and classification of benthic samples some issues relevant to this are discussed below.

As the community represents a complex, multi-dimensional system, and the goal of classification is the allocation of a single *one*-of-$N$ class, some

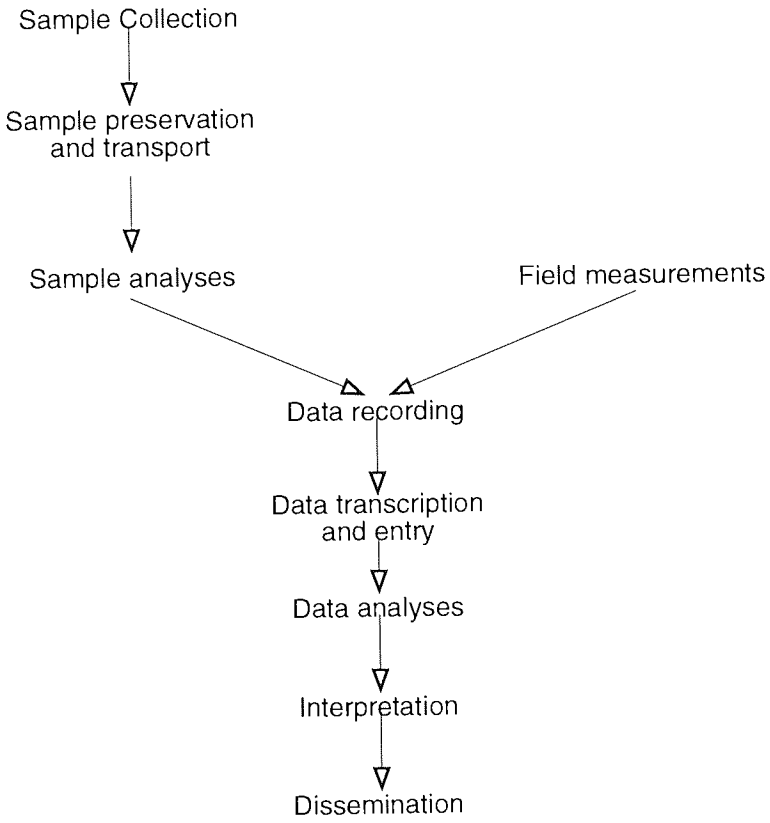



Figure 4.5: Typical elements of a monitoring programme. All are possible sources of information loss (after Norris & Georges [119])

information loss cannot be avoided. However, some biologists see the use of quantitative data as detrimental to effective biological monitoring. They claim that when too much reliance is placed on quantitative systems the biology is seemingly relegated to second place behind the score or index derived from them [39]. They view the use of scores as a convenience, to explain the data succinctly to managers, and that it is a futile procedure to reduce the data to simple numbers that do not convey the information that is present within a sample. The line of argument may be valid, but it is not helpful to think in these terms if biological monitoring is to have an impact on the quality of classification. A reliable and efficient method is required to interpret and summarise ecological data, without such a method the biological monitoring procedures will never gain an equal recognition with their chemical counterparts.

One source of information loss (or shortage, since it was not available) in this project was that of physical information on the sample site (biotope). The relevance of this when interpreting a benthic sample is succinctly stated by De Pauw and Hawkes [24]:

> "River benthic invertebrates are only of use as indicators of river water quality when considered in the context of the biotope in which they were found."

However, none of the biomonitoring systems of scores or indices currently in general use has specifically incorporated biotope type, despite the fact that there is a definite difference in the fauna associated with pools and riffles [93, 16]. Thus, biotope information is very relevant to the classification of water quality, as has been recognised by the developers of RIVPACS. Biomonitoring methods have long favoured the use of riffle samples because the riffle communities provide the most reliable means of classification. This is because in riffles water quality is the most important factor affecting the community structure. At pool locations environmental forces (e.g. substratum type) become much more dominant. The riffle invertebrate allow greater discrimination between differences in water quality than corresponding pool communities.

Riffle sites can, by their nature, support a more diverse fauna than a pool site. In a riffle community an increase in organic matter can lead to either an increase or decrease in the diversity of the community. For example small, good quality streams (e.g. Welsh head streams) support only a limited or restricted community. With the introduction of a limited amount of organic matter, the nutrient availability increases. The net result of the input of more energy into the system is that there is an increased diversity, with a larger and more varied assemblage of animals resulting. Further increases in organic load cause a reduction in the number of the sensitive taxa, leading to a less diverse structure. This is shown schematically in Figure 4.6. Pool sites, which have naturally higher level of organic enrichment, exhibit a more monotonic relationship between organic pollution and diversity. An increase in organic pollution is almost always accompanied by a decrease in the diversity of the community.

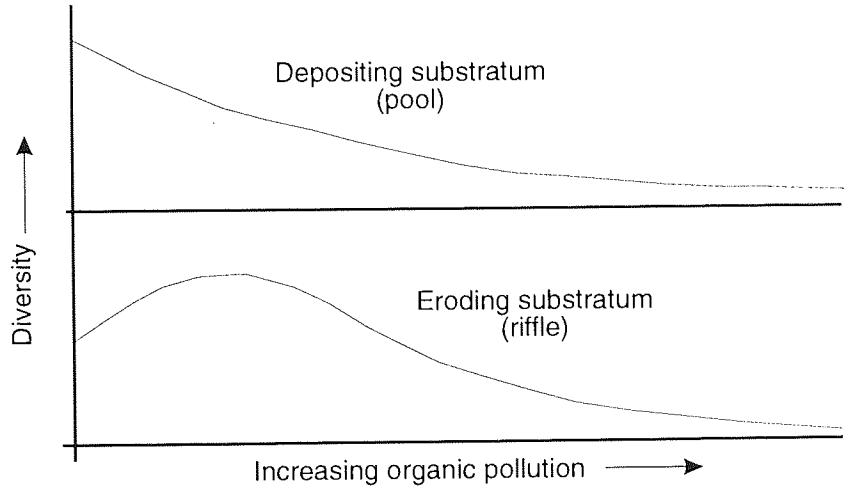



Figure 4.6: Schematic illustration of riffle and pool diversity.

Another area of uncertainty is sampling. There are two aspects to this problem; the first is what do you sample (riffles, pools or a proportion of both), and the second is how do you sample it? The first question is inextricably linked to the biotopes present at the site, and is particularly pertinent to the Severn-Trent data described in the following section. Current NRA sampling policy is that a sample taken at a site should be representative of all the biotopes present. The samples resulting from the different biotopes are combined into one species list. The single species list does not reflect any single biotope, so additional uncertainty is added to any classification or interpretation placed on these data because there is no indication as to the biotope from which any particular taxon originated. This could only be avoided by keeping the samples separate, but this would lead to an increase in costs associated with storage and handling. In riffles, kick heel sampling is popular, but not all sites have riffles. For slow flowing or deep rivers there are a number of alternative possibilities, but different sampling techniques can produce widely different species list for the same location, which leads to difficulties in the interpretation [45].

The problems of information loss and uncertainty are always likely to be present in any form of biological monitoring, so there will never be a definite or precise system developed. The best recourse is to identify the potential areas of information loss, and to try and minimise their extent. This will only be achieved through good design of monitoring programmes.

## 4.3 Severn-Trent Data Set

### 4.3.1 Base Data

The full database contained 1 376 routinely taken benthic invertebrate samples taken from Upper Trent area of the Severn-Trent Region of the NRA. Each record contained the invertebrate sample, as well as additional physical and geographical information (Table 4.2). The invertebrate records consisted of four fields, these were the taxon's name, its abundance level, Maitland code and sample number (see Appendix A1 for the database's full taxonomic list). The Maitland code [96] is an eight figure number which is split into four groups of two digits, and is useful for sorting and searching for taxa within a database. For example in the Severn-Trent database the Maitland code for *Gammarus pulex* is 37 14 02 06: 37 is the order Amphipoda, 14 is for the family of Gammaridae, 02 is the genus *Gammarus* and finally the species name *pulex* is represented by 06. Thus, any taxon which is in the Gammaridae family will have a Maitland code of 37 14 xx xx, and any species of the genus *Gammarus* will be 37 14 02 xx. Each species has a unique Maitland code number, but unfortunately there were subtle differences between the Severn-Trent Maitland code numbers and those of the National NRA, which meant some of the sorting algorithms had to be rewritten when the National database was investigated.

The presence of a taxon was banded into six levels, with Table 4.3 showing the total number of occurrences for each of the abundance levels in both the full database and the 292 sample subset. These numbers, defining the abundance levels, were slightly different to those used in the direct knowledge elicitation (*cf.* Table 4.1), so that there was no one-to-one correspondence between the abundances of the Severn-Trent samples and the BERT knowledge base, but this did not unduly affect the neural network experimental work. For convenience the six abundances of the Severn-Trent data were grouped into three states (*present*, *few*, and *com+*) which were the best approximation to the BERT states of *rare*, *established* and *abundant*. The mapping from Severn-Trent states to BERT states was: *present* $\rightarrow$ *rare*; *few* $\rightarrow$ *established*; and (*common*, *abundant*, *very abundant* and *1000+*) $\rightarrow$ *abundant*.

| Database Field | Format (range) |
|---|---|
| Water course | Text |
| Site description | Text |
| National grid reference | Alphanumeric |
| Location | Text |
| Sample ID | 5-digit number |
| Date | Day, month and year |
| BMWP score | Numeric (1-166) |
| ASPT | Numeric (1.0-7.6) |
| No. of BMWP taxa | Numeric (1-30) |
| Total number of taxa | Numeric (1-36) |
| TBI score | Numeric (1-10) |
| Pebbles | % or a/p |
| Boulders | % or a/p |
| Sand | % or a/p |
| Silt | % or a/p |
| Flow level | Above, normal or below |
| Flow speed | Fast, moderate, slow or static |
| Clarity | Clear, cloudy or turbid |
| Odour | None, slight or strong |
| Shade | None, little or much |
| Depth (cm) | Numeric (0-200) |
| Width (m) | Numeric (0-40) |

Table 4.2: Sample information available in Severn-Trent database.

| Level of abundance | Full database | | 292 Samples | |
|---|---|---|---|---|
| Absent (0) | | | | |
| Present (1-2) | 6416 | (31.27%) | 1299 | (31.15%) |
| Few (3-9) | 8618 | (42.00%) | 1787 | (42.85%) |
| Common (10-49) | 4193 | (20.44%) | 829 | (19.88%) |
| Abundant (100-499) | 824 | (4.02%) | 157 | (3.76%) |
| Very abundant (100-999) | 386 | (1.88%) | 83 | (1.99%) |
| 1000+ | 80 | (0.39%) | 15 | (0.36%) |

Table 4.3: Description of abundance codes in the Severn-Trent database. The number of occurrences of each state in the full and 292 sample database are also given.

Figure 4.7 shows a representative sample from the Severn-Trent database. Some of the physical characteristics are missing and the substratum materials are recorded as present/absent, not as a percentage showing how much of each was present. The main difficulties, however, are with the taxonomic list; some of the animals have been identified to species, some to genera and others to family level (family level was the minimum, apart for difficult to identify groups, e.g. Hydracarina). This in itself is not a problem, but the inconsistency from sample to sample was. Throughout the database this was the case, and the main factor governing the level of identification beyond family level appeared to be the particular biologist examining the sample. The group of organisms which were of particularly interest to him/her were the ones identified to the species or genera levels.

Another problem is revealed by the Glossiphoniidae, *Glossiphonia complanata* and *Helobdella stagnalis* entries. The Glossiphoniidae entry is recorded as *present* and thus implies that there were only 1-2 individuals of the family found in the sample. But, this was clearly not the case because two species of the family, *G. complanata* and *H. stagnalis* are recorded as being *few* (3-9 individuals) and *present* (1-2) respectively. This implies that the Glossiphoniidae entry should be read as meaning 'everything else in the family Glossiphoniidae expect for species or genera recorded elsewhere'. This causes problems with data processing on two counts. Firstly, the entries in the database are no longer independent, and therefore require cross-referencing within the sample, thereby increasing the complexity of any searching and sorting routines. Secondly, the abundances of, in this case, three groups cannot be combined to form a single abundance level for the whole family. What exactly is the abundance of a family that is recorded in three separate components as present(1-2), present(1-2) and few(3-9)? Although such samples caused some concern, an even more worrying case was where the subclass Oligochaeta was recorded as *present* (1-2) and one of its families Tubificidae as 1 000+.

### 4.3.2 Construction of 292 Sample Database

The first step in constructing the database for experimentation was the removal of all the samples originating from canals. The remaining samples were

Severn-Trent Region — Upper Trent Area
Sow R. - Eccleshall

| | | | |
|---|---|---|---|
| Sample No. | : BI 75416 | Date | : 04/02/1991 |
| Water Width | : . m | Grid Ref | : SJ 831 296 |
| Sample Depth | : 20 cm | Location | : 70262820 |
| Flow | : Normal, Medium | Colour | : None |
| Oil | : | Odour | : None |
| Bould/Cobb | : | Clarity | : Clear |
| Pebb/Grav | : | Algae | : Present |
| Sand | : | Macrophytes | : - |
| Silt | : | Sew Fungus | : - |
| Substrate | : Rocks, Stones, Gravel, Sand | | |
| Reason | : | Saline | : |
| Lane use | : | Shade | : None |

| Taxa Group | Taxa Recorded | Abundance |
|---|---|---|
| Oligochaeta | Tubificidae | Few |
| --- | Sialidae | Present |
| Mollusca | Valvata sp. | Present |
| | Potamopyrgus jenkinsi | Common |
| | Lymnaeidae | Present |
| | Ancylidae | Present |
| | Sphaeriidae | Few |
| Coleoptera | Dytiscidae | Present |
| | Gyrinidae | Few |
| Ephemeroptera | Baetidae | Present |
| | Caenidae | Present |
| Diptera | Dicranota sp. | Present |
| | Chironomidae | Few |
| | Simuliidae | Few |
| Malacostraca | Asellus aquaticus | Few |
| | Gammarus pulex | Common |
| Hirudinea | Glossiphoniidae | Present |
| | Glossiphonia complanata | Few |
| | Helobdella stagnalis | Present |
| | Erpobdellidae | Few |
| Trichoptera | Leptoceridae | Present |

| | | | |
|---|---|---|---|
| Total Taxa | : 21 | BMWP Score | : 81 |
| BMWP Taxa | : 19 | ASPT | : 4.26 |

Figure 4.7: A typical sample from the Severn-Trent database.

| Class | Freq. | Percentage |
|---|---|---|
| B1a | 14 | 6.8% |
| B1b | 67 | 32.7% |
| B2 | 104 | 50.7% |
| B3 | 19 | 9.3% |
| B4 | 1 | 0.5% |

Table 4.4: Classification of 205 sample Severn-Trent database.

randomised and every sixth sample was selected, the result being a set of 205 samples. At the start of the project it was envisaged that approximately 300 samples would be sufficient for training the network models, however in the light of progress made in neural network studies this now appears to be on the small size. The 205 samples were classified by the Expert into five biological classes, as described in Section 4.2.4. Table 4.4 shows the distribution of the classes of this 205 sample data set. The percentages in the third column are the best estimates of the prior probabilities of each class occurring in the Upper-Trent catchment, based upon the available data. It can be observed from this table that the distribution of the classes was biased heavily towards those classes of intermediate quality. It was felt that this base set of 205 should be augmented with additional samples from the remaining data to diminish problems that could arise from the under representation of some classes.

The additional samples were chosen heuristically in two ways: one to select more good quality samples (B1a and B1b) and the other was designed to select more poorer quality samples (B3 and B4). The two criteria were respectively:

- TBI >= 9 and ASPT >= 6.0 for the good quality samples, and
- TBI <= 3 and ASPT < 2.5 for the poor quality samples.

The extra samples made the total data set up to 293 samples. Closer examination of this data revealed that there was one particularly unusual sample. This sample contained a single taxon, namely Limnephilidae, and was classified as B2 by the Expert. As the Expert considered this sample to be very unusual and unrepresentative it was decided to remove it from the database. This left a database of 292 classified samples for use in the experimental work,

| Class | Freq. | Percentage |
|---|---|---|
| B1a | 58 | 19.9% |
| B1b | 71 | 24.3% |
| B2 | 103 | 35.3% |
| B3 | 35 | 12.0% |
| B4 | 25 | 8.6% |

Table 4.5: Classification of 292 sample Severn-Trent database.

and Table 4.5 shows the frequency of the classes within this data set.

A map showing the location of the 292 samples is given in Figure 4.8. The samples constitute a good mix of upland and lowland sites, and are representative of the whole of the Upper Trent Region. Also, both quality extremes are represented; for example the good quality streams of the upland catchments (e.g. R. Dove and R. Churnet) as well as the poorer quality ones of the industrial West Midlands (e.g. R. Tame). Perhaps the most important aspect of the Severn-Trent data was that the Expert was familiar with the area, as he had spent most of his career working in the Severn-Trent region [168].

The Expert expressed some concern about classifying some of the Severn-Trent samples, since the data did not strictly meet the riffle biotope assumptions. The uncertainty concerning the biotope was unavoidable as the NRA's present sampling policy is to sample all available biotopes within a site.

The new system of thirteen classes still used the original five class system, but samples were either unaltered or adjusted up(+) or down(−) within their original class as appropriate (see Figure 4.1). The best quality class was still a B1a, as the Expert could not conceive of anything better than a B1a, while the poorest class was still B4, as likewise the Expert could not picture a sample of poorer quality classification than a B4. The reliability of these finer classifications is clearly lower than the original coarser one, as is indicated by the fact that the new groups (+,−) are under represented in the data set (Figure 4.9). Of the 13 classes, 8 are suffixed by + or −, but of the 292 samples only 95, less than a third, fall into these bands. A reason for this disparity was the method used to obtain the extra classifications. The Expert originally classified the 205 samples to five classes. When the extra samples were added these were

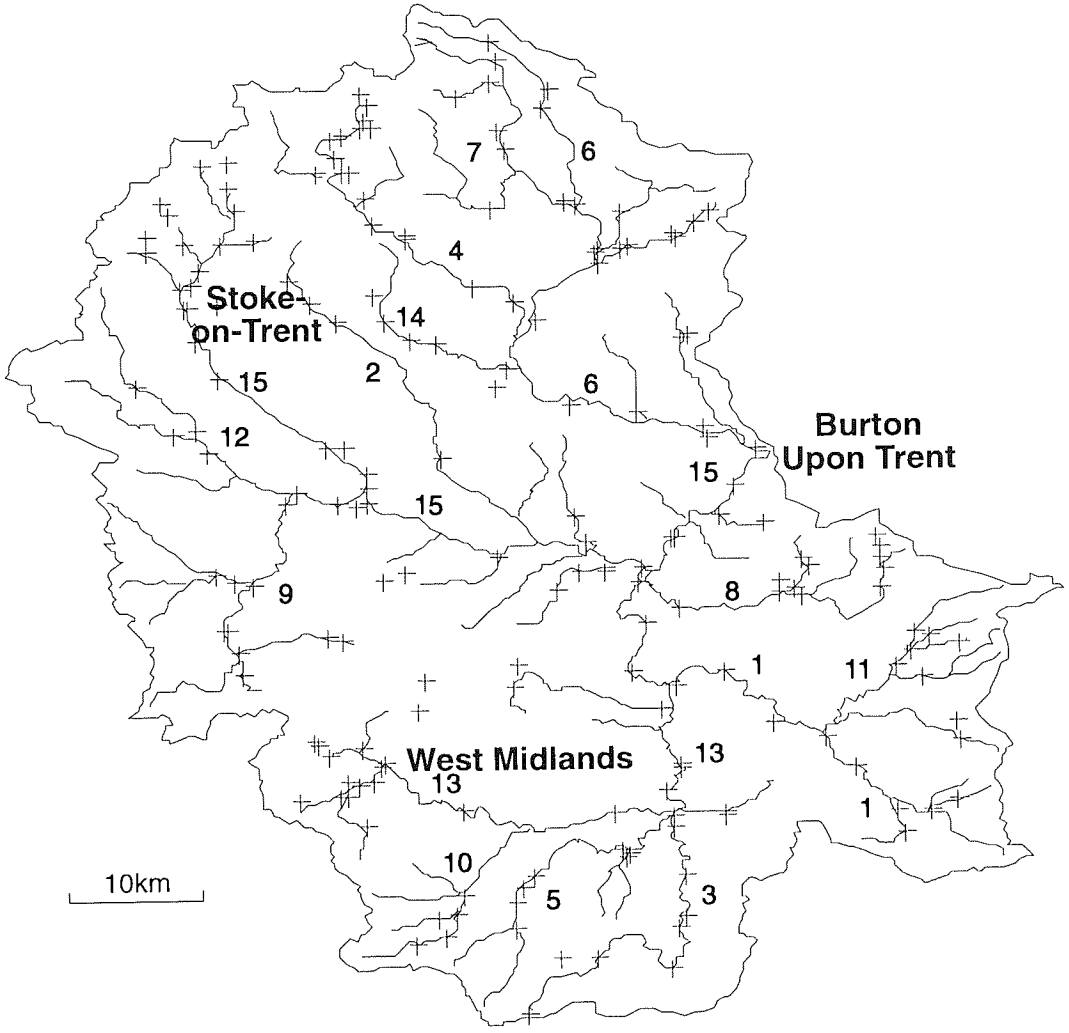



Figure 4.8: Map showing the distribution of the 292 sample database taken from the Upper Trent Catchment of the NRA Severn-Trent Region.
Key: 1) R. Anker, 2) R. Blithe, 3) R. Blythe, 4) R. Churnet, 5) R. Cole, 6) R. Dove, 7) R. Manifold, 8) R. Mease, 9) R. Penk, 10) R. Rea, 11) R. Seance, 12) R. Sow, 13) R. Tame, 14) R. Tean, 15) R. Trent.

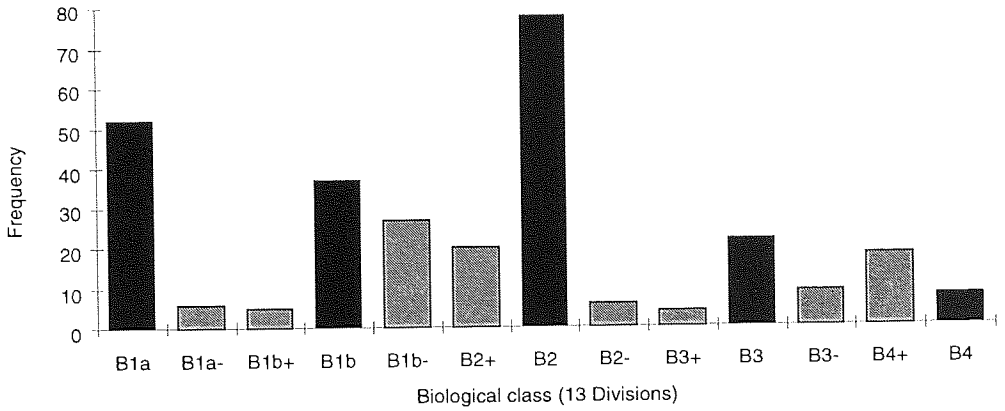



Figure 4.9: Histograms showing the frequency of the biological water quality classes (13 divisions) in the Severn-Trent database.

classified to the 13 class scale. The first 205 were then classified to the new 13 classes, with the original class available to the Expert. With this set the Expert adjusted only a small percentage of the samples. If the exercise were to be repeated then the samples would be classified to the 13 classes from the start. Despite these weaknesses the finer classifications represent a gain in information relative to the coarser classifications.

### 4.3.3 Confirmation Tests

An important question to be asked concerns the consistency and precision of the Expert's classifications, as the Expert himself is a source of error. Ideally, the Expert would be consistent from session to session with respect to his interpretation of the quality class of a given sample. This is, however, unlikely to occur in practice. In order to assess the Expert's consistency a confirmation exercise was conducted. Using a set of test data from a neural network run, all samples that differed from the Expert's classification by more than one division were listed. This gave a set of 22 samples. Six small groups of samples were formed and the Expert was asked to place the samples in each group into quality order, and then to classify them. The original classifications were withheld from the Expert until the completion of the exercise.

The results of the confirmation exercise, Table 4.6, are interesting for a number of reasons. The largest difference between the original classification

| | Expert (Orig.) | | Expert (New) | | Comments |
|---|---|---|---|---|---|
| | Order | Class | Order | Class | |
| **Group 1** | | | | | |
| 76084 | 1 | B1b | 1 | B1b | A good B1b |
| 74700 | 1= | B1b | 2 | B1b- | *Last two are |
| 75395 | 3 | B2+ | 3 | B2+ | * closest |
| **Group 2** | | | | | |
| 74743 | 3 | B2+ | 1 | B1b | |
| 74941 | 1= | B1b- | 2= | B2+ | * This is slightly |
| 75005 | 1= | B1b- | 2= | B2+ | * better than next |
| **Group 3** | | | | | |
| 76023 | 3 | B3- | 1 | B3 | |
| 75616 | 4 | B4 | 2 | B3- | * These are all quite close |
| 74897 | 1= | B3 | 3 | B4+ | * Presence of *Limnophora* |
| 75108 | 1= | B3 | 4 | B4 | * in 74897 is significant |
| **Group 4** | | | | | |
| 75881 | 2= | B2 | 1= | B2 | * Little difference |
| 75532 | 1 | B2+ | 1= | B2 | * between these |
| 75340 | 2= | B2 | 3 | B3+ | |
| **Group 5** | | | | | |
| 75334 | 1 | B1b- | 1 | B2+ | |
| 75072 | 2= | B2 | 2= | B2 | |
| 75790 | 2= | B2 | 2= | B2 | |
| 75935 | 2= | B2 | 2= | B2 | *C. riparius* present |
| **Group 6** | | | | | |
| 74710 | 4= | B3+ | 1 | B2- | |
| 75610 | 1= | B2- | 2= | B2/B3 | |
| 74829 | 1= | B2- | 2= | B2/B3 | |
| 75541 | 4= | B3+ | 2= | B2/B3 | |
| 76020 | 1= | B2- | 5 | B3 | Clearly the worst |

Table 4.6: Results of confirmation tests, with additional explainatory comments from the Expert.

and the amended class is two gradations on the thirteen class (minor) scale, which is quite acceptable. It is also worth noting that the Expert was much more at ease ranking the list of samples in a set than he was classifying them individually. The ranking of the samples is more in keeping with his usual methods of interpretation. The Expert was definitely influenced by the abundance of some taxa, and commented to this effect frequently, especially when finer differences between samples were being considered. The problems of inconsistent levels of identification of taxa was also commented upon, a particular case being Chironomidae and *Chironomus riparius*. A sample containing *C. riparius* was considered to be of relatively poorer quality than similar samples containing Chironomidae. This is because *C. riparius* is more tolerant of pollution than the other members of its family, so identification to species level in this case provides much more specific information. Additionally, some taxa not within the 41 key groups influenced the Expert's opinion, especially in the better quality samples where the diversity of the species lists was implicitly taken into account in this classification.

### 4.3.4 Comparison of Biological Classification with the BMWP Score, ASPT, TBI and Number of Taxa

This section compares the biological classification, based on the NWC, to the BMWP score, ASPT and TBI systems that are commonly used within the water industry. In addition to these, the 'Number of Taxa' in a sample was also studied, to give an indication of diversity within different biological classes. Due to the absence of an absolute standard the biological classification was taken to be the reference standard. Justification for favouring the Expert classification, as pointed out by Walley [168], comes from the fact that the Expert (H.A. Hawkes) was the chairman of the working sub-group of the Biological Monitoring Working Party (BMWP), so is very well acquainted with both the BMWP score and ASPT. He also spent most of his career working in the Trent region, hence he is familiar with the TBI. The Expert was thus not overly biased towards any of these systems, so none should be particularly disadvantaged.

Grouping the 292 samples by the five biological classes, the maximum, minimum, mean and standard deviation of BMWP score, ASPT, TBI and Number of Taxa (NOT) were calculated (Figure 4.10). To summarise the graphs:

**BMWP Score**

The mean descends from class B1a to B5, however there is a large overlap between the B1a, B1b and B2 classes. The BMWP score discriminates poorly between the lower quality B3 and B4 classes, and has a particular high variance for the B2 class, which covers over 80% of the range of the BMWP scores for this particular data set. The classes of good-to-intermediate quality (B1a-B2) show the highest ranges. The resulting plot is typical of the BMWP score because of its cumulative effect, and higher weighting of sensitive taxa.

**ASPT**

Again, the mean value of the ASPT for each class decreases with poorer biological quality class. There is better discrimination between the five biological classes, but there is still an appreciable overlap, albeit much smaller than that of the BMWP score. There is a better range of values for the poorer quality classes (B3,B4), however it would be difficult to discriminate between the two classes when given an ASPT of 2 to 2.5. Like the BMWP score, the range of the intermediate classes are larger than those at either end of the quality spectrum, but the standard deviations of the ASPT with respect to biological class are smaller and more consistent than those of the BMWP score. From these graphs it is possible to say that the ASPT is a more reliable indicator of organic pollution than the BMWP score.

**TBI**

Considering only the B1a class, it can be seen that there is good agreement between the TBI and the Expert. All of the samples classified as B1a had either a TBI of 9 or 10. For the poorer quality classes, again, the B2 has the highest range. Also there is some overlap between the B1b/B2 and the B3/B4 classes.

**Number of Taxa**

This is not usually used for classification purposes, but it does provide a feeling for the diversity of the sample, although care must be taken because of the methods used by the Severn-Trent sampling and identification procedures. Again, the mean of scores for each class decrease with respect to increasing

organic pollution, but both the B1b and the B2 have a higher number of taxa than the better quality B1a class. This confirms the fact that diversity increases with small amounts of organic pollution prior to decreasing with larger amounts. There is again a large overlap between the B1a, B1b and B2 classes.

To a certain extent these results are not surprising. The biologoical classification is based on the Expert's assessment of organic pollution, and as the ASPT and TBI were designed to relate to organic pollution this explains the resonable agreement between these and the biological classification. The BMWP and NOT relate not only to organic pollution but also to environmental stresses, hence they do not match the Expert as closely as the ASPT and TBI.

It appeared during the elicitation sessions that there was a similar line of reasoning being taken by the Expert to that of the TBI, but it was not known whether the Expert had been conditioned into this by prior use of the TBI. It appeared that the overall classification was decided upon by looking for the sensitive species that were present, these were usually only a small subset of the whole set of taxa, and an approximate benchmark was set depending on the outcome of this search. Then by further examination of the other taxa which were present the benchmark was either adjusted up or down according to diversity of the sample or other factors which the Expert considered important. It should be noted that the TBI was developed for rivers within the Midlands, which is also the region most familiar to the Expert and also the source region of the data. Thus it would be fair to say that good agreement could be expected between the two.

A similar analysis using the 13 class scale results in slightly different interpretation, Figure 4.11. The most noticeable difference is that the mean values of the indices for the classes no longer descend uniformly; these graphs can be used for the justification of the increasing diversity associated with a riffle site (see Figure 4.6). The results are representative, but not conclusive because of the small sample size of some of the classes (see Figure 4.9 and associated text).

Figure 4.12 shows the relationship between BMWP score and ASPT, with the biological class of each point also denoted. The high variance of the BMWP score is apparent for the better quality sites. The main feature of the graph is

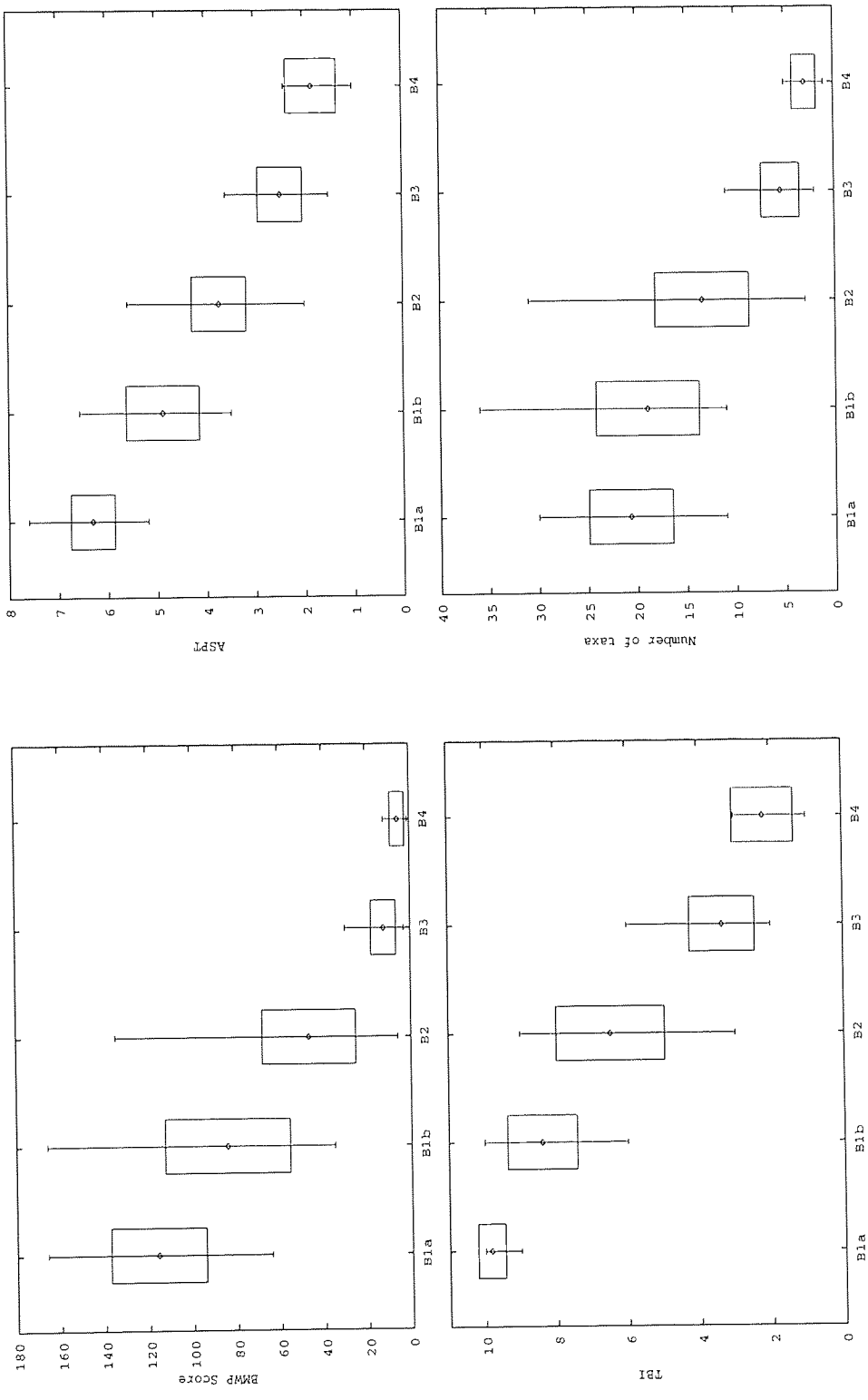



Figure 4.10: Summary of BMWP Score, ASPT, TBI and 'Number of Taxa' in terms of five biological classes. The mean, maximum, minimum and standard deviations are shown for each class.

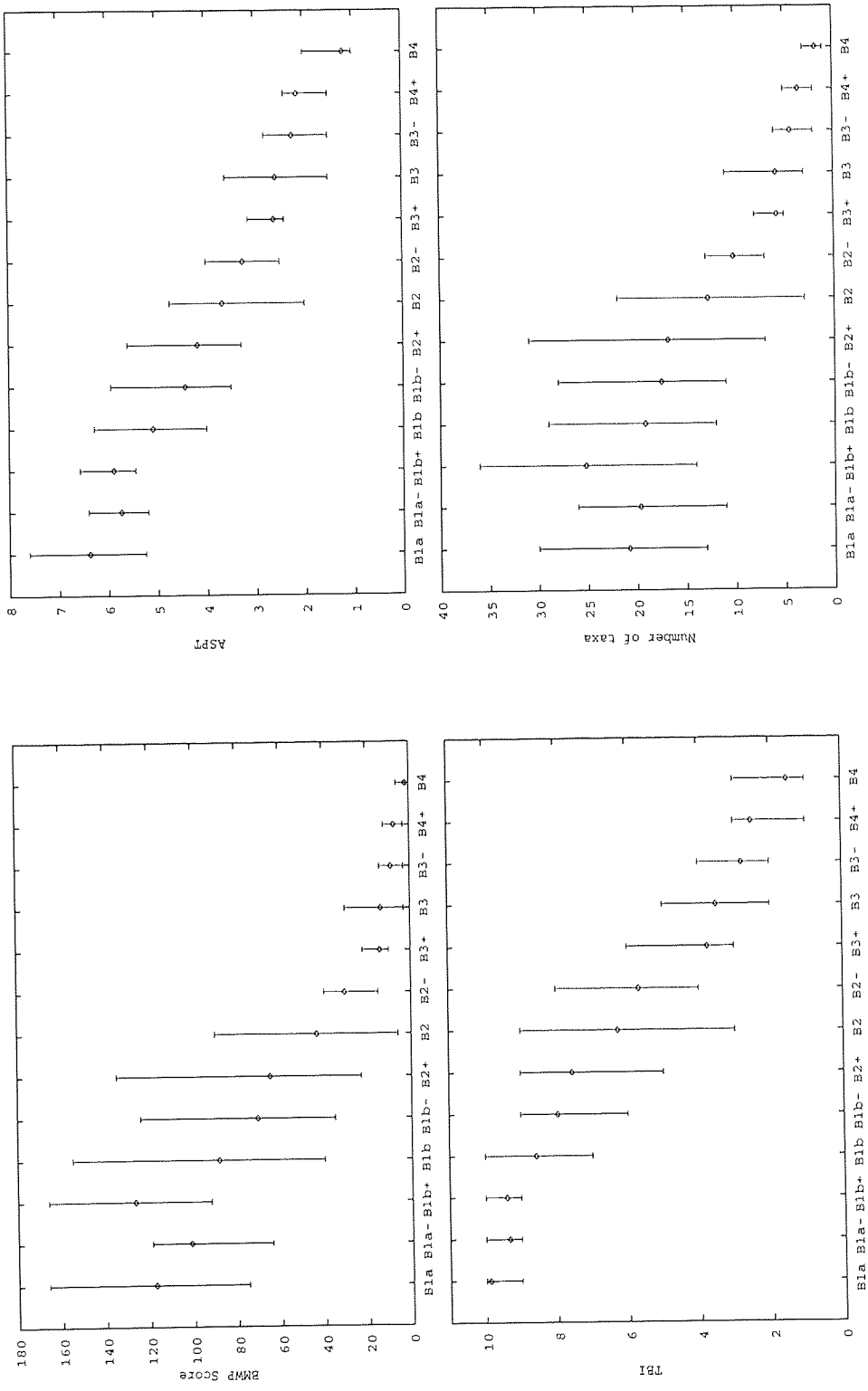



Figure 4.11: Summary of BMWP Score, ASPT, TBI and 'Number of Taxa' in terms of thirteen biological classes. The mean, maximum and minimum are shown for each class.

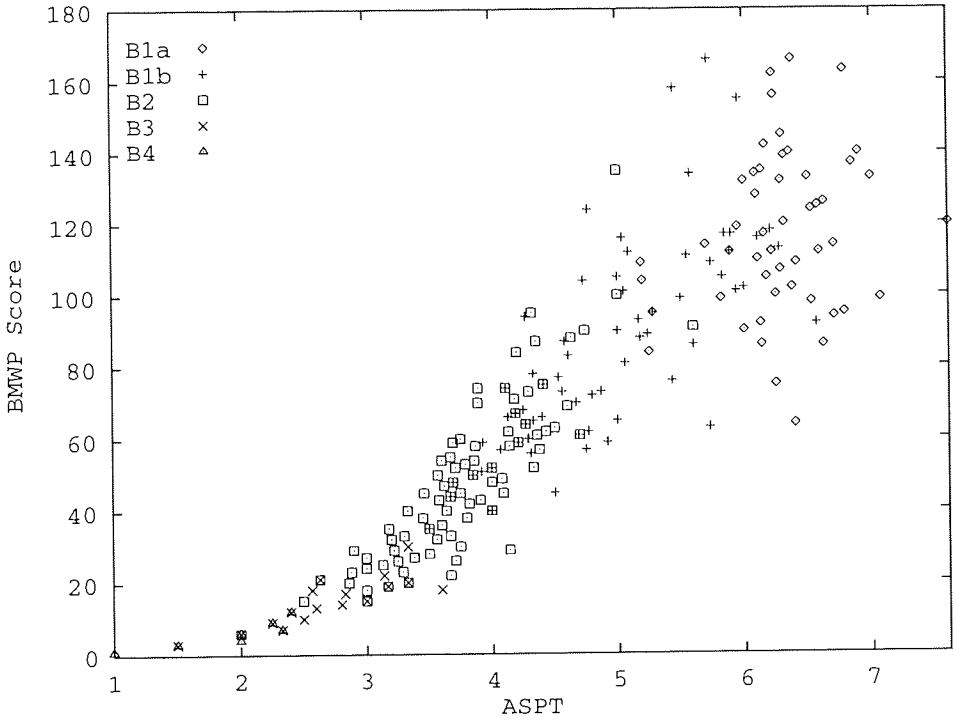



Figure 4.12: A plot of BMWP score against ASPT showing biological quality class.

that there is high overlap of the samples with respect to the biological classes. This implies that it would be difficult to predict the biological class if given just the ASPT and the BMWP score for a particular sample.

### 4.3.5 Frequency of the Original BERT Taxa

When the Severn-Trent data set was obtained and the preliminary analysis conducted there appeared to be a distinct difference between the assumptions made in the BERT elicitation sessions and the Severn-Trent with regard to the frequency of the taxa. It appeared that the frequency of the BERT taxa in the 292 data was lower than expected. To examine this further, the frequency of the original BERT taxa (Table 4.1) was calculated, along with a modified set, given in Table 4.7.

From Table 4.7 it is apparent that a few taxa are particularly infrequent, for example *Ancylus fluviatilis* and *Baetis rhodani*. The frequencies of species in the 292 Severn-Trent database are much lower than corresponding ones

| Original Taxa | Frequency | Modified Taxa | Frequency |
|---|---|---|---|
| Polycelis nigra | 9 | Polycelis spp. | 33 |
| Dendrocoelum lacteum | 13 | Dendrocoelidae | 16 |
| Potamopyrgus jenkinsi | 128 | Potamopyrgus jenkinsi | 128 |
| Bithynia tentaculata | 0 | Bithynia spp. | 12 |
| Lymnaea peregra | 103 | Lymnaeidae | 163 |
| Planorbis spp. | 0 | Planorbidae | 54 |
| Ancylus fluviatilis | 16 | Ancylidae | 130 |
| | | Sphaeriidae | 161 |
| Sphaerium spp. | 44 | Sphaerium spp. | 44 |
| Pisidium spp. | 19 | Pisidium spp. | 19 |
| Tubificidae | 273 | Tubificidae | 273 |
| Lumbriculidae | 26 | Lumbriculidae | 26 |
| Glossiphonia spp. | 95 | Glossiphonia spp. | 95 |
| Helobdella stagnalis | 31 | Helobdella stagnalis | 31 |
| Erpobdella octoculata | 99 | Erpobdellidae | 154 |
| Hydracarina | 85 | Hydracarina | 85 |
| Asellus aquaticus | 170 | Asellus aquaticus | 170 |
| Gammarus pulex | 173 | Gammarus pulex | 173 |
| Baetis rhodani | 3 | Baetidae | 153 |
| Rhithrogena spp. | 8 | Heptageniidae | 75 |
| Heptagenia spp. | 2 | | |
| Ecdyonurus spp. | 34 | Ecdyonurus spp. | 34 |
| Ephemerella ignita | 30 | Ephemerella ignita | 30 |
| Caenis spp. | 50 | Caenis spp. | 50 |
| Amphinemura sulcicollis | 0 | Nemouridae | 39 |
| Leuctra spp. | 9 | Leuctridae | 47 |
| Isoperla grammatica | 31 | Perlodidae | 57 |
| Haliplidae | 50 | Haliplidae | 50 |
| Dytiscidae | 115 | Dytiscidae | 115 |
| Elminthidae | 89 | Elminthidae | 89 |
| Sialis lutaria | 8 | Sialis lutaria | 8 |
| Rhyacophila dorsalis | 17 | Rhyacophilidae | 67 |
| Glossosoma spp. | 7 | Glossosomatidae | 9 |
| Agapetus spp. | 2 | | |
| Polycentropidae | 26 | Polycentropidae | 26 |
| Hydroptilidae | 4 | Hydroptilidae | 4 |
| Hydropsyche angustipennis | 2 | Hydropsyche angustipennis | 2 |
| Other Hydropsychidae | 110 | Other Hydropsychidae | 110 |
| Limnephilidae | 87 | Limnephilidae | 87 |
| Ceratopogonidae | 9 | Ceratopogonidae | 8 |
| Chironomus riparius | 21 | Chironomus riparius | 21 |
| Simulium ornatum | 0 | Simuliidae | 125 |
| | | Atherix ibis | 30 |

Table 4.7: Comparison of the frequency of BERT taxa with a modified set of taxa.

reported in the data set by Wright et al. [181]. There are some differences between the two sets of data, notably the reliance of the sample being drawn from 'good quality' sites for the Wright et al. data. But the most probable reason for the disparity between the frequencies is the effort put into identifying the animals. Perhaps the biggest criticism of the BMWP score is that it only requires family level identification, a good example of this is shown by *Ancylus*. There are two aquatic species of Ancylidae occurring in the UK, and it would be a fair assumption that all the occurrences of Ancylidae were *Ancylus fluviatilis*. Thus the relatively small effort required to generate a BMWP score for a sample is reflected in the taxonomic lists of 292 Severn-Trent data, although other constraints, typically time and money, hinder the biologists work.

In order to investigate the effect of the under representation of BERT taxa in the 292 data, a simple index was developed. This index, the 'union value', was simply the number of taxa in a sample which were BERT taxa divided by the total number of taxa in the sample. Thus a union value of 0.0 meant that no taxa in the sample were common to the BERT taxa, while a value of 1.0 represented complete coverage. Ideally, the higher the union value for a sample the better, as this means that more of the taxa would be utilised for classification within the model. The union value was calculated for each sample, and the summary of mean and standard deviation of union value for the whole data set and each of the five individual classes in given in Figure 4.13. Four sets of taxa were considered, these being the original BERT taxa, the modified set (Table 4.7), and both the BERT taxa and the modified set taken to family level.

There are two distinct trends in this figure. The first trend is that between the different groups of taxa. The mean of the union values are consistently ordered, with lowest being the BERT taxa, then the modified taxa, the BERT taxa families to, finally, the modified taxa to families. This would be the predicted result, as the modified set were chosen with the Severn-Trent data in mind, and that the families levels are more frequent than the mixed (frequently species) level of the BERT and modified taxa. The second trend is with respect to the quality classes, with the union value increasing as the quality decreases. The implications of this trend is that a higher number of taxa are being ignored in the good quality classes, and consequently their evidence is not being utilised

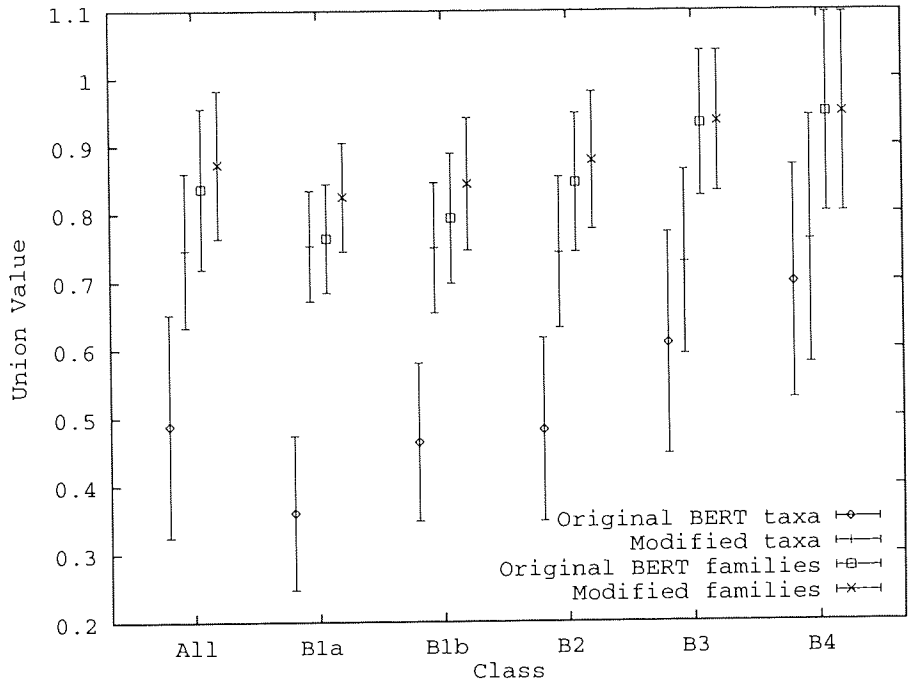



Figure 4.13: Mean and standard deviation of union value based on class for Severn-Trent database.

to form the classification. So using the BERT taxa with the 292 data set, would lead to less than 50% of the taxa in a B1a class being used, and would hinder any resulting classification. Thus for the neural network experimentation the modified taxa were used, while the family level data was used for the work on indicator taxa (Chapter 6).

### 4.3.6 Recapitulation

The Severn-Trent data are typical of data arising from most biological monitoring programmes, but it is a unique data set. This is the first time that an invertebrate data set has been classified in such a manner by a domain expert, and it is also unusual for any data, yet alone invertebrate data, to have accompanying probabilistic domain knowledge as well.

## 4.4 Other River Data

### 4.4.1 Yorkshire Water Authority Data

The first data set to be studied in the project originated from the Yorkshire Water Authority, sampled in the early 1970s. It included invertebrate species lists, to a mixed level of identification, the names of the water course from which the samples were taken and the dates of sampling. The data were just of sufficient quality to be usable, but no better than that. It was the only data set available at the start of the project and it formed the basis of some of the early experimental work, which was reported by Ruck et al. [148].

This paper was the first work to describe the use of neural network models for the direct classification of river water quality from benthic macroinvertebrate data. Two sets of data were used, one consisting of the raw input data, the other a set of principal component transformations. The models were trained using leave-one-out cross validation, and were found to correctly classify the testing data to just under 70%. These results were promising considering the small size and the relatively poor quality of data. This work should be viewed as the pilot study to the more extensive tests reported in this dissertation.

### 4.4.2 Synthetic Data

This section describes work on a set of synthetic data based on the conditional probabilities derived from the direct elicitation sessions (Section 4.2.3). These conditional probabilities describe, as fully as possible, the distribution of the invertebrate taxa with respect to the biological water quality class, and by working backwards it is possible to generate representative invertebrate samples from these probabilities.

There were two main objectives behind the creation of the synthetic data set. The first was to enable some prior knowledge to be used in the training of the networks (see Section 5.6). The second was to allow for some larger scale neural network experimental work to be conducted, to augment the study of the Severn-Trent data. In essence, the synthetic data provided a means to test various methodologies as applied to freshwater biomonitoring, which would

not have been possible with the Severn-Trent data.

#### 4.4.2.1 Sample Generation from the Conditional Probabilities

To generate a representative sample the conditional probabilities $P(e_{ik}|H_j)$ needed to be 'conditioned' with the 'known' distribution of the quality classes ($H_j$'s). The $P(e_{ik}|H_j)$ represent the probability of finding the $i$th taxon in the $k$th state given the $H_j$th quality class. Thus, by specifying the desired distribution of biological classes the probability of the taxon being *absent*, *rare*, *established* and *abundant* can be calculated. The resulting distribution was sampled to select the state of the given taxon to be included in the sample.

The $P(e_{ik}|H_j)$ values were taken from the histograms that were originally elicited for the *absent*, *established* and *abundant* states. The distribution for *rare* was arbitarily taken to be 40% of that of *established*, with the probabilities re-normalised to ensure that they summed to unity.

The conditioned $P(e_{ik}|H_j = C)$ distribution, where $C$ is the quality class, for each taxon can then be found from:

$$P(e_{ik}|H_j = C) = \sum_j P(e_{ik}|H_j)P(H_j) \tag{4.1}$$

where the $P(H_j)$ are set to reflect the desired class $C$. If the desired class was a B2, then the $P(H_j)$'s would be as follows: $P(H_1) = 0$, $P(H_2) = 0$, $P(H_3) = 1$, $P(H_4) = 0$, and $P(H_5) = 0$. The above formula also allows for interpolation between classes. For example, if the desired class was a B1b+, then the $P(H_j)$'s could be described as follows: $P(H_1) = 0.33$, $P(H_2) = 0.67$, $P(H_3) = 0$, $P(H_4) = 0$, and $P(H_5) = 0$.

To aid understanding a brief example will be worked through. Table 4.8 shows the elicited $P(e_{ik}|H_j)$ values for *Gammarus pulex*. To find the probability of each state, $P(e_{ik})$, for *G. pulex* in a sample of, say for example, a B1b+ class we use Equ. 4.1. The result is that *G. pulex* has a 9.3% chance of being found *absent*, 20.0% *rare*, 25.9% *established* and 44.8% *abundant*. This distribution can then be sampled by generating a number between 0.0 and 1.0 using a uniform random number generator and picking the appropriate state, which is graphically depicted in Figure 4.14.

| Class | Abs | Rare | Estb. | Abund. |
|---|---|---|---|---|
| B1a | 0.18 | 0.24 | 0.38 | 0.20 |
| B1b | 0.05 | 0.18 | 0.20 | 0.57 |
| B2 | 0.30 | 0.20 | 0.26 | 0.23 |
| B3 | 0.81 | 0.12 | 0.04 | 0.03 |
| B4 | 0.89 | 0.11 | 0.00 | 0.00 |

Table 4.8: Conditional probabilities, $P(e_{ik}|H_j)$, for *Gammarus pulex* for generation of synthetic data.

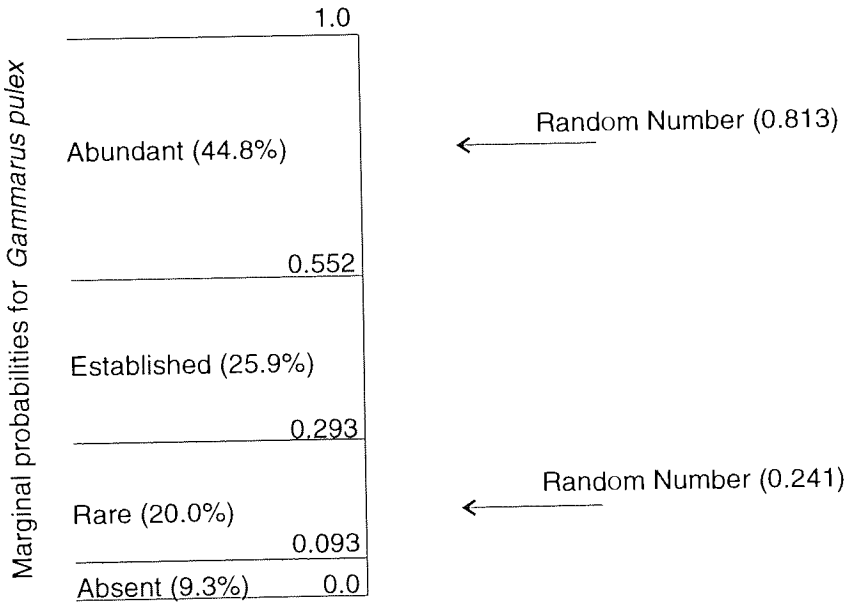



Figure 4.14: Sampling of conditioned distributions, $P(e_{ik}|H_j = B1b+)$, for *Gammarus pulex* in a B1b+ quality class. The values of $P(e_{ik}|H_j = B1b+)$ for each state ($k$) are given in brackets, with the cumulative mass to the right of these. Two example random numbers are shown. The first, 0.813, would generate a sample with *G. pulex* abundant; the second, 0.241, would give *G. pulex* rare.

| Taxon | Riffle | Pool | Taxon | Riffle | Pool |
|---|---|---|---|---|---|
| Polycelis nigra | √ | √ | Ephemerella ignita | √ | √ |
| Dendrocoelum lacteum | √ | √ | Caenis spp. | √ | √ |
| Potamopyrgus jenkinsi | √ | √ | Amphinemura sulcicollis | √ | |
| Bithynia tentaculata | √ | √ | Leuctra spp. | √ | |
| Lymnaea peregra | √ | √ | Isoperla grammatica | √ | |
| Planorbis spp. | √ | √ | Haliplidae | √ | √ |
| Ancylus fluviatilis | √ | | Dytiscidae | √ | √ |
| Sphaerium spp. | √ | √ | Elminthidae | √ | |
| Pisidium spp. | √ | √ | Sialis lutaria | √ | √ |
| Tubificidae | √ | √ | Rhyacophila dorsalis | √ | |
| Lumbriculidae | √ | √ | Glossosoma spp. | √ | |
| Glossiphonia spp. | √ | √ | Agapetus spp. | √ | |
| Helobdella stagnalis | √ | √ | Polycentropidae | √ | √ |
| Erpobdella octoculata | √ | √ | Hydropsyche angustipennis | √ | |
| Hydracarina | √ | √ | Other Hydropsychidae | √ | |
| Asellus aquaticus | √ | √ | Hydroptilidae | √ | √ |
| Gammarus pulex | √ | √ | Limnephilidae | √ | √ |
| Baetis rhodani | √ | | Ceratopogonidae | √ | √ |
| Rhithrogena spp. | √ | | Chironomus riparius | √ | √ |
| Heptagenia spp. | √ | | Simulium ornatum | √ | |
| Ecdyonurus spp. | √ | | | | |

Table 4.9: Occurrence of taxon in pools and riffles.

This is repeated for all the taxa in the species list using different random numbers to create a synthetic sample. This process can then be used as many times as desired to create a series of samples, representative of riffles, for any of the quality classes. The database created for the neural network experimentation contained 5000 samples in total, with 1000 being drawn from each of the five quality classes.

A second data set, again comprising 5000 samples, was generated using the above method to model pool biotopes. As no knowledge elicitation (in the form of histograms) was available for pool biotopes some (sweeping) assumptions were made on the occurrence of the 41 taxa. The ones which were considered to inhabit only riffles were removed from the data (Table 4.9). With guidance from the Expert, sixteen taxa were removed, and this left twenty-five taxa to be included into the pool samples. This was unrealistic as other new species

| Physical Variable | Riffle | Pool |
|---|---|---|
| Froude number | 0.51 ± 0.26 | 0.10 ± 0.10 |
| Velocity/depth ratio | 4.69 ± 3.98 | 0.66 ± 0.83 |
| Velocity (m/s) | 0.62 ± 0.32 | 0.20 ± 0.20 |
| Slope | 0.016 ± 0.01 | 0.004 ± 0.0005 |
| Depth (m) | 0.17 ± 0.12 | 0.39 ± 0.32 |
| Pebble, Boulder, Gravel | 60-100% | 0-30% |
| Sand, Silt | 0-40% | 20-100% |

Table 4.10: Hydraulic characteristics (means and standard deviations) of riffle and pool biotopes, after Jowett [78].

would be found, such as Tipulidae in the Severn-Trent database, but did allow for the creation of a data set that was appreciably different from the synthetic riffle data set.

The BMWP scores and ASPT were calculated on a class by class basis for the two synthetic sets of data, and these are shown in Figure 4.15. Comparing riffles to pools it can be seen that both the BMWP scores and ASPT are higher for riffles than for the pool samples. Both of the BMWP score figures have a similar shape, with higher scores recorded for the B1b samples reducing towards the B4 samples. The BMWP score for the riffle data shows good discriminatory power between the B2, B3 and B4 classes, while there is a greater degree of overlap with the B1a, B1b and B2 graphs. The ASPT, however, does show a good discrimination between the five classes, with no overlap between any of the standard deviations. For the pool samples, the BMWP scores exhibits a similar relationship to that of the riffle sites, but on a more compressed scale, with the ASPT, also mirroring the riffle samples on a reduced scale.

Additionally a set of physical variables were created for both the pool and riffle biotopes. Using figures from Jowett [78] the physical characteristics can be summarised by Table 4.10. These figures were used for the generation of artificial physical variables, using Gaussian functions with mean and standard

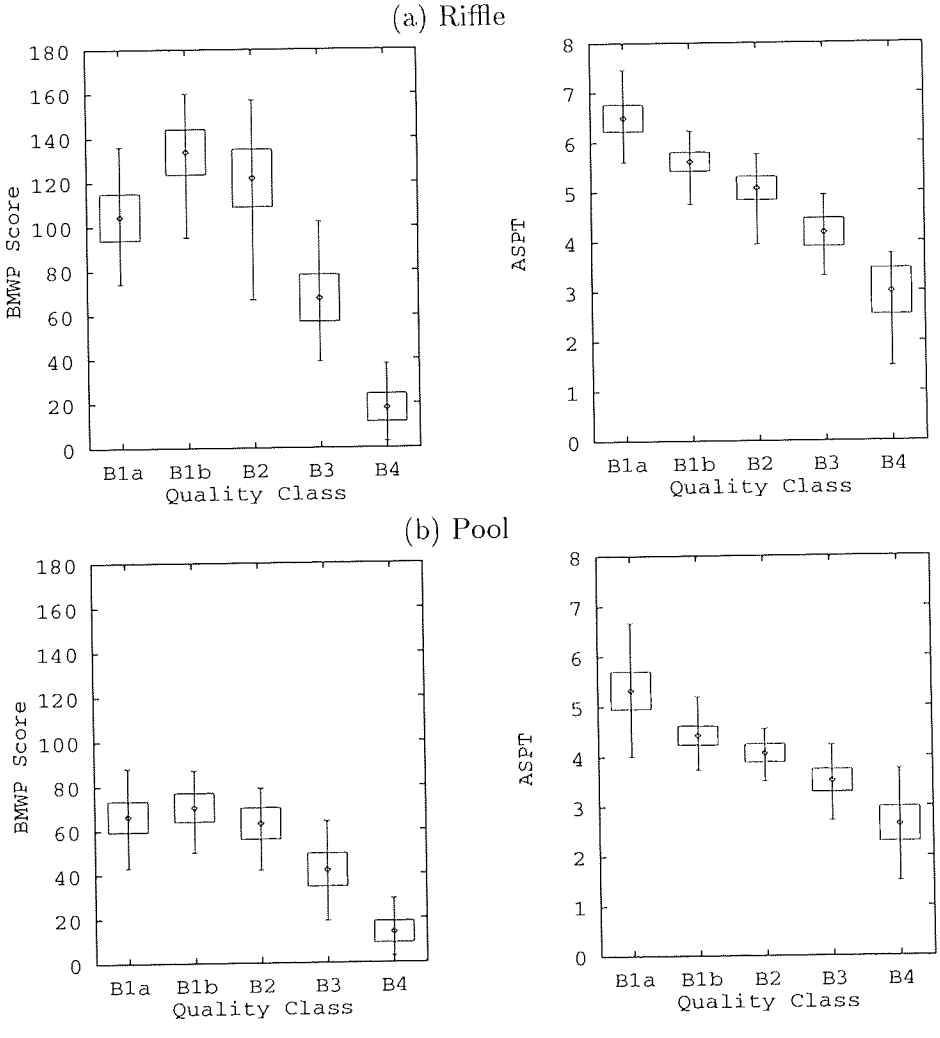



Figure 4.15: Summary of BMWP Score and ASPT for synthetic data. The mean, maximum, minimum and standard deviation are shown for each class.

| Expected Class | Expert's Classification | |
|---|---|---|
| | Riffle | Pool |
| B1a | B1a | B1b |
| B1a | B1a | B2 |
| B1b | B1b | B2 |
| B1b | B1b | B2 |
| B2 | B2 | B2 |
| B2 | B1b | B2 |
| B3 | B2/B3 | B3 |
| B3 | B2 | B3 |
| B4 | B3 | B3/B4 |
| B4 | B4 | B3 |

Table 4.11: Classification of artificial samples. Two sets of 10 samples drawn from different biotopes (with 2 samples from each class) were classified by the Expert.

deviation as given in Table 4.10. Each variable was treated as independent (even though they are obviously dependent in real life), and any values that were less than zero were removed and an alternative value was generated. The physical variables were used for the mixtures of experts experiments (Section 5.4).

#### 4.4.2.2 Expert Classification of Synthetic Data

In order to gather feedback on the artificial samples, the Expert was asked to classify a subset of samples from each of the synthetic data sets. Ten samples were drawn at random from each set, but it was ensured that there were two samples from each class. This gave a total of twenty samples for the Expert to classify, which were randomised prior to being given to the Expert. Prior to the exercise the Expert was not informed that the samples had been drawn from different distributions, he was just asked to classify the samples directly into the biological classes from the species list. The results of the elicitation are shown in Table 4.11.

It was anticipated that the Expert's classifications of the riffle data would compare closely with the expected classifications, but that there would not be such a close comparison in the case of the pool data. Inspection of Table 4.11

indicates that this was indeed the case. There was good agreement between the riffle data and the Expert's classifications, especially in the top quality classes, B1a and B1b. Note that it has been assumed that the Expert was correct since even though a sample may have been generated using probabilities for, say a class B3, the resulting species may have been indicative of a class B2 sample. In contrast to the riffle data, the Expert's classifications for the pool data were consistently on the poor side for the good quality sample.

#### 4.4.2.3 Discussion

The synthetic data were thought to be generally plausible, except for a few samples which, although correctly classified, were deemed to have a slightly artificial 'feel to them'. Most samples that were considered artificial had several species in abundance, this was especially the case for the B1a and B1b classes for the riffle data. Usually it is common to have 3 or 4 species in abundance in sample, but for a couple of the samples this was exceeded. These anomalies can be attributed to the method in which the samples were generated.

The method considered each taxon as independent (i.e. each was considered in isolation) but within the community there would be competition between similar ecologically niched species and this competition would tend to inhibit the abundance of other competitors. This was not taken into account when the samples were generated, and would require knowledge about the community structure interactions for it to be introduced. No autecological information was included in the generation of the synthetic samples, so it was perfectly possible to create a sample that would not occur in nature. If this ecological knowledge were to be built into the method of generating the random samples then a lot of addition knowledge elicitation would have been required. This would have reduced the number of improbable samples that were generated.

During the stage when the Expert was classifying the samples it became apparent that this also provided a good method of direct knowledge elicitation. The original histograms contained probabilistic information from the Expert, but using this method more traditional 'rules' could be elicited. When classifying samples there were occasionally unusual features, mainly concerned with the biotope/sampling problems. For example, "I wouldn't expect to find that in a riffle" was a frequent remark made by the Expert. This occasionally oc-

curred when classifying the Severn-Trent data. With the synthetic samples, however, there were more occasions for the Expert to express his knowledge in terms of the relationship between the different species, for example "I wouldn't expect to find this taxa to be abundant when this one is established".

There was a lot more conflicting evidence within the synthetic samples. A few samples were identified as, for example, B2/B3, and it was apparent that this was different from the traditional B2/B3 classification. If a real sample were to classified B2/B3 then the evidence provided by the taxa present would suggest that the quality is borderline between the B2 and B3 classes, possibly a B2- or a B3+. But with the synthetic data some B2/B3 classifications suggested that some of the evidence was indicating a B2 class, while other evidence was pointing to a B3 class, there was no accord across the whole sample. This is exemplified by a few samples of the 20, where 1 or 2 species were inconsistent with the rest. For example, one of the samples contained a number of good quality indicators as *abundant* and also *Hydropsychidae angustipennis* as *established.* If this were an actual sample then this would be considered conflicting evidence, because *H. angustipennis* needs mild organic pollution for it to be *established.* In this case the evidence provided by *H. angustipennis* was pulling the quality class lower to a poorer level. The opposite effect also occurred when one or two sensitive species were present while the rest of the sample, mainly tolerant species, indicated a higher pollutional load.

As a practical exercise, the classification of the synthetic data by the Expert was beneficial. It should be noted that since the artificiality of the data was apparent, then this may be reflected in the classification rates produced in the experimental work. As a method of knowledge elicitation the classification of synthetic samples did bring out more structured rule based ecological knowledge from the Expert.

### 4.4.3 National Data

This section uses the NRA database of the 1990-92 National Survey. The distribution of BMWP scores and ASPT through all 10 NRA Regions is investigated, as is the distribution of taxa through the 10 regions. The database comprises benthic invertebrate sample records from the years of 1990-92, iden-

| Region | Number of samples | | | Abundance Codings |
|---|---|---|---|---|
| | 1990 | 1991 | 1992 | |
| Anglian | 3030 | 3155 | 1948 | 1-6 |
| Northumbria | 1150 | 1221 | 0 | 1-6 |
| North West | 2060 | 2879 | 987 | 1 |
| Severn-Trent | 2830 | 0 | 0 | 1-6 |
| Southern | 1189 | 1206 | 848 | 1-6 |
| South West | 1485 | 1426 | 1477 | 1-6 |
| Thames | 928 | 975 | 563 | 1-6,11,12 |
| Welsh | 2271 | 1600 | 1584 | 1,6 |
| Wessex | 1117 | 0 | 0 | 1 |
| Yorkshire | 1293 | 1074 | 0 | 1-6 |
| Total | 17353 | 13536 | 7407 | |

Table 4.12: Summary of the National NRA database showing, for each region, the total number of samples and the taxa abundance coding adopted.

tified to family level (see Appendix A2 for the full list of recorded families), and some chemical and physical characteristics of the sampling sites. Unfortunately, the data were only available late on in the project, and due to time restrictions only a limited study was undertaken. It was not possible to compare the chemical or biological classifications on a site-by-site basis, and no attempt was made to investigate any relationship between them. This in itself would have been a long term research project.

A preliminary data analysis showed a number of inconsistencies within the database, mainly between the chemical and biological sampling records. Due to the time restrictions all suspect samples were ignored. This, in hindsight, was possibly a little too 'ruthless' but did not significantly reduce the number of samples available. The total number of samples which were used in the analysis is shown in Table 4.12. Three years of samples were contained in the database, but as Table 4.12 shows not all the regions were represented in all years, thus it was decided to concentrate solely the 1990 samples. A further problem was that there was inconsistency between the regions on the abundance codings that were used, with some regions having 6 levels (which could be mapped to the values used in the Severn-Trent database, Table 4.3)


Table 4.13: River quality in 1990 by NRA region [112].

while others only recorded the taxa as absent/present. The lowest common denominator was used, which meant that all taxa records were reduced to absent/present for the following analysis.

#### 4.4.3.1 Comparison of Scores and NWC Classification

In this section the BMWP score and ASPT are compared to the NWC classification on a region-by-region basis using the national database and the chemically-based results of the 1990 survey [112]. The main assumption is that the biological samples are fully representative of each region, and that all the regional variations are adequately covered. This is important as the statistics are collected together for each region and not gathered and compared on a site by site basis. Table 4.13 shows the summary of the percentage of river length in each NWC class results for all of the 10 NRA regions. There is a large variation between the regions for the percentage length of class 1a's, but the percentage of 'good & fair' (1a, 1b and 2) are all above 80% apart for North West. The percentage of 'poor & bad' are all below 10% except for North West, Severn-Trent, South West and Yorkshire.

Figures 4.16 and 4.17 show the distribution of the BMWP score and ASPT, derived from the 1990 national database samples for each of the 10 regions. Looking at the BMWP scores (Figure 4.16) it is apparent that there are

large differences in the distribution of the BMWP scores between the regions. The South-West region has the only distribution which is skewed to the right (i.e. higher BMWP scores), while both Anglian and Severn-Trent regions are strongly skewed to towards the lower BMWP scores. The South-West region results are unusual in that, from the histograms, it would appear that South-West has the best quality streams, but this is not the case when the NWC classification is considered. In this particular case the chemistry and biology are revealing different pictures, with the chemistry classification lower than the biology. For the Anglian region this is reversed, from the BMWP score histograms the biological system is inferring that, comparatively, the region scores poorly, while the chemistry informs us that the rivers are better quality than this.

The ASPT figures are more discriminating than the BMWP scores. Anglian region's ASPT histogram is strongly peaked about 4, with virtually all the ASPT falling between 3 and 5. Recalling Figure 4.10 which showed the variation of ASPT and BMWP score with the biological classification (B1a, ..., B4) the ASPT histograms would reveal that nearly all of the classes are B1b and B2. Similarly for South West the ASPT histogram reveals that the majority of the sites fall between an ASPT of 5 and 7, which would mean that the most (i.e. 85% or more) of the samples occupied classes B1a and B1b, which is conflicting with the information from the chemical classification in which only 52% are 1a or 1b. The South West has the most strongly skewed histogram towards the higher scoring ASPT, but it has the lowest percentage of 1a and 1b's out of the 10 regions, which is contrary to expectations.

The distribution of the ASPT and BMWP scores between the NRA regions is interesting as it shows the tremendous variation, which, if the results of the 1990 national survey are to be relied upon, implies that geography has the major effect on the variation as opposed to water quality (which is approximately equivalent for all regions). The figures demonstrate the extent to which the variation is prevalent, and to investigate this further, the distributions of a number of taxa was calculated for each of the 10 regions.

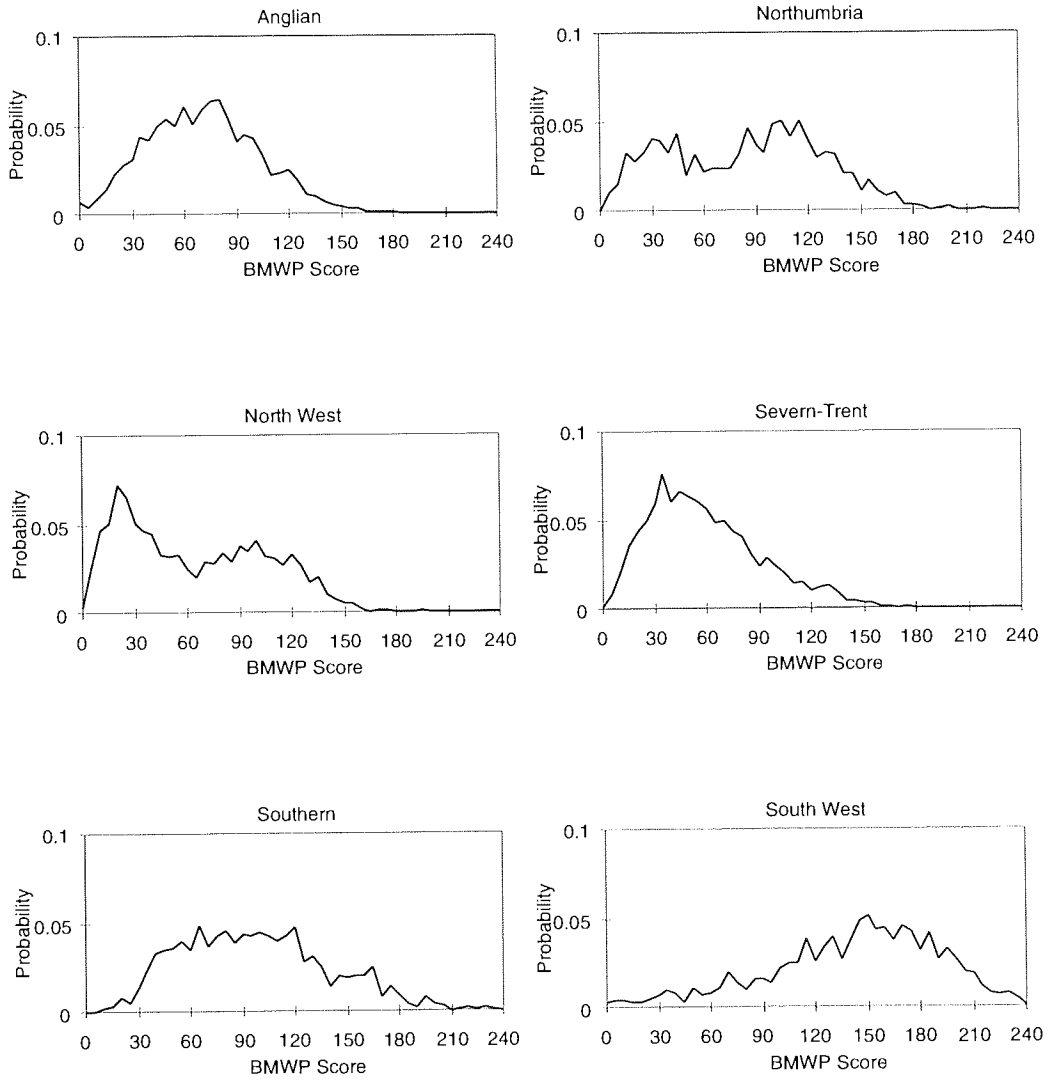



Figure 4.16: Distribution of BMWP Scores within each NRA region.

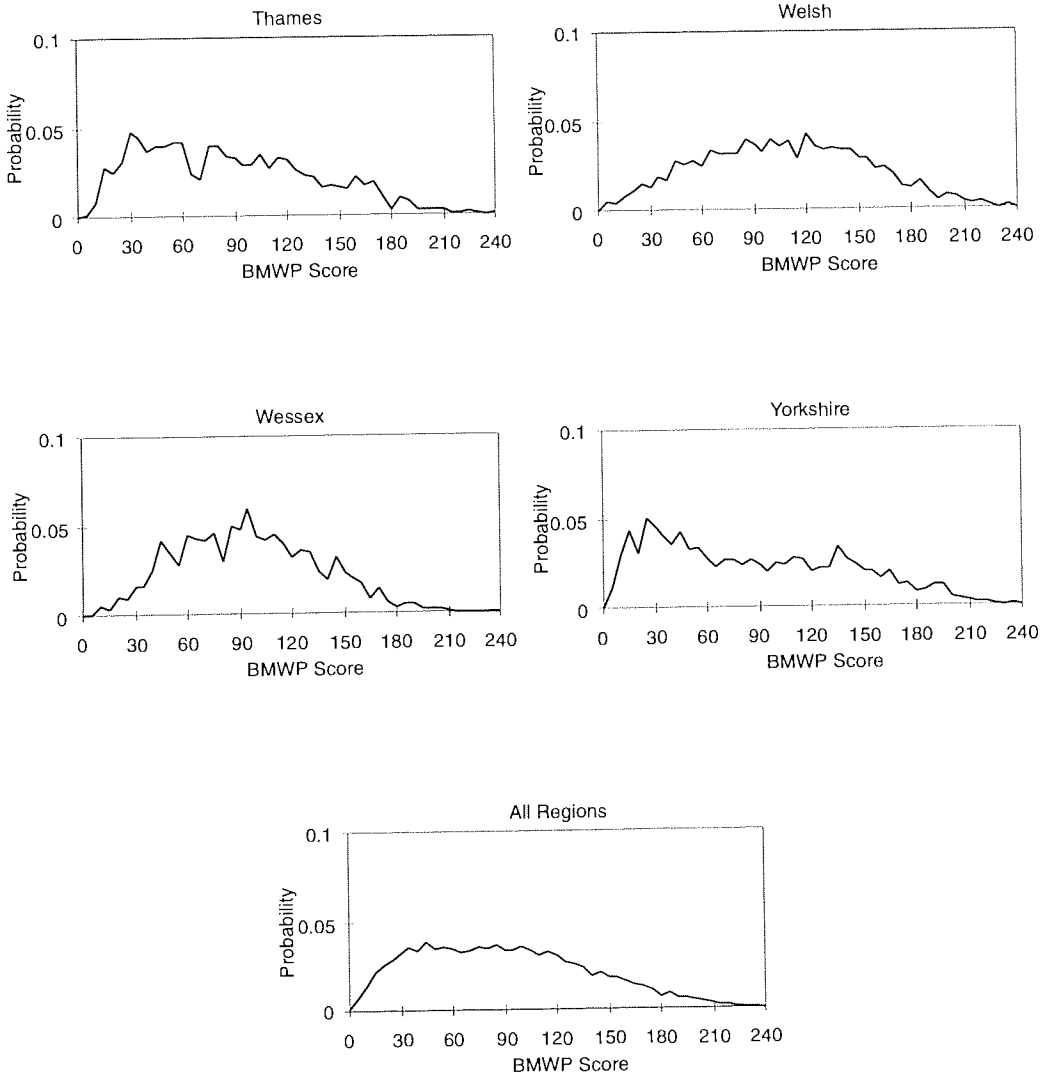



Figure 4.16: Distribution of BMWP Score within each NRA region (cont'd.)

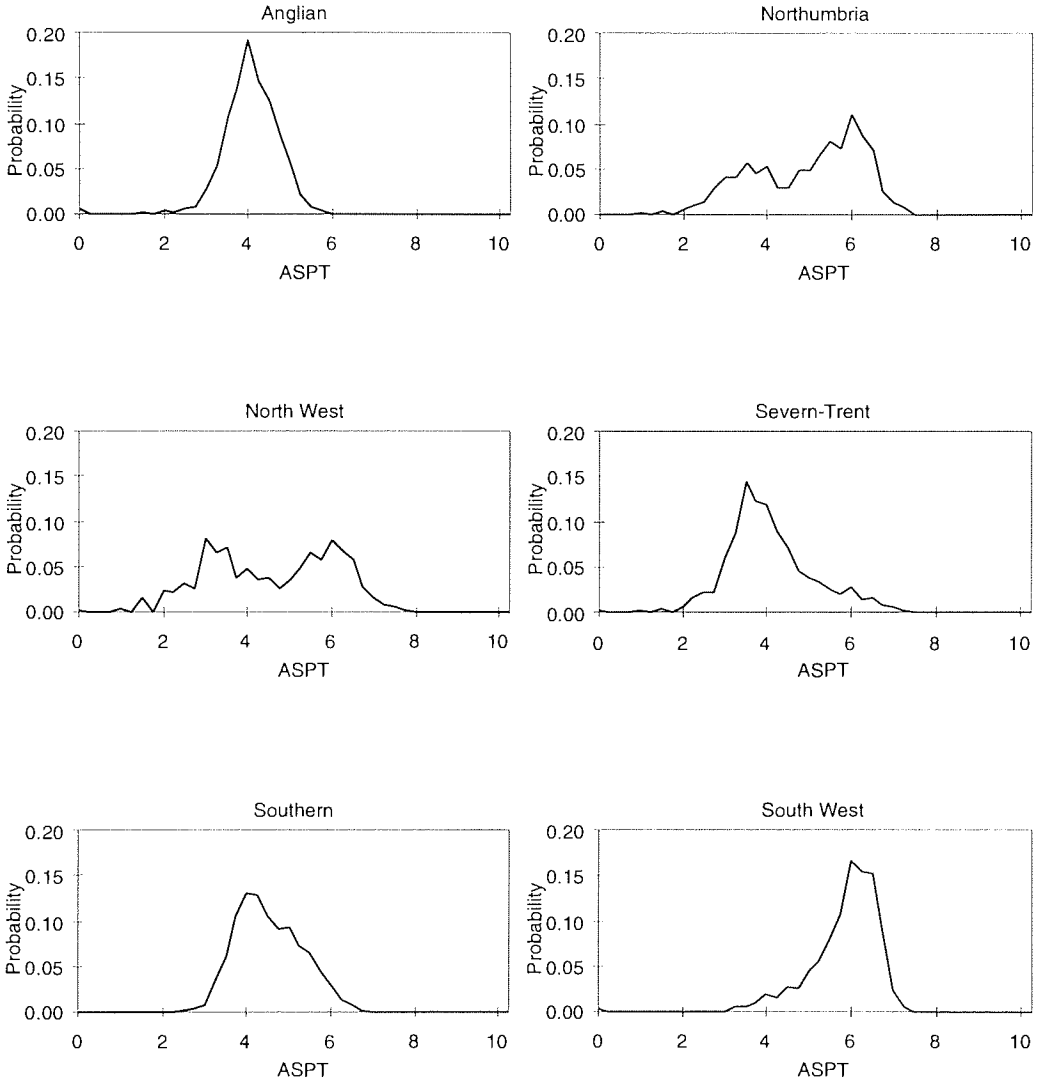



Figure 4.17: Distribution of ASPT with each NRA region.

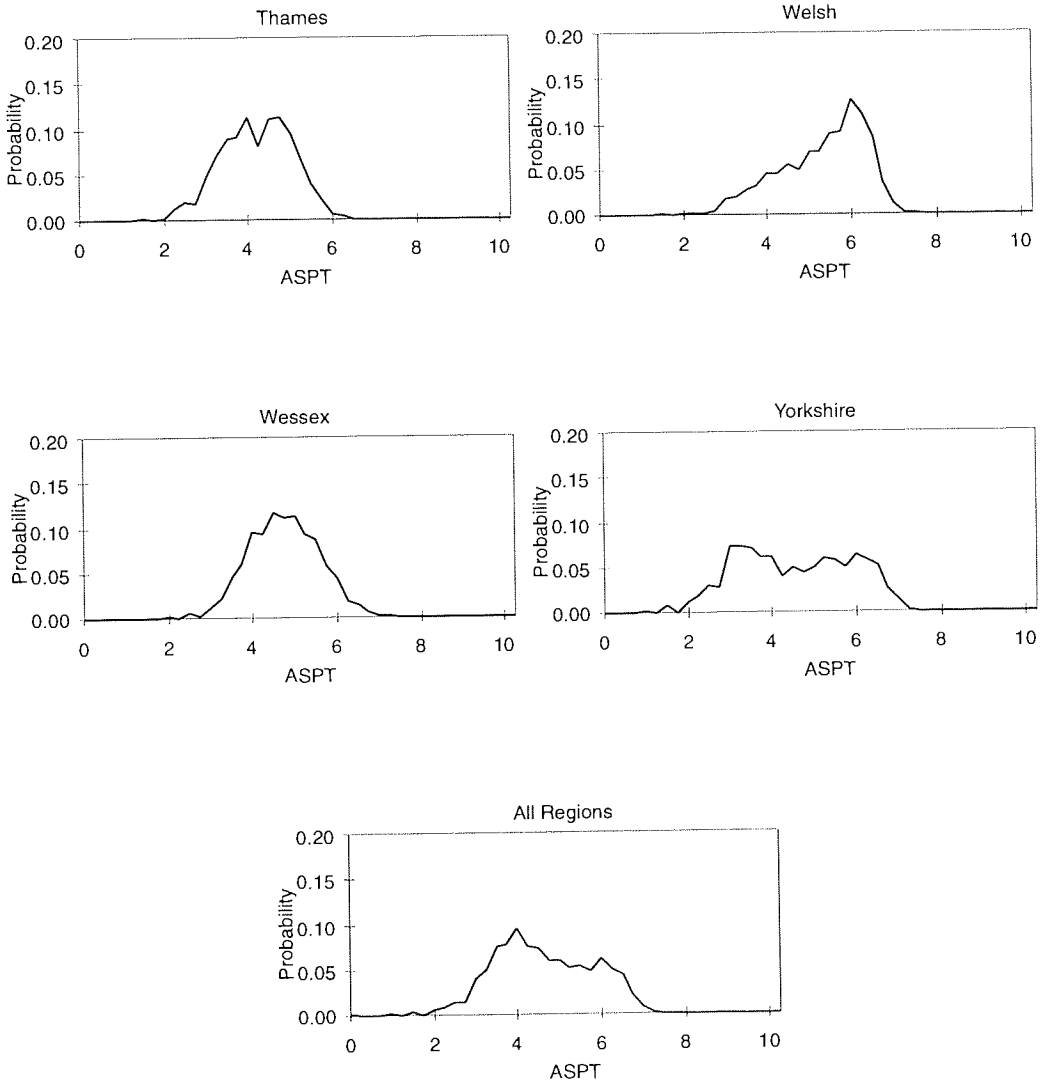



Figure 4.17: Distribution of ASPT within each NRA region (cont'd.)

#### 4.4.3.2 Distribution of Taxa by Region

The distribution of different taxa between regions was probably unknown at the time when the BMWP system was designed, and even today, there is a degree of uncertainty to the national incidence of taxa. With the advent of electronic storage and national monitoring programmes more research would be possible into the variation of the taxa and the effects of the local environment, and the following figures will no doubt be improved upon in the near future. But at the present time they do provide some unique insights into the distribution of aquatic invertebrates throughout the UK.

Figures 4.18 and 4.19 show the percentage occurrence of seven families of invertebrates in all of the samples of that region. This is simply the percentage of the total number of occurrences disregarding abundance level (i.e. a taxon was either absent/present) to the total number of samples for that region. Of the seven taxa the more tolerant organisms, Gammaridae, Asellidae and Baetidae all have a have a frequency of over 30% in all regions (Figure 4.18). This is in contrast to the four more sensitive families in Figure 4.19 where there is a much greater variation over the regions. Five regions (Northumbria, North West, Yorkshire, Welsh and South West) all have all four of the sensitive families in at least 10% of their samples, while Anglian and Thames have less the 10% in total. It appears that the distributions are strongly influenced by geology and topography of catchments, and physiographic characteristics of rivers, which are the main regional differences.

These graphs, in conjunction with the BMWP score and ASPT distributions demonstrate the problems of national based monitoring systems. This naturally high variation, on a national scale, is an important factor to be considered with respect to the implementation of any national monitoring system. The establishment of national databases allows for the review of methods of biological monitoring to a much greater extent than which previously would have been possible. For example, the graphs of the percentage occurrences of taxa over the whole of the UK highlights the huge regional variation that is present, even when only looking at family level. The variation in incidence results in differences in BMWP score and ASPT even when a similar range

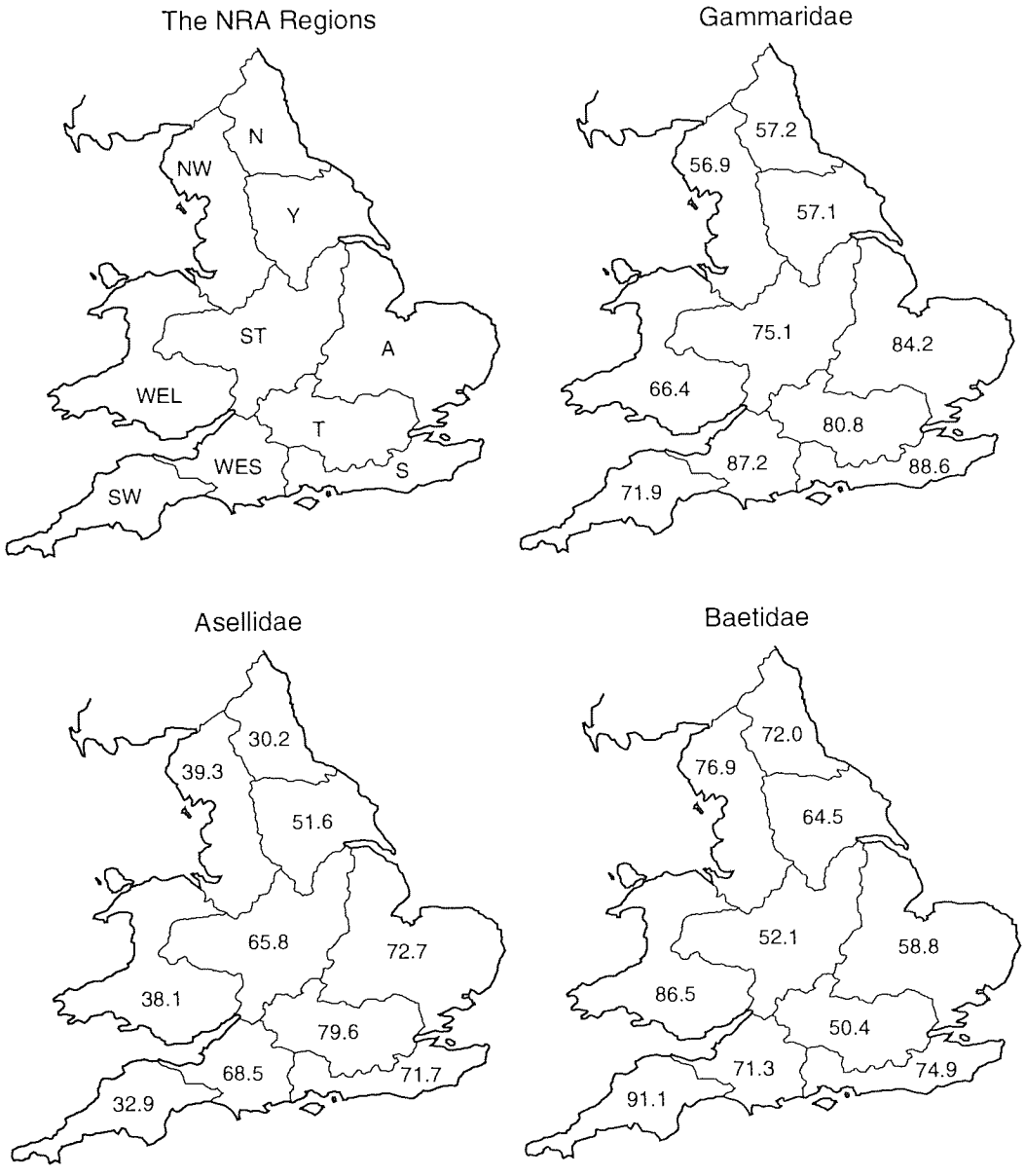



Figure 4.18: Percentage occurrence of Gammaridae, Asellidae and Baetidae within the 10 NRA regions.
Key: (A) Anglian, (N) Northumbria, (NW) North West, (ST) Severn-Trent, (S) Southern, (SW) South West, (T) Thames, (WEL) Welsh, (WES) Wessex, (Y) Yorkshire.

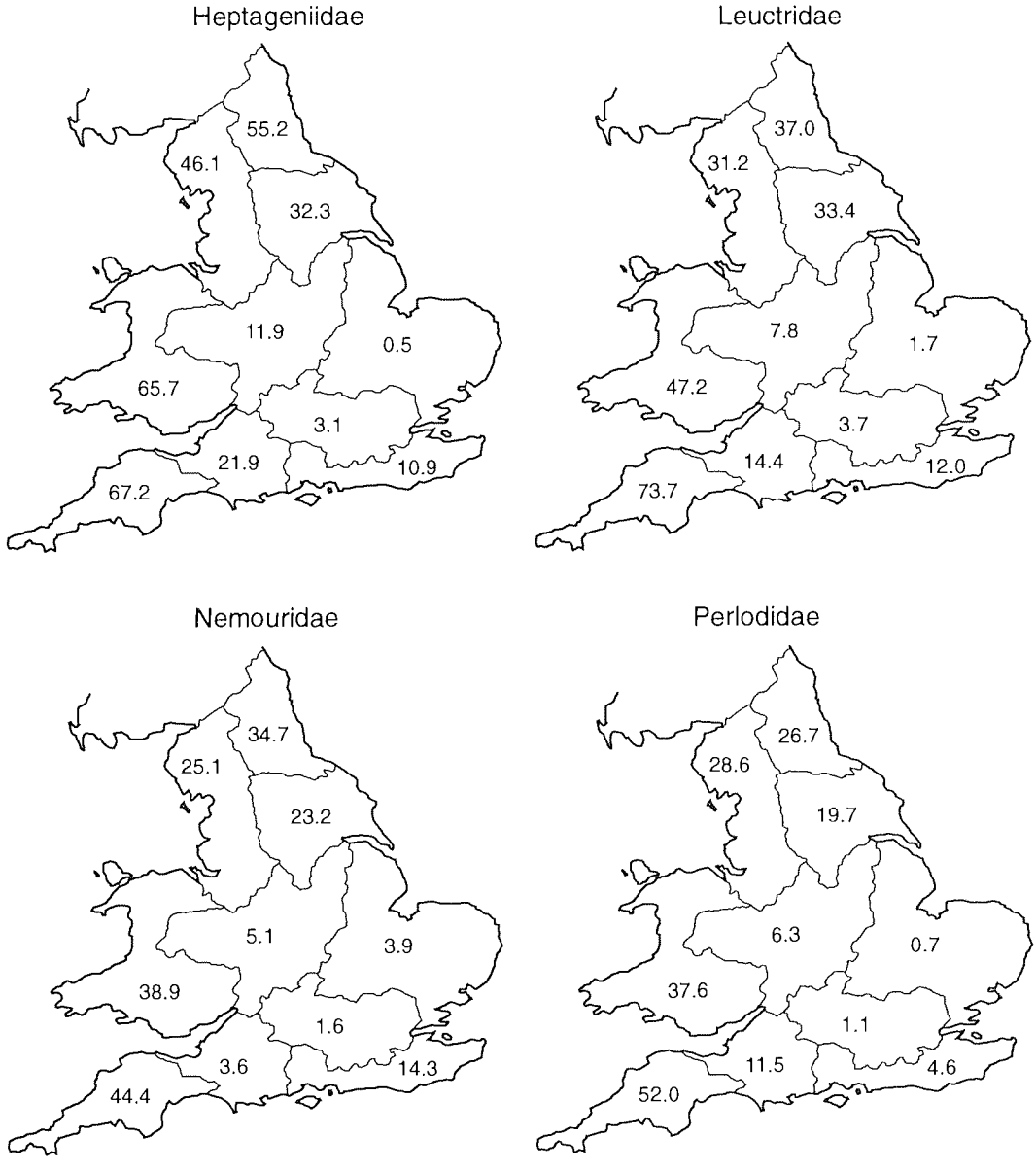



Figure 4.19: Percentage occurrence of Heptageniidae, Leuctridae, Nemouridae and Perlodidae within the 10 NRA regions.

| Data Set | Number of Samples | Comments |
|---|---|---|
| Yorkshire River Authority | 50 | Used for preliminary neural network experiments. Inconsistent mixed level identification |
| NRA Severn-Trent | 292 (1378) | Classified into biological class by the Expert. Used for neural network experiments and indirect elicitation of conditional probabilities. Inconsistent mixed level identification. |
| NRA National | 6000+ | Used to show distribution of taxa, BMWP score and ASPT over 10 NRA regions. Family level identification. |
| Conditional Probabilities | n/a | From direct elicitation. Used in Chapter 6 for the selection of indicator taxa, and as basis for synthetic data. Mixed level identification. |
| Synthetic Data | 4800(x3) | Used for neural network experiments. Mixed level identification. |

Table 4.14: Summary of river invertebrate data.

of water qualities are considered, which reaffirms the proviso that the BMWP score and ASPT should not be used to report comparisons between sites.

## 4.5 Summary

Within this Chapter several river invertebrate data sets have been described and studied. The range of data studied in this project is exceptionally broad, ranging from real world data, to elicited probabilistic domain knowledge and synthetic data. Table 4.14 gives a summary description of the river data that have been studied. The Severn-Trent and synthetic data constitute unique invertebrate data sets, and will be used extensively in the following two chapters.

The foundations of this project are described in Section 4.2, which draws on the early work of the BERT system. Within this section the conceptual basis of the biological classification system adopted was described. The biological classification was designed to mirror the present NWC classes but also has

a degree of similarity with the Saprobic system which is used elsewhere in Europe. The importance of information loss and handling uncertainty are also commented upon.

The Severn-Trent data, which was this project's main 'real world' data set, was detailed in Section 4.3. This data is typical of many environmental data sets, having many small inconsistencies in the recorded data and incomplete knowledge about the sample. The construction of a database suitable for neural network experimentation from the Severn-Trent was described. The reliability of the Expert's classifications was assessed, and was found to be consistent to within two grades on a thirteen grade classification. The biological classification was compared to four other commonly used biotic systems. These comparisons demonstrated that, using the Expert's classification as a benchmark, a non-monotonic relationship exists between water quality and both the BMWP score and 'Number of taxa'. The ASPT and TBI showed a more linear relationship with absolute quality, but there was a high variance for intermediate quality classes.

In addition to the Severn-Trent data set three other river invertebrate data sets were used (Section 4.4). An early pilot study used data from the Yorkshire Water Authority, and found that MLP models could achieve classification rates of 70% via the direct classification of invertebrate sample. The creation of two synthetic data sets, using conditional probabilities from the BERT knowledge base, were described. The two sets of data represented different communities 'typical' of riffle and pool biotopes, with a complimentary set of physical characteristics also being developed. A randomly selected set of the synthetic data were classified by the Expert, and were thought to be sufficiently realistic to be of use in experimentation. Finally, the 1990 National Survey database was used to show the variation of ASPT and BMWP scores between the 10 NRA regions. The large degree of variation was in contrast to the results of the 1990 regional classification's based on the NWC, which were broadly similar across all regions. This was not the case for the distributions of ASPT and BMWP score. The national distribution of seven families of taxa were studied, and, these also showed a large variations both between and within the families. The more sensitive animals had the higher variance, while the more tolerant organisms were more consistently distributed throughout the 10 regions.

The invertebrate data sets are typical of many sets of environmental data. Typically, large amounts of data are available, with the data being more 'observational' in nature than the data generated from designed experiments. The data represented complex cause-effect relationships, with both seasonal and geographic variation. They suffered from a lack of standardised methods of measurement, with both systematic and random errors being present. The systems described in the following chapter, while still subject to the above inadequacies, do demonstrate that the reliable classification and interpretation of environmental data, especially benthic invertebrate data, is possible.

# Chapter 5

# Neural Network Experiments

## 5.1 Introduction

This chapter reports the neural network experimental work undertaken using some of the river data described in Chapter 4. A brief review of the implementation details is given, as well as the general methodology adopted for the experiments. Following this, the work concentrates on the classification of biological water quality via the direct interpretation of invertebrate samples. The results of a series of experiments which investigated the affect of modifications to the networks on classification performance are reported. The effect of the method used to encode the input data upon the overall performance is examined, as is the utility of combining model predictions and using balanced data sets. Applications other than classification are then considered; including the detection of novel samples, the handling of data from different biotopes and the graphical representation of biological class. A qualitative comparison between these neural network models and the BERT system is made. The chapter concludes with a summary of its main findings.

## 5.2 Preliminaries

### 5.2.1 Implementation

The main neural network simulator that was used was the Xerion library [165] of routines. The Xerion library is a collection of routines specifically written for neural network researchers, and was found to be ideal for implementing many of experimental procedures described later on in this chapter. It was

used extensively for all the MLP networks and also for the mixtures of experts experiments (Section 5.4). The library is distributed as source code, which is comparatively easy to install and compile, provided that you have a reasonably well maintained system.

Other software was also used for the neural network experimentation. The SOM_PAK program [84] was used for the work on Self-Organising Maps. This again was provided as source code, which was easy to install, compile and use. For the work on the detection of novel samples (Section 5.5) some computer code had to be written, but this was a straight forward exercise. An alternative method would have been to implement the algorithm in a specialist package, such as MATLAB.

The majority of the neural network experimental work was conducted on a SUN IPX machine runnning a UNIX operating system. This was the preferred system (as opposed to a DOS/WINDOWS based one) as the Xerion library could only be compiled for UNIX based systems. Memory management is also less of a concern on UNIX systems.

In this dissertation the network models under consideration were all relatively small, so training times were not prohibitive (in the order of minutes and seconds rather than days and hours). The storage requirements were also not onerous, with 50MB being sufficient for the data sets, models, executables and scripts. As none of the data sets was particularly large they were stored in a database on a personal computer, and the various manipulations of variables (e.g. standardisation) were completed within the database environment. The transformed data files were exported as ASCII text files for the experiments.

A small library of shell scripts (batch files) were written to allow for the full automation of the experimental cycle. The library provided scripts for searching and collating results from the log files, handling the cross-validation of the data sets (see Section 5.2.2), general training and testing of the various models and also the control of various graphical displays of the models if required. The automation of the experimental cycle lead to the elimination of data handling errors, which would have most likely occurred if any data manipulations were carried out by hand. It also meant that experiments could be run overnight, thus making better use of the available resources.

## 5.2.2 Experimental Methods

### 5.2.2.1 Training, Validation and Testing

As commented upon in Chapter 3, a model's performance should be assessed using data independent of the data used to estimate the model's parameters. This typically requires the use of three sets of data, referred to as the training, validation and testing sets. The training data is used to estimate the model's parameters, this estimation process is typically halted when an error measure, usually the mean square error (MSE), calculated for the validation set starts to worsen. The model's performance is then assessed using the testing set. A common alternative method, known as early stopping, uses only the training and testing sets. Here the learning procedure is continued until some *ad hoc* point is reached, for example when 200 epochs (iterations of the minimisation algorithm) have been completed or a MSE of 0.01 has been reached.

To investigate the relationship of MSE with the training, validation and testing simulations some networks were trained with a small weight-decay $\lambda$ of 0.01, using 8 hidden nodes and conjugate gradients. The topology of the network was 41-8-5, that is 41 inputs, 8 hidden nodes and 5 outputs. Figures 5.1 and 5.2 show the MSE and error rate[1] for the training, validation and testing data using the synthetic riffle data (Section 4.4.2).

The MSEs of the three curves are similar over the first few iterations, but then the MSE for the training data begin to decrease faster than those of the validation and testing data (Figure 5.1). Improvement of the training MSE continues over the full number of iterations, however there appears to be a tailing off of the performance for the validation and testing data. Both of these latter sets exhibit a similar relationship, which is as expected since both were independent of the training data and drawn from the same underlying statistical distribution. What is slightly unusual for the validation and testing data is that the MSEs level off but do not begin to increase. One possible reason is that the data is synthetic and all three data sets closely resemble the underlying statistical distribution, so the differences between each is small and overfitting does not have a noticeable effect on the validation and testing

[1]Error rate = 1.0 - classification rate, where the rates are expressed as percentages between 0.0 and 1.0 (e.g. 20% = 0.2).

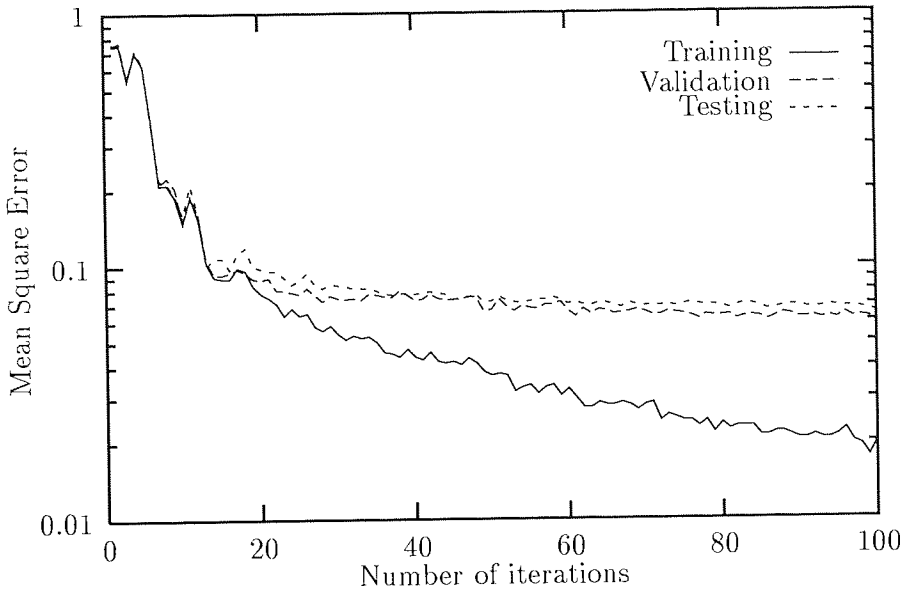



Figure 5.1: MSE plotted against number of iterations for training, learning validation and test data sets taken from the synthetic data.

performances.

The plot of the error rate, Figure 5.2, is analogous to the MSE except that that the curves are slightly smoother. As can be seen by comparing the two graphs there is a clear relationship between the error rate and the MSE, but this does not necessarily mean that this is always the case. Typically, the MSE is used as the criterion on which the validation is based, and it is this which is used for the rest of the dissertation.

#### 5.2.2.2 Cross Validation

For small data sets there can be a problem dividing up the data into the three sets since this leaves insufficient data on which to estimate the model's performance. To overcome this problem and to maximise the utility of the data, cross validation can be used.[2]

The $k$-fold cross validation algorithm involves three steps (after Efron & Tibshirani [28]):

[2]Note that cross validation and learning validation are two separate and distinctly different concepts.

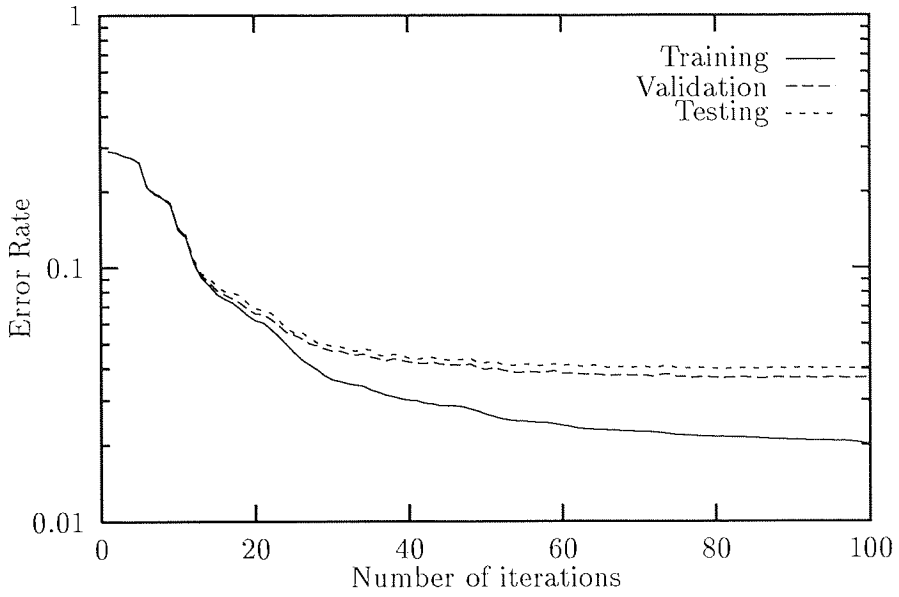



Figure 5.2: Error rate plotted against number of iterations for training, learning validation and test data sets taken from the synthetic data.

*i.* split the data into $k$ (roughly) equal parts,

*ii.* for the $k^{th}$ part train the model on the other $k-1$ parts, and test its performance by calculating the error rate of the fitted models when predicting the $k^{th}$ part,

*iii.* repeat step *ii.* for all $k$ parts and combine the $k$ estimates of the error rate.

Leave-one-out cross validation omits a single example at a time, and is the most rigorous method that can be adopted without resorting to a bootstrap analysis.

A problem with $k$-fold cross validation is that instead of a single model, $k$ models are generated and it is not possible to combine all $k$ models into a single one, except for the case where all the models are linear. Also by training $k$ models there is an associated increase in the computational effort required to conduct a series of experiments. Additionally, the estimate provided by any cross validation procedure has a high variance but a low bias. As always there is a trade-off between the bias and variance of a model (in a neural net context see Geman et al. [43]).

### 5.2.3 Minimisation

The adjustment of the weights of the network (learning) is equivalent to unconstrained optimisation. It is unconstrained as there are, typically, no constraints on the values which the weights can take, but by the use of regularisation the distribution of the weights can be indirectly manipulated. The optimisation is a function minimisation in terms of the weights of the network, so if there are 100 weights then the problem involves minimisation in 100 dimensions. This is a little over simplified as parameters other than the weights can be adjusted during minimisation and there are also various pruning algorithms, which remove weights, and ontogenic models, which grow or shrink the network to an appropriate degree. The derivative information (obtained from the back-propagation algorithm) can also be used in the minimisation. The original back-propagation algorithm was described using a simple steepest descent method to update the weight values:

$$w_{i+1} = w_i - \eta \frac{\partial E}{\partial w_i} \tag{5.1}$$

It is well documented that simple steepest descent can be a very inefficient, relatively slow algorithm, which is also prone to getting trapped in local minima.

Most minimisation schemes can be described by the following simple algorithm [38]:

*i.* determine a new search direction $\mathbf{s}^k$,

*ii.* find $\alpha^k$ which minimises $f(\mathbf{w}^k + \alpha^k \mathbf{s}^k)$, and

*iii.* set $\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha^k \mathbf{s}^k$.

For example, steepest descent uses a fixed $\alpha$ (which is the learning rate $\eta$), and a search direction $\mathbf{s}$ which is the vector of steepest descent (hence the name). The minimisation can be broken down into two stages: the first is the determination of the new search direction, the second being a line search along this new search direction to determine the starting for the next iteration. The line search is the most important aspect of the algorithm as this is where most of the computational time is spent during minimisation. Most minimisation

algorithms require first order-derivative information; however as is it possible to obtain second order derivative information ($\partial^2 E/\partial w_{ij}^2$) this can be incorporated into the minimisation algorithm. Buntine and Weigend [18] review second order derivative methods. Also some pruning schemes use the second order derivative information to eliminate redundant weights from the network after learning [90, 48].

In the neural network community the two most common methods for finding the new search direction (excluding steepest descent) are the conjugate gradient and quasi-Newton methods. Only conjugate gradients have been considered in this dissertation, with the Polak-Ribiere method being used [44, 38, 131]. To test the minimisation algorithms a simple experiment was undertaken using the synthetic data to compare the optimisation methods. The methods tested were steepest descent, quickprop (which is another popular method [34]) and the conjugate gradient algorithm.

Figure 5.3 shows the relationship between mean square error (MSE) and time for a typical training run. Of the three minimisation methods considered the conjugate gradients code is both quicker (in that it reduces the MSE faster) and finds a better solution than either steepest descent or quickprop. The fact that the conjugate gradient code reaches a better minimum is, perhaps, not surprising, but it was a surprise that it was quicker as it is a more complex calculation. Quickprop performed well, but steepest descent did not learn at all on this large data set. A number of different strategies were experimented with, involving changing both the learning rate and the momentum term, but none was successful.

The relative complexity of computation of the three different minimisers is shown in Figure 5.4, where the number of weight updates are plotted against time. The Quickprop algorithm is the least computationally complex, while conjugate gradients is the most. The conjugate gradient curve is not quite linear, but has a slight curvature which indicates that more time is being spent on the line search routine as the minimisation is nearing completion (all other aspects of the minimisation remain constant except for the time spent on the line search). These experiments demonstrated that the most practical minimiser was the conjugate gradients, as this was both the most reliable and found the best minima of the methods tested. So for the remainder of the

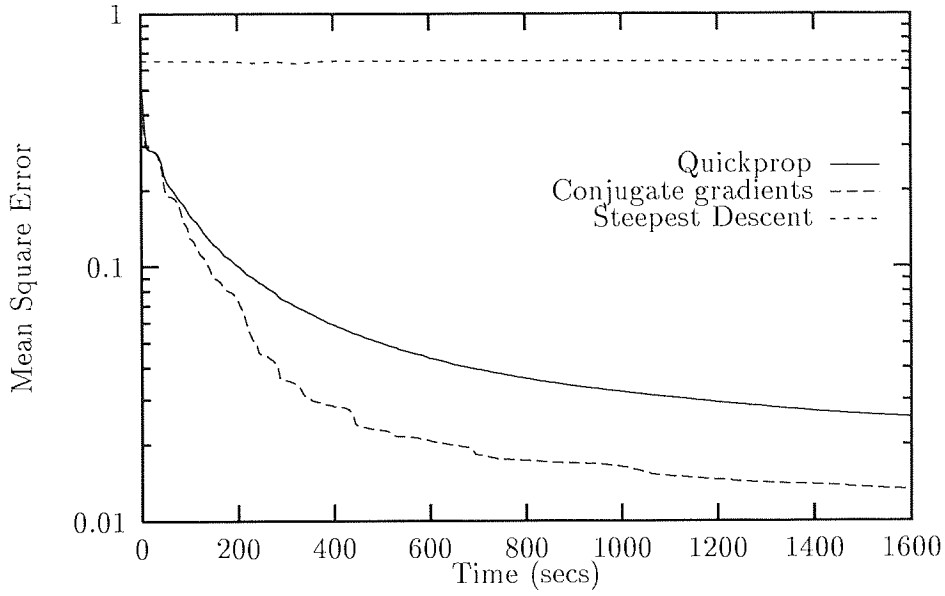



Figure 5.3: MSE plotted against time for Quickprop, conjugate gradients and steepest descent. The results are for the synthetic data. Note that steepest descent failed to learn in this example.

experimental work the Polak-Ribiere conjugate gradient method was used for minimisation.

## 5.3 Direct Interpretation

### 5.3.1 Overview

This section considers the use of neural networks for direct interpretation of water quality class from the invertebrate community. A thorough investigation is presented which looks at a number of issues concerning the implementation of MLP models, including scaling of input and output variables, model regularisation and the combination of several models.

### 5.3.2 Hidden Units

The number of hidden units which are used in a MLP can be critical to the performance of the model. If too few hidden units are used, the model will be unable to learn (i.e. it will not be flexible enough to model the desired

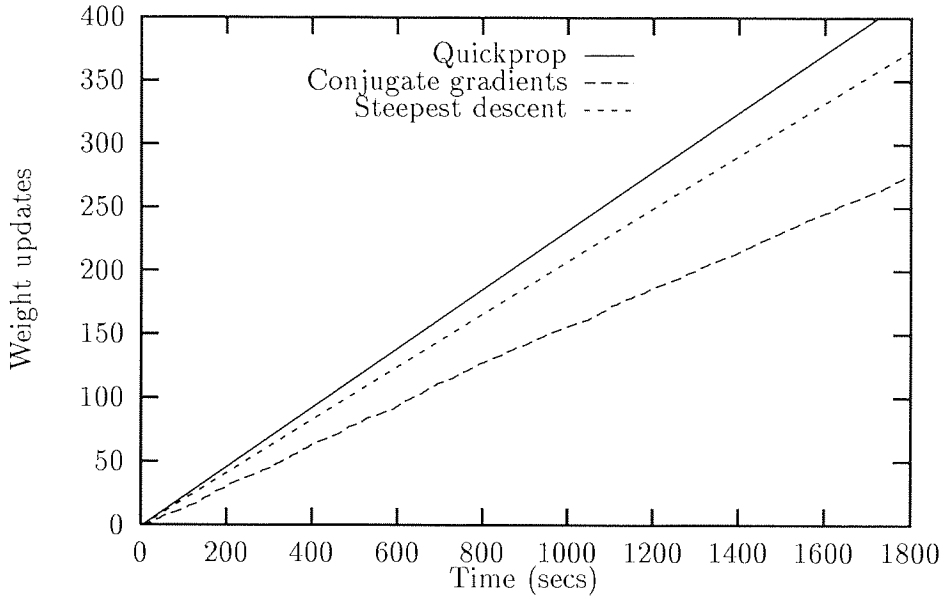



Figure 5.4: Number of weight updates plotted against time for Quickprop, conjugate gradients and steepest descent. The results are for the synthetic data.

mapping). If too many are used, the extra parameters (weights) will increase the likelihood of the model to overfit the data, and will also slow down the computation.

To investigate this, a series of models were trained with 2, 4, 6, 8, 10, 12, 15, 20, 25 and 30 hidden units. Two weight decay parameters were also considered, and each network was trained 10 times using different randomised values for the weights in each trail. The Severn-Trent data, transformed into 16 dimensions using a Principal Component Analysis (PCA), was used for training the models (see Section 5.3.4). The results of the MSE at the point where the validation error starts to degrade is given, averaged over the 10 trails in Figure 5.5. For the validation data it is apparent that the number of hidden units had little effect on the accuracy of the mapping achieved, except for 2 hidden units which had a much poorer performance. The weight decay parameter also had little effect on the distribution. There is a slight trend of increasing MSE with the higher number of hidden units. For the rest of the experiments 8 hidden units were generally used in the MLP models.

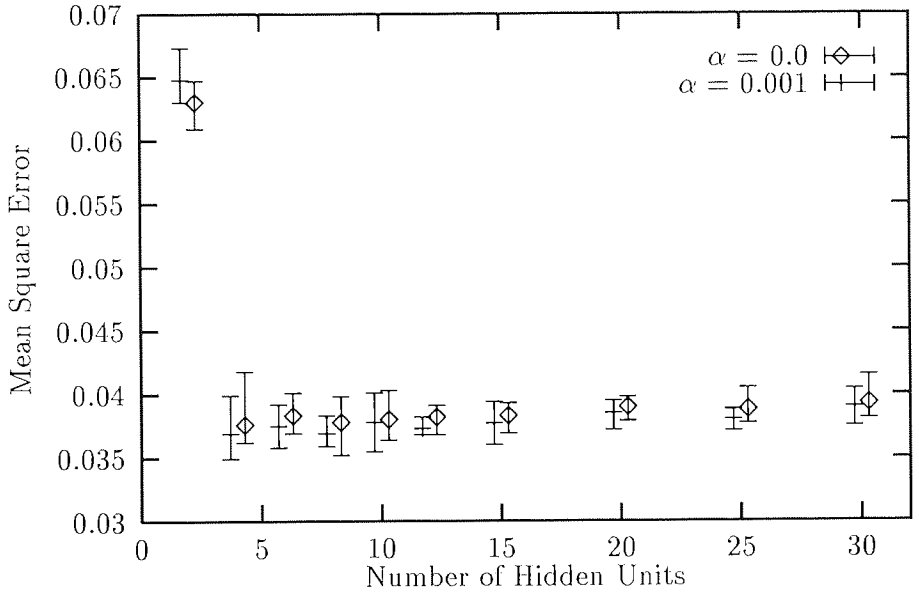



Figure 5.5: Minimum MSE for the validation data plotted against hidden units for 16 PCA Severn-Trent data.

### 5.3.3 Regularisation

The regularisation of models is a commonly used technique and can be important in preventing overfitting of data (see Section 3.4). Two methods of regularisation were considered for the MLP models, namely weight-decay and soft-weight sharing [120]. Weight-decay is the most commonly used regulariser, and is very easy to implement. Soft-weight sharing is a more complex regulariser that models the weights as a mixture model of Gaussian distributions. Again, a series of MLPs were trained with different regularisers. Eight regularisers were considered, five weight-decay $\lambda$'s (1.0, 0.1, 0.01, 0.001, 0.0001) and three configurations for the soft-weight sharing (mixture models of 2, 5 and 10 Gaussians). Finally, an unregularised MLP was trained for use as a benchmark. The average results for 10 runs of each regulariser are given in Figure 5.6.

It appears that the regularisers have little effect on the MLP performance for the Severn-Trent data, except for the case where $\lambda = 1.0$. The variances are small as well. The $\lambda$ of 0.1 had marginally the best performance and this value was used in the subsequent experimental work.

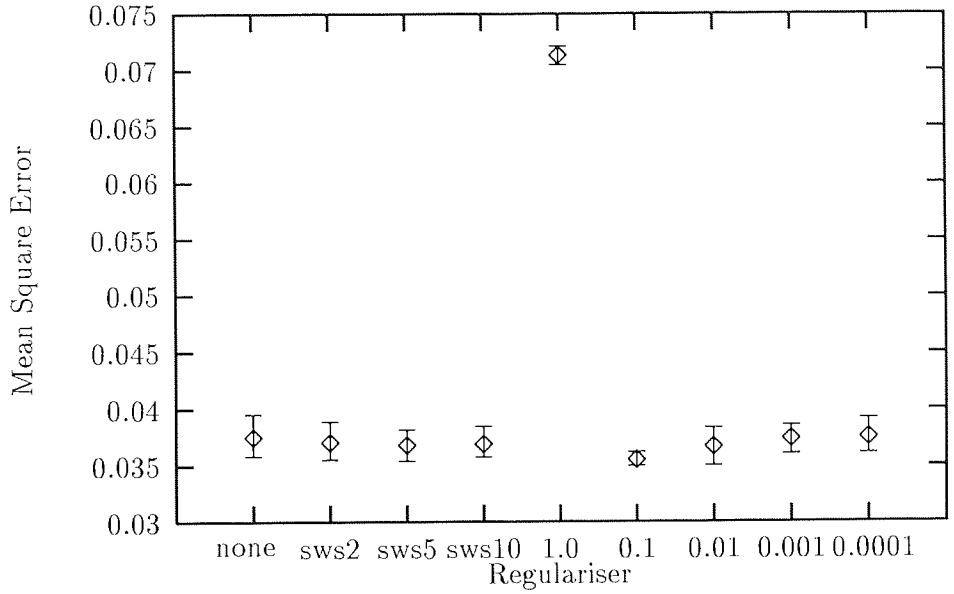



Figure 5.6: MSE plotted for a selection of regularisers, averaged over 10 runs, for the 16 PCA Severn-Trent data. Key: 'none' refers to unregularised MLP, sws to 'soft weight sharing' and the rest to the weight decay $\lambda$'s.

### 5.3.4 Data Encoding

Perhaps the most under reported aspect of neural network research is the pre-processing measures which are implemented before the data is presented to the network. This is surprising as it is probably the most important single factor which affects the performance of the model. This section looks at different pre-processing measures that can be adopted for the invertebrate data (input) and the quality classification (output). The aim of this section is to show that it is important to consider pre-processing the input data prior to its use in the network.

For most continuous variables it is popular to present standardised (zero mean and unit standard deviation) values to the model. However, for most of the data used in this thesis this is not applicable. The various qualitative abundances of the taxa (*absent*, *present*, *few* and *com+*) form an ordinal group of discrete intervals, so the standardised transformation is not applicable. A popular form of dimensionality reduction is that of principal component analysis (PCA). The algorithm transforms the coordinate axes of a system (an

orthonormal basis) into another basis which is aligned to maximally explain the variance in progressing dimensions. The first dimension of the principal component space explains the highest degree of variance that is possible in a single dimension, the second dimension explains the highest proportion of the remaining variance that is possible, and so on down to the final dimension. As there is an ordering to the components it is possible to pick only those which have a good explanatory power. This can be achieved either through examination of a scree plot or by using only the eigenvalues which are greater than unity; both were employed in this study. Essentially, a PCA is an unsupervised system (i.e. it takes no account of class labels) that is used to reduce the dimensionality of the data. Non-linear principal components are also occasionally used.

#### 5.3.4.1 Procedure

Four different methods of scaling the input data were considered, which are shown in Table 5.1. The first two methods differ in how absent evidence is weighted, with the three other states being the same. The other two methods are based on principal component transformations, as described in the proceeding section.

For the output classification two methods were used: the usual *one*-of-*N* coding was used (i.e. B1a = [1 0 0 0 0], referred to as 'Hard Class') for the five category class system, and a probabilistic interpretation of the 13 class system (see Section 4.3) (e.g. B1b+ = [0.33 0.67 0 0 0] or B3− = [0 0 0 0.67 0.33], 'Prob. Class'). For both of these output encoding systems a softmax constraint was used on the output layer nodes. Softmax constrains each output unit to between 0.0 and 1.0, and ensures that the sum of the outputs is unity [14]. This allows the output distribution to be interpreted probabilistically.

One possible problem with the representation of the classification system is that the relationship between the classes is not explicit, because the network does not take into account the ordering of the output units. For example, the order of the output classes are typically B1a, B1b, B2, B3 and B4, which is the intuitive ordering, but from the network's 'point-of-view' this is equivalent to an order of, for example, B2, B1a, B4, B3 and B1b. The model performance is not hindered by the apparent mis-ordering of the output classes. A method

| Name | Dim. | Description |
|---|---|---|
| 41 Absent(0.0) | 41 | Absent(0.0), present(0.33), few(0.66), com+(1.0) |
| 41 Absent(-0.33) | 41 | Absent(-0.33), present(0.33), few(0.66), com+(1.0) |
| 16 PCA | 16 | Cut-off taken as eigenvalues greater than unity |
| 5 PCA | 5 | Cut-off taken as knee of scree plot |

Table 5.1: Input data for pre-processing experiments.

to overcome this is to collapse the classification into a single linear index, but this loses the posterior probability distribution across the networks outputs. For the single index two methods were considered. The first used the five main biological classes mapped to integer values (B1a was mapped to 0.0, and B4 to 4.0, this is referred to as 'Hard Linear'). The second used the thirteen classes, with the '+' increments reducing the class score by 0.33 (i.e. B1b+ = 1.0 - 0.33 = 0.67), while the '−' incremented the class score by 0.33, and this is referred to as 'Prob. Linear'. For the single network output a linear activation function was used.

The network models were trained using a training, validation and testing approach, with the data set split into four (i.e. four-fold cross validation). The results are all in terms of classification rates on the testing data. A weight-decay term of 0.1 was used, along with 8 hidden units for all the models.

#### 5.3.4.2 Results

The results are shown in Table 5.2. The input data which used the 16 dimensional PCA scaling had the best performance, with the 5 class PCA the worst. There was little difference between the two full 41 input sets. Considering the output classifications the 'hard' systems where marginally more successful than the 'probabilistic' system, with classification (five outputs) being more successful than prediction (one output).

#### 5.3.4.3 Using Additional Information

To improve upon the present classification rates, which tend to be in the region of just below 70% for test data, a measure of each sample's diversity was used

| Name | Hard Class | Prob. Class | Hard Linear | Prob. Linear |
|---|---|---|---|---|
| 41 Absent (0.0) | 65.8% | 64.2% | 64.5% | 62.8% |
| 41 Absent (-0.33) | 66.4% | 64.3% | 65.3% | 60.7% |
| 16 PCA | 70.9% | 67.9% | 65.8% | 63.4% |
| 6 PCA | 61.2% | 60.4% | 61.4% | 60.7% |

Table 5.2: Effects of classification accuracy for different input and output encodings of the Severn-Trent data, averaged over 10 runs. See accompanying text for details of data sets.

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 54 | 4 | 0 | 0 | 0 |
| | B1b | 9 | 49 | 13 | 0 | 0 |
| | B2 | 0 | 14 | 88 | 1 | 0 |
| | B3 | 0 | 0 | 4 | 26 | 5 |
| | B4 | 0 | 0 | 0 | 8 | 17 |

Classification rate = 80.1%

Table 5.3: Confusion matrix for MLP trained with 16 PCA Severn-Trent data and the number of taxa as additional input variables.

to augment the input data. As the input data consisted of only a subset of all of the available taxa, some taxa were not being utilised by the models to form the classification. To rectify this, the number of taxa in the sample was added to the input file as an additional input to the 16 PCA data set and the preceding experiment was repeated using the hard five classification scale for the output.

The result of this single addition is demonstrated by Table 5.3, where the classification rate is over 80%. Averaged over 10 runs the networks had an average classification rate of 78.8% on the testing data.

#### 5.3.4.4 Discussion

This section has demonstrated that the choice of representation for the data is an important aspect of the overall performance of the MLP models. The

major factor affecting the performance of the model was the choice of suitable features. The simple count of the number of taxa improved the classification rate by over 10%, which is an appreciable gain. The Expert probably used the diversity information implicitly when making his decision on the likely quality class, thus reflecting this in his resultant classification.

Examination of the misclassifications showed that the same samples were being consistently misclassified, always to an adjacent class. It is interesting to note from Table 5.3 that most of the misclassifications occurred around the B1b and B2 classes, which are the two classes with the highest variance in terms of both sample diversity and biotic indices (see Figure 4.10). It appears that some of these B1a and B2 samples are close to the class boundaries or even that the boundaries are not clearly distinguished in the Expert's mind, and hence more easily misclassified. This is reinforced by Figure 4.9 which shows a higher proportion of intermediate samples, in this case B1b− and B2+, between the B1b and B2 classes than the other pairs of classes (B1a/B1b, B2/B3 and B3/B4).

### 5.3.5 Combination of Models

#### 5.3.5.1 Overview

This section investigates the use of committees (or ensembles or stacking) of models. The combination of the predictions or classifications of multiple models into a single outcome has been widely discussed in the literature [120, 70, 178, 47, 88, 142]. Typically a number of topologies, learning rules, minimisation methods are used and a single 'best' model is selected for implementation into a system. This training of multiple models is particularly common in neural network studies, where considerable experimentation is carried out during the model selection process.

An important aspect in the combination of multiple models is to find the optimal mixing proportions. The are two general approaches to determine this, namely Bayesian and Cross Validation. Bayesian methods can take advantage of the evidence provided by each model, while cross validation determines the mixing coefficients by using a cross validation procedure. A problem with the cross validation method is that the training data has to be split into an

extra set for the determination of the mixing coefficients, and this can lead to problems when the training set is small. One simple method is to weight each model equally (this has been adopted in Chapter 7 for use with the Canadian data).

In this section three methods of combining models were investigated. These were:

**Simple Average:** The average of the output vector for all models was calculated, and the highest value was taken to be the classification of the committee. Notationally;

$$c_i^p = \frac{1}{m} \sum_{j=1}^{m} y_{ij}^p \tag{5.2}$$

where $c_i^p$ is committee output for the $p$th pattern and class $i$, $y_{ij}^p$ is the $j$th model output for pattern $p$ and class $i$, and there are $m$ models in total.

**Product Average:** This is similar to the preceding method, except that a product over the outputs is taken, thus:

$$c_i^p = \prod_{j=1}^{m} y_{ij}^p \tag{5.3}$$

The resulting $c_i^p$'s are then normalised, and the largest is taken as the committees's classification.

**Rogova's Method:** Rogova [144] suggests using Dempster-Shafer Theory for combining the results of a number of classifiers. The proposed method introduces a mechanism for the calculation of evidence reflecting the abilities of each individual classifier. Following the notation of Rogova, for a set of $N$ classifiers, $f^n$, the output vector $\bar{\mathbf{y}}^n$ is given by $f^n(\bar{\mathbf{x}}^n)$, where $\bar{\mathbf{x}}^n$ is the input vector. Assuming that we have $K$ classes it is usual that we assign class $j$ to the input vector $\bar{\mathbf{x}}^n$ if $y_j = \max_{1 \leq k \leq K} y_k$. This method is suitable for a scheme such as the majority voting where it is only required to know the single decision class for each network, however it does not take into account how categorical the classification has been.

A majority voting scheme would not take into account the difference between an output of (1,0,0) and an output of (0.26,0.24,0,24), which can be considered as unsatisfactory in some situations.

To overcome this Rogova [144] introduces a measure of evidence for each classifier $f^n$ and class $k$, denoted by $e_k(\bar{\mathbf{y}}^n)$ which is equivalent to $e_k(f^n(\bar{\mathbf{x}}))$. The input data corresponding to class $k$, $\{\bar{\mathbf{x}}_k\}$, is used to find the mean of the output vectors for class $k$ and classifier $f^n$, this mean vector being $\bar{\mathbf{E}}_k^n$. A proximity measure, $d_k^n$ is found between each $\bar{\mathbf{E}}_k^n$ and a output vector $\bar{\mathbf{y}}^n$ which is used to determine the evidence $e_k(\bar{\mathbf{y}}^n)$. The term $d_k^n$ is given by a function $\phi(\bar{\mathbf{E}}_k^n, \bar{\mathbf{y}}^n)$, which has the properties of varying between 0 and 1, and has a maximum value when an output vector, $\bar{\mathbf{y}}^n$, is equal to one of the mean vectors, $\bar{\mathbf{E}}_k^n$. A number of possible functions are put forward for the function $\phi$, and for the following experiments Equ. 5.4 is used:

$$d_k^n = \cos^m(\alpha_k^n) \tag{5.4}$$

where $\alpha_k^n$ is the angle between $\bar{\mathbf{E}}_k^n$ and $\bar{\mathbf{y}}^n$, and $\phi$ can be calculated by:

$$\phi(\bar{\mathbf{E}}_k^n, \bar{\mathbf{y}}^n) = \frac{(\sum_{1 \leq i \leq K} E_{ik}^n y_i^n)^2}{\|\bar{\mathbf{E}}_k^n\|^2 \|\bar{\mathbf{y}}_k^n\|^2} \tag{5.5}$$

Using Dempster-Shafer Theory the evidence $e_k(\bar{\mathbf{y}}^n)$ is then given by:

$$e_k(\bar{\mathbf{y}}^n) = \frac{d_k^n \prod_{i \neq k}(1 - d_i^n)}{1 - d_k^n \left[1 - \prod_{i \neq k}(1 - d_i^n)\right]} \tag{5.6}$$

The last step is to combine the evidence from each classifier, which can be simply written as:

$$e_k(\bar{\mathbf{x}}) = C \prod_n e_k(\bar{\mathbf{y}}^n) \tag{5.7}$$

where $C$ is the normalising factor. The final class assignment of the input vector $\bar{\mathbf{x}}$ is achieved by picking the class $j$ which satisfies $e_j = \max_{1 \leq k \leq K} e_k(\bar{\mathbf{x}})$.

#### 5.3.5.2 Balancing of Data Sets

The balancing of data sets refers to the distribution of the output classes in the training set. If one class is under represented in the data then it is likely that the model will have a high error rate for that particular class. There are a number of possible measures which can be taken to compensate for the low frequency of some classes, for example, the artificial replication of classes with an additional random component, the presentation of low frequency classes more often and the building of constraints into the error term [135]. The approach which was adopted for the following experiments was to present the low frequency classes more often to the network, with the effect that all five classes occurred with equal frequency within the training set.

#### 5.3.5.3 Procedure

20 networks were trained and ranked in order of classification rate for both the 'plain' data and the balanced data. The networks were trained using a training, validation and testing method, so training was stopped when the error on the validation started to increase. The 292 samples were split into four sets of equal size, with a four-fold cross validation strategy being used. The resulting sets of 20 networks were combined using the three methods described above (Section 5.3.4).

#### 5.3.5.4 Results

The results for both the plain data and the balanced data are presented in Table 5.4. There is a small difference in the classification rates between the two data sets, with the balanced data having a slightly lower overall classification rate. Of the three methods used to form the committees, Rogova's method resulted in a slightly, but not significantly, better performance.

#### 5.3.5.5 Discussion

Based on the analysis of the Severn-Trent data, the value of using committees of networks is debatable. The committees performed only fractionally better (usually 1 or 2 additional patterns correctly classified) than the indi-

| Method | Top 10 Mean (Std. Dev.) | Top 20 Mean (Std. Dev.) |
|---|---|---|
| Plain Data | | |
| Simple Average | 66.8% ±1.1 | 66.2% ±1.8 |
| Product Average | 66.7% ±1.9 | 65.5% ±1.4 |
| Rogova's Method | 68.2% ±1.9 | 67.1% ±0.9 |
| Balanced Data | | |
| Simple Average | 66.2% ±1.1 | 65.6% ±1.2 |
| Product Average | 66.4% ±1.4 | 61.8% ±1.4 |
| Rogova's Method | 67.7% ±1.8 | 67.3% ±1.4 |

Table 5.4: Classification rates using committees of networks for the 16 PCA Severn-Trent data.

vidual models. This is in contrast to the sediment toxicity experiments where committees were shown to be more useful (Section 7.5).

The effect of balancing results in a slightly lower overall classification rate, but there was a small redistribution of the correct and misclassified samples. The balancing did have the effect of improving the classification of the lower frequency B3 and B4 classes but to the detriment of the B2 classes. In effect, the misclassifications for the balanced data are more evenly distributed over all five classes.

There are some other considerations in the use of committees of models. One is that if the model is not capable of learning the mapping then there is no point producing stacked models from it. This occurs with, for example, multi-valued functions, this is where there are two or more possible output values for a given input. If a function is multi-valued then the network learns the average of the outputs (note that the average result may not be an acceptable solution), hence by stacking the models the average is improved, but is still wrong. The other consideration is that of the mixing proportions. If these are fixed, the probability of using an individual model's output is independent of the data presented to it, which may not be the best policy. For example if there are two models, (e.g. one is good at identifying riffle samples and the other pool), then some account should be taken of this when the mixing proportions are selected. This is considered in detail in Section 5.4.

### 5.3.6 Discussion: Direct Interpretation

A full and rigorous examination of all the aspects of generating a good neural network model for the direct interpretation of river water quality has been made. Somewhat unexpectedly the number of hidden units, methods of regularisation or combination of models were found to have little effect on the overall classification of an invertebrate sample. The pre-processing of the input and output data did have some effect, with a 16 dimension PCA set using a *one*-from-$N$ classification showing the best overall performance. This provides some support for keeping the input dimensionality small. The addition of a simple measure of sample diversity had a noticeably beneficial effect on the classification performance.

## 5.4 Classification within Different Biotopes

### 5.4.1 Introduction

In the previous section only the classification of riffle biotopes was considered. For a model to be of any practical value it must be able to reliably classify data from different biotopes. The problem with using a single model for a number of biotopes is that interference effects may exist which diminish the classification rates [121]. Thus, a model may get 90% of classifications correct using data from one source and 85% correct from a different source, when trained separately for each source. But if the model were trained using data from both sources together then the classification rates may fall to, say, 80% and 70% respectively.

It is possible to overcome interference effects by using a modular architecture where a number of models (or experts) are trained simultaneously with another network, referred to as the gating network, which selects the most appropriate model (expert) to use for the output classification. Each model can be considered as an 'expert', and the whole system can be described as an adaptive mixture model or a (hierarchical) mixture of experts. A number of authors have described ideas based around this philosophy, for example Nowlan [120], Nowlan and Hinton [121], and Jacobs et al. [70], and it is receiving an appreciable amount of attention in the literature. It should be emphasised that there

is a fundamental difference between the adaptive mixtures and the combination of models as discussed in Section 5.3.5. The mixing proportions within the committees of models are determined by validation methods and are fixed after the learning phase is over, thus remaining the same for all future patterns to be classified. In the adaptive mixture systems, the gating (selection) mechanism is conditioned on the input data, and leads to the use of different expert models for different input patterns. The conditioning of the selection of the models on the input data leads to a partitioning of the classification task between the models, (i.e. networks become specialised on subsets of the input space) and in effect focuses each on the classes which it is good at identifying, and neglects examples where its mixing proportion is small.

### 5.4.2 Procedure

Using a method as described by Nowlan and Hinton [121] an experiment was conducted into the utility of these methods for classifying invertebrate samples from both pools and riffles. The basic model is schematically illustrated in Figure 5.7. Two MLPs performed the task of the expert networks with the gating network being represented by a linear model. The two experts receive identical input, but the input to the gating network can either be the same as that to the experts or a completely different set of data. For this study, two scenarios were considered for the gating network input, the first was to use the invertebrate sample as input (i.e. give the gating network the same input as the two experts), while the second used the physical characteristics of the site as input. The inputs to the expert networks were samples taken from the riffle and pool synthetic data sets.

Both riffle and pool data were transformed into a 5-dimensional PCA space, and a training, validation and testing approach was used to monitor model performance. The training, validation and test data had 1800, 1600 and 1600 samples respectively for both riffle and pool sets. This gave the total number of training patterns as 3600 (riffles and pools combined). In order to benchmark the problem, and to quantify the interference effects, a single MLP was trained using the combined riffle and pool samples.

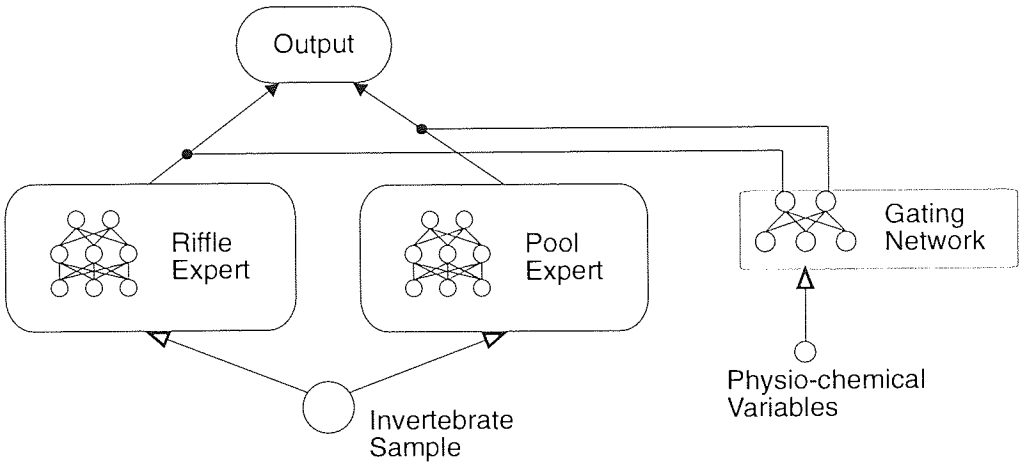



Figure 5.7: Schematic illustration of the mixtures of experts network for the classification of samples taken from riffle and pool biotopes.

The system was trained by using the following error function:

$$E^p = -\log \sum_i g_i e^{-\|\mathbf{t}^p - \mathbf{y}_i^p\|^2/2\sigma^2} \tag{5.8}$$

where $E^p$ is the error for training pattern $p$, $g_i$ is the output of the gating network for expert $i$, $\mathbf{t}^p$ is the target pattern and $\mathbf{y}_i^p$ is the output vector of expert $i$, and $\sigma$ is a constant term. Equ. 5.8 is suitable for gradient based minimisation algorithms as the first derivatives can be calculated. Each output of the gating network represents the probability that the associated expert is correct. In theory the true output of the network, for a given input, is stochastic, and to properly interpret the network stochastic sampling should be performed on the output. In this work the output vector was determined by using the expert with the highest mixing proportion, and not by combining the experts in proportion to their gating weights. In practice the gating network tends to make very categorical choices as to which expert to use, and the winning expert rarely has a mixing proportion of less than 0.9. This is explained by the effect of the error term, Equ. 5.8, which tends to raise, for each particular case, the mixing proportion of the experts that perform well on that example.

a) Riffle test data

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 265 | 53 | 2 | 0 | 0 |
| | B1b | 45 | 252 | 23 | 0 | 0 |
| | B2 | 0 | 56 | 237 | 27 | 0 |
| | B3 | 0 | 0 | 25 | 281 | 14 |
| | B4 | 0 | 0 | 0 | 10 | 310 |

Classification rate = 84.1%

b) Pool test data

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 247 | 59 | 14 | 0 | 0 |
| | B1b | 0 | 239 | 72 | 9 | 0 |
| | B2 | 0 | 2 | 215 | 85 | 18 |
| | B3 | 0 | 0 | 3 | 269 | 48 |
| | B4 | 0 | 0 | 0 | 7 | 313 |

Classification rate = 80.2%

Table 5.5: Interference effects for a MLP trained on the combined synthetic riffle and pool data. This network exhibits interference effects as both classification rates are lower than for a corresponding network trained on a single set (i.e. either riffle or pool) only.

### 5.4.3 Results

The results from the single layer MLP trained on the combined riffle and pool data are given in Table 5.5. The overall classification rate of the riffle data is 84.1%, whilst that of the pool data is 80.2%. Both of these figures are approximately 10% below the classification which can be achieved by networks trained separately as riffle and pool models. This suggests that there is some interference occurring within the combined model between the riffle and pool data. This is also suggested by the spread of misclassifications in the confusion matrices.

The results from the first experiment, where the gating network received the same input as the expert's, are given in Table 5.6. Both sets of data are

a) Riffle test data

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 309 | 11 | 0 | 0 | 0 |
| | B1b | 4 | 300 | 16 | 0 | 0 |
| | B2 | 0 | 23 | 280 | 17 | 0 |
| | B3 | 0 | 0 | 11 | 307 | 2 |
| | B4 | 0 | 0 | 0 | 6 | 314 |

Classification rate = 94.4%

b) Pool test data

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 309 | 11 | 0 | 0 | 0 |
| | B1b | 4 | 284 | 32 | 0 | 0 |
| | B2 | 0 | 23 | 268 | 29 | 0 |
| | B3 | 0 | 0 | 21 | 295 | 4 |
| | B4 | 0 | 0 | 0 | 6 | 314 |

Classification rate = 91.9%

Table 5.6: Classification of synthetic riffle and pool data where gating network had invertebrate sample as input.

classified better than for the single MLP, with overall classification rates of 94.4% for the riffle data and 91.9% for the pool data. The classification rates for the second experiment, where the gating network's input were physical variables describing the biotope, are shown in Table 5.7, again both sets of data are well classified. The riffle data had an overall classification rate of 96.2% and the pool data 94.3%, which is slightly better than the first experiment.

### 5.4.4 Discussion

The results demonstrate that there is a sizable benefit in using a mixture of experts architecture when data from different biotopes are being classified. This has traditionally been difficult to implement as species provide different information depending upon the biotope in which they were found [24]. In this example pool and riffle samples only were considered, but there is no reason

a) Riffle test data

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 317 | 2 | 1 | 0 | 0 |
| | B1b | 4 | 302 | 14 | 0 | 0 |
| | B2 | 0 | 20 | 292 | 28 | 0 |
| | B3 | 0 | 0 | 5 | 312 | 3 |
| | B4 | 0 | 0 | 0 | 4 | 316 |

Classification rate = 96.2%

b) Pool test data

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 315 | 4 | 1 | 0 | 0 |
| | B1b | 6 | 285 | 29 | 0 | 0 |
| | B2 | 0 | 16 | 294 | 13 | 0 |
| | B3 | 0 | 0 | 16 | 301 | 3 |
| | B4 | 0 | 0 | 0 | 6 | 314 |

Classification rate = 94.3%

Table 5.7: Classification of synthetic riffle and pool data where gating network had physical variables as input.

to be limited just to these.

An interesting result was obtained by examining the behaviour of the gating network over the two experiments. With the first system, where the gating network had the same input as the experts, the two experts were found to classify different qualities, one classified the B1a and B1b classes while the other handled the B2, B3 and B4 classes. It was quite a strong split with over 95% of samples being handled this way. The second system, where the gating network was given physical information on the biotope, the division of the problem between the two experts was between riffle and pool sites and this was a 100% split. So in both cases there was a divergence of what the two individual experts classified.

Another consideration is that the total number of parameters in such models can be particularly high, especially if a large number of experts is used.

This is not such a great problem in reality as the effective number of parameters is much less than the total number of weights, due to the effect of the gating network [121].

## 5.5 Detection of Novel Samples

### 5.5.1 Overview

Once a network model has been developed and is applied in a working system, it can be expected to perform reliably only if the future data presented to it are similar to that used in its training. In more exacting terms, the underlying statistical distribution of new data presented to a trained model should be the same as that data used to estimate the model's parameters. If the data is unusual then the performance of the model is likely to decrease, and this will lead to misclassifications and inaccurate predictions [10]. Ideally, this situation should not arise if the training data is drawn from a sufficiently wide range of conditions, but this cannot be guaranteed. Thus, the ability to derive some measure of the novelty of input patterns would provide a basis for a degree of confidence in the performance of the model.

In networks where there is no normalisation of the output classifications it may be possible to assume a 'doubt' class. For example, if none of the outputs is greater than 0.8 then the pattern classification could be judged to be unreliable. This is almost certainly a poor strategy as there is no guarantee that the outputs will conform in this manner, since it is just as likely that an unusual sample may strongly activate a particular output (this is not the case for localised hidden layer functions, e.g. Gaussian RBFs). Typically the outputs are normalised internally within the network processing, using a softmax construct. With this the outputs are guaranteed to sum to unity, and it would be possible to use the normalising factor to assess the reliability of the classification.

An alternative approach is to use a filter prior to the network to detect novel patterns in the input data. Bishop and James [10] suggested a non-parametric density estimation method based upon Gaussian kernel functions (Parzen windows), and used Bayes' theorem to calculate *a-posteriori* probabil-

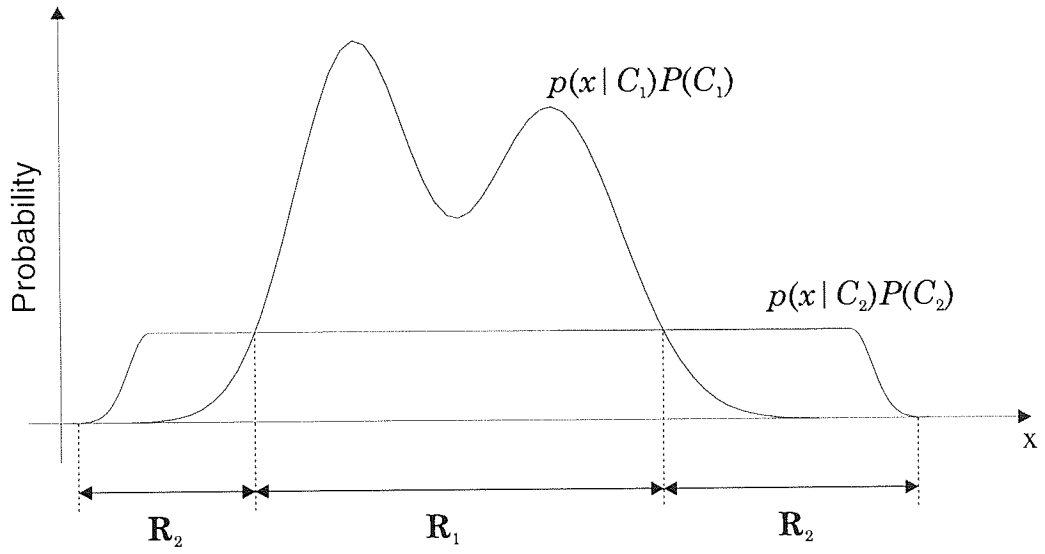



Figure 5.8: Schematic illustration of novelty detection (after Bishop and James [10]). Data which fall in region $\mathcal{R}_2$ are classified as novel.

ities for the predictor vectors. The input data is drawn from one of two classes, each being described by a fixed probability distribution. Class $\mathcal{C}_1$ contains the training and testing data used to infer the model's parameters, while class $\mathcal{C}_2$ denotes the novel data configurations. These two classes are exhaustive and thus the *a-priori* probabilities of class one $P(\mathcal{C}_1)$ and class two $P(\mathcal{C}_2)$ sum to unity, $P(\mathcal{C}_1) + P(\mathcal{C}_2) = 1$. Thus sites can be rejected or highlighted as novel if $P(\mathcal{C}_2|\mathcal{X})$ is greater than $P(\mathcal{C}_1|\mathcal{X})$, where $\mathcal{X}$ denotes the input vector. That is, if the probability that a site belongs to class $\mathcal{C}_2$, given the data, is greater than the probability that it belongs to class $\mathcal{C}_1$, then reject it (Figure 5.8).

### 5.5.2 Procedure

To test out these ideas in the context of this dissertation the synthetic data sets (Section 4.4.2) were used. The training set consisted of 1800 samples of the riffle data. Two test data sets were used, both of which had 1600 samples, one set representing riffle sites (drawn from the same underlying distribution as the training data) and the other representing pool sites (which is appreciably different from training data).

A Parzen estimator, using Gaussian kernels, was used for the density estimation, with the total likelihood for a particular data vector $\mathbf{x}$ given by:

$$p(\mathbf{x}) = \frac{1}{n(2\pi)^{(d/2)}\sigma^d} \sum_{p=1}^{n} \exp\left\{-\frac{|\mathbf{x} - \mathbf{x}^p|^2}{2\sigma^2}\right\} \tag{5.9}$$

where $\mathbf{x}^p$ is a pattern from the training set (which has $n$ patterns in total), $d$ is the dimensionality of the input space and $\sigma$ is the smoothing parameter that controls the width of the Gaussian kernel. A well documented problem associated with the above models is that of increasing dimensionality of the density estimation space, and for this reason the dimensionality of the data sets were reduced from their original size of 41 dimensions (each dimension representing a single taxon) to a five dimensional space using a PCA analysis. In addition a MLP having 5 inputs, 5 hidden and 5 output units was trained using the training data and a validation data set. Training was halted when the validation error began to increase, and at this point the two test data sets were processed through the network.

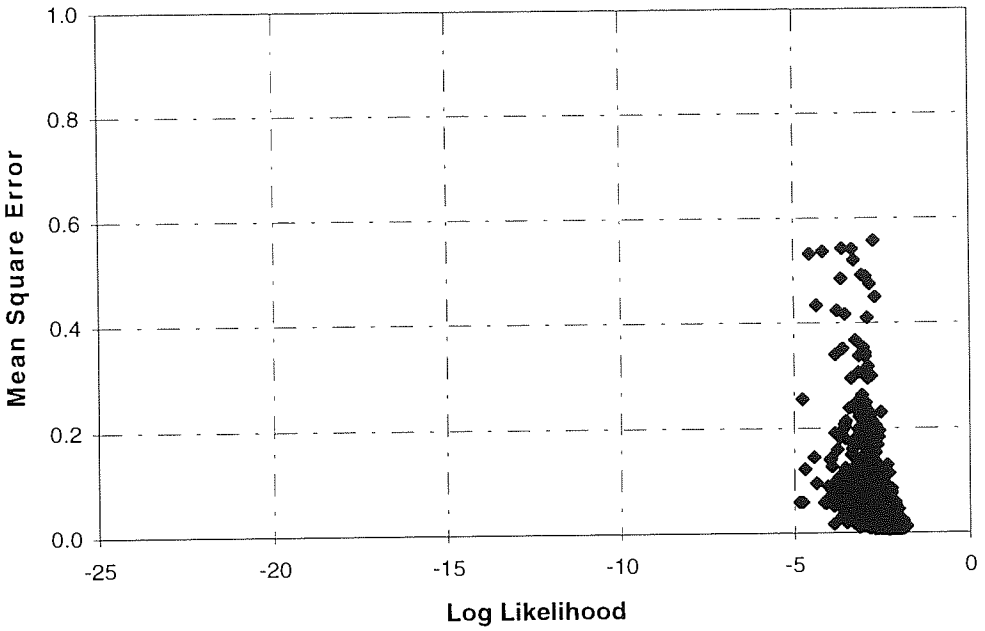
The width parameter for the window functions was set using the following simple heuristic; set the standard deviation $\sigma$ to the average distance of the five nearest neighbours.

### 5.5.3 Results

Table 5.8 shows the confusion matrices for the combined riffle and pool test data as derived from a MLP trained on riffle data only. As expected the MLP produced a greater classification rate for the riffle data (93.8%) than the pool data (71.9%). The riffle sites were all classified correctly to within a single class, unlike the pool sites where there was a noticeable degradation of the classification rate because of the absence of the more sensitive indicators. The MSE from these classifications were plotted against the log likelihood of the unconditional probability density (Equ. 5.9) from the novelty detector. These are shown in Figure 5.9.

Examination of Figure 5.9 (a) shows that most (i.e. more than 90%) of the data fall within a small region of the graph. These points have small MSE's, which is indicative of correctly classified patterns. The important item to

a) Riffle test data




b) Pool test data

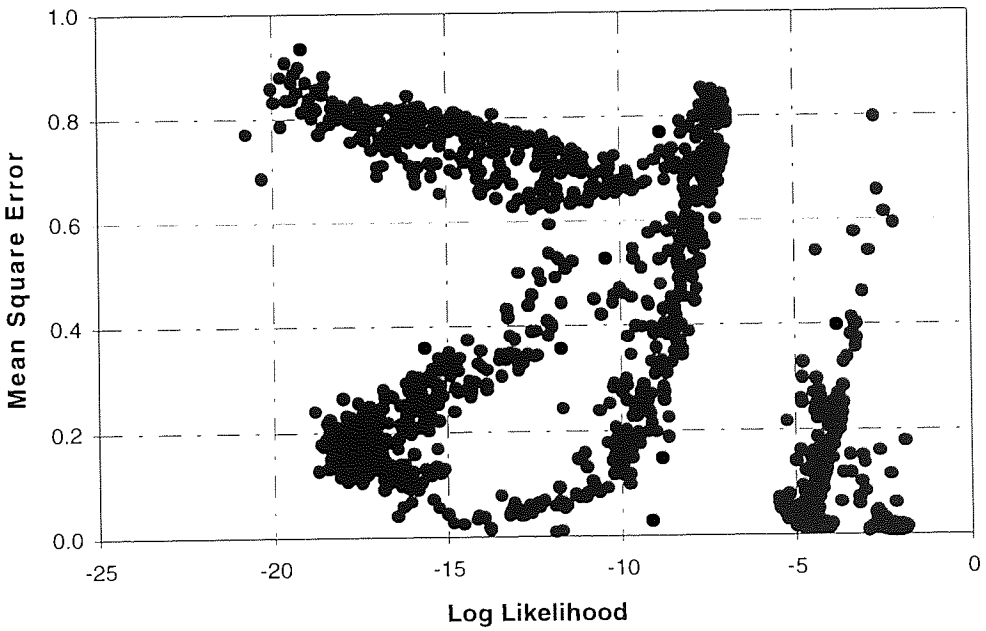



Figure 5.9: Plot of MSE against log likelihood from the density estimation for synthetic riffle and pool test data. The region to the left of a log likelihood value of -6 can be regarded as representing novel data.

a) Riffle test data

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 310 | 10 | 0 | 0 | 0 |
| | B1b | 3 | 293 | 24 | 0 | 0 |
| | B2 | 0 | 20 | 278 | 22 | 0 |
| | B3 | 0 | 0 | 9 | 304 | 7 |
| | B4 | 0 | 0 | 0 | 5 | 315 |

Classification rate = 93.8%

b) Pool test data

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 178 | 138 | 14 | 0 | 0 |
| | B1b | 5 | 181 | 103 | 31 | 0 |
| | B2 | 0 | 1 | 208 | 89 | 22 |
| | B3 | 0 | 0 | 18 | 268 | 34 |
| | B4 | 0 | 0 | 0 | 5 | 315 |

Classification rate = 71.9%

Table 5.8: Confusion matrices for synthetic riffle and pool test data for a MLP trained only with riffle data.

note is that all the points have a log likelihood of greater than $-5.0$. This is in contrast to the pool data of Figure 5.9 (b) where the majority of the sites have a log likelihood of less than $-7.0$. There is a clear division between normal (riffle) and novel (pool) data, and the majority of the significant misclassifications would be eliminated if a threshold for the log likelihood of -6.0 was adopted. For reference, in Figure 5.9 (b) the majority of the points with a log likelihood of greater than $-6.0$ (i.e. data which can be considered as normal) are from classes B4 and B3, and it is these classes which have the greater degree of similarity between the riffle and pool data.

### 5.5.4 Discussion

The above results demonstrate that it is possible to define a measure of normality associated with the input data. For this particular example, normal

was assumed to be representative of a riffle community and it would be possible to pick out samples that differed from this assumption. But the technique is not limited to just the riffle/pool situation, it may be possible to pick out other examples (e.g. heavy metal pollution) where the sample is unusual. This should be viable, as throughout the various elicitation sessions the Expert could identify unusual samples and comment to the effect that the original 'organic pollution only' assumption did not apply to that particular sample. The experiment assessed novelty using the invertebrate data only, so if biotope information were included a greater level of discrimination between riffle and pool data could be expected, and thus a corresponding increase in the difference between normal and novel. In addition, it should be possible to detect novelty in the physical characteristics of sites, which could be useful in models which incorporate this type of information (e.g. Section 5.4).

The question of what constitutes novel data is an important one. Whichever method of novelty detection is used it will require an implicit assumption to be made about the distribution of the outliers. In this particular example, a uniform distribution was assumed over the 5-dimensional PCA space, but this may not be applicable in all situations, especially where different scalings or transformations are used for the various components of the input vector. Additionally, there are implicit assumptions associated with each model. The main one associated with Parzen windows is that the window width and shape is the same for all regions of the space of interest. This may restrict the efficient modelling of the unconditional data density. Other models, for example Gaussian mixture models, could have been used for the density estimation, but there are a number of trade offs between the complexity, speed of construction and evaluation of the model that must be taken into account.

The clustering of the data in Figure 5.9 is an artifact of the synthetic data. If 'real' data were to be used, a more uniform density for the log likelihood would occur, and the boundary between normal and novel could be less clear cut. However, the simple heuristic used to set the width parameter for the Parzen windows could be improved upon to yield better discrimination, simply by using cross validation data. A model has been described in the literature [143] which determines the threshold as part of the training process, but

this uses a more complex algorithm which incrementally adjusts the number of components of the density estimator.

## 5.6 Encoding Prior Information

### 5.6.1 Overview

This section is concerned with the incorporation of prior knowledge within a neural network. Up to this stage the neural net models have not utilised any information which has been elicited from the domain expert (except for the sample classifications). It appears that the knowledge-based systems and the neural networks occupy different ends of the spectrum as far as the use of prior information is concerned. The knowledge-based systems use only the subjective, probabilistic domain knowledge and do not modify their knowledge base to reflect the test data, while the neural networks use only information from the data and make no use of the elicited probabilistic information. Learning in probabilistic knowledge-based systems has been described [159], as has the use of objective probabilities (i.e. frequencies from data sets) to refine the knowledge base in an expert system [161]. But subjective domain knowledge has not, if ever, been explicitly used to train (or prime) a neural network to classify real world data.

The use of prior information in neural networks to explicitly encode the underlying functional relationship between input and output has been described (e.g. monotonic functions [73]). Prior knowledge has also been used to determine suitable topologies. These are both more implicit in their approach to the use of prior information, the estimation of the parameters is still only reliant on the data sets. The principal idea in this section is that it should be possible to use subjective domain knowledge in the training process to facilitate the development of a more robust classification tool.

### 5.6.2 Procedure

The most practical method of introducing prior knowledge was to augment the Severn-Trent training data with samples of the synthetic riffle data. For

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 46 | 11 | 1 | 0 | 0 |
| | B1b | 9 | 45 | 16 | 1 | 0 |
| | B2 | 0 | 19 | 60 | 22 | 2 |
| | B3 | 0 | 0 | 7 | 24 | 4 |
| | B4 | 0 | 0 | 0 | 15 | 10 |

Classification rate = 63.4%

Table 5.9: Confusion matrix for the Severn-Trent test data from an MLP using a ratio of 2:1 Severn-Trent/synthetic riffle data for the training data.

the preliminary experiments a number of different ratios of Severn-Trent to synthetic data, for use as training data, were experimented with. The MLPs were trained using the data sets of Table 5.1 with 8 hidden units and a weight-decay $\lambda$ of 0.1.

### 5.6.3 Results

It soon became clear that use of the synthetic data was not improving any of the models' performances. For the Severn-Trent data the average classification rate, using only the invertebrate sample as input, would be around 65%-70% (see Table 5.2). Virtually every network with a ratio of Severn-Trent/synthetic data of less than 1:1 (i.e. 1:2, 1:4), that is more synthetic data than real, had a classification performance of less than 50%. The classification rate improved to over 50% only when there was proportionally more Severn-Trent data than synthetic in the training set. The most successful ratio was that of 2:1 Severn-Trent/synthetic, an example of which is given in Table 5.9. Here the classification rate improved to over 60%, which is still lower than was previously obtained with just the Severn-Trent data. The misclassifications are spread uniformly over the 5 classes.

Again, of the different encoding formats the most successful was the 16 PCA method.

### 5.6.4 Discussion

Unfortunately the results were disappointing and contrary to expectation. The use of prior information (i.e. the synthetic riffle data) did not improve the classification accuracy of the MLP for the Severn-Trent data. It is difficult to pinpoint the reason (or reasons) why this should be so, but perhaps the most likely explanation is that there is some interference effects occurring between the Severn-Trent and synthetic datasets. This is similar to those effects demonstrated by the riffle and pool data (see Section 5.4 and Table 5.5).

The use of subjective expert knowledge for training neural networks will, undoubtedly, become more common place. For example, Abu-Mostafa [1] refers to systems learning with 'hints', which is very much along the lines of the efficient use of prior knowledge. But as demonstrated by the results of this section, care must be taken with the use of prior knowledge, as it can degrade, as well as improve, the models performance.

## 5.7 Self-Organising Maps

### 5.7.1 Introduction

Self-organising maps (SOMs) constitute a popular tool for the visualisation of complex experimental data [83]. They form a topology preserving nonlinear projection from the dimension of the input vector to, typically, a two dimensional space. Input vectors which are similar will be mapped to a similar region of the two dimensional map. They use an unsupervised learning algorithm, which means that classification labels are not used to form the mapping. If the classification is known, then it is possible to use a corresponding supervised learning algorithm, called Learning Vector Quantization (LVQ). The main utility of the SOMs is that they are powerful visualisation tools, especially useful for the analysis of time series problems where a trajectory can be followed on the feature map. Other unsupervised visualisation tools are available, for example Sammon's mapping [152] and various multi-dimensional scaling algorithms.

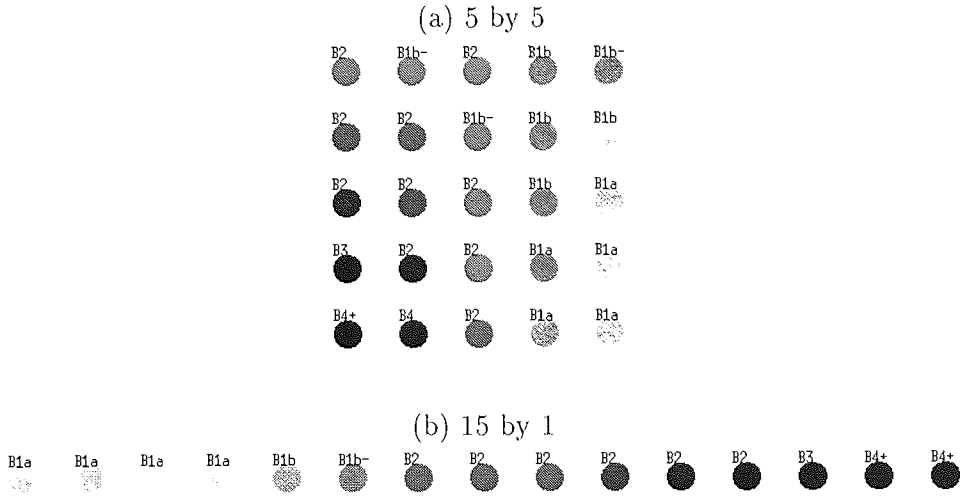



Figure 5.10: Two Self-Organising Maps generated from the Severn-Trent data.

### 5.7.2 Mapping the Severn-Trent Data

Using the Severn-Trent data (Section 4.3) two different topologies of feature map were trained (5x5, and 15x1) using the SOM_PAK software. This was an explorative exercise to see whether the SOM were able to extract a meaningful interpretation of the biological classification without prior knowledge of this classification. Figure 5.10 shows the two trained feature maps. The thirteen division classification was used to interpret the maps, but these labels were only added after the training process had been completed.

#### 5x5 Map

This map, Figure 5.10 (a), shows a good ordering of the classes from the lower right hand corner, round in an anti-clockwise direction to the bottom left hand corner. There is a fold (or discontinuity) in the map starting from the bottom edge, lying vertically up two layers. There is a clear trend from the best quality classes (B1a) to the poorest (B4), with the B2 class occurring with the greatest frequency.

### 15x1 Map

This linear feature map, Figure 5.10 (b), shows an excellent progression from good quality classes on the left, through the intermediate classes to the poor quality classes on the right. The ordering is clearly apparent and in good agreement with the classification system. As was previously stated, although the classes are ordered (i.e. B1a > B1b > ... > B4) there is no information on the spread of the class. This is reflected somewhat in the proportion of the B2 which are present (some 6 out of 15 units). It could be argued that the reason for the large presence of B2s was they formed a larger proportion of the data set, but if this was the case then the B1bs should be next best represented class, but this is clearly not so.

## 5.7.3 Discussion

As can be seen from Figure 5.10 the two SOMs do represent mappings that can be intuitively interpreted with respect to the biological classification used in this dissertation. Although, this is not conclusive or incontrovertible evidence that the particular classification is the best method of summarising the data, it does provide supportive evidence that this classification offers a powerful means of interpreting the invertebrate communities.

If the labels are known, then it is possible to prime the SOM's so that the position of key classes or groups are fixed beforehand. This would enable the desired representation of samples to be forced onto specific locations. For example, this would be useful for the mapping described by Walley [168], where a truncated diamond topology (with good quality sites at the top and poor sites at the base) for the biological classification of river water quality is suggested. For a surveillance program of a single site it would be possible to plot the trajectory of the classification over time, and this trajectory would be an excellent visual means of presenting the results of a surveillance program.

Manual calibration of the regions of the resulting map was aided by knowing the classification of individual samples. If these classes were unknown, then interpretation of the resulting maps would be considerably more difficult. There would be a possibility of reading too much meaning into the structure

of a SOM, for it is possible that some latent artifact of the data influenced the resultant mapping.

## 5.8 Comparison to the BERT System

Although outside the scope of this dissertation from an experimental perspective, a qualitative comparison is made here between the neural network models of this project and the knowledge-based approach of the BERT system [170, 169]. The work in this dissertation was influenced by the BERT system, and draws on many of its strengths, although there are some significant differences.

The main difference is in the process by which each method obtains its 'knowledge', which enables it to perform the classification. The BERT system is primed with the expert knowledge, in fact the main component is the knowledge base, the construction of which was not a trivial exercise. At present, BERT does not update its database of probabilities in the light of 'real world' data, that is it does not 'learn'. This is in contrast to the typical neural network approach, where no prior knowledge is utilised, and the learning is solely governed by the available 'real world' data. Two half-way-houses are possible; one where the BERT system learns, the other with some prior knowledge encoded into the neural networks. The latter was found to produce poor results in Section 5.6, while the former has not been tried, but would not be difficult to implement.

Perhaps the biggest obstacle for the BERT system would be its extension to incorporate other information, such as environmental attributes, in the derivation of the classification. With the MLP's this would not be a problem as additional inputs could easily be used. Also, the Bayesian based BERT system is readily formulated in terms of discrete attributes, but its extension to continuous values would increase its mathematical complexity quite substantially, whereas this would present no difficulty for the MLP networks.

Although the identification of unusual samples was discussed earlier (Section 5.5), the BERT system takes a different approach to this problem by identifying individual taxon as either rogues or victims if the evidence provided by that taxon is not in keeping with that provided by the other taxa in

the sample. Rogues are taxa which are unexpectedly present and distort the conclusion of the system by providing contradictory evidence. Rogues can be identified and removed by the use of the conformity index. The conformity index can also be used to identify victims and to define sample consistency. Victims are taxa which would be expected to be present in the sample but are absent. Sample consistency provides a measure of the agreement between all of the individual pieces of evidence provided by the indicator taxa. Decisions on rogues and victims are made within a probabilistic framework, relying on the knowledge base, and offer a more detailed interpretation than that possible with the density estimation method discussed in Section 5.5.

The Bayesian methods of BERT and the neural networks are essentially complementary in nature, and there is no reason why the two methods could not be combined. For example, where the classifications of a number of models are averaged the models do not have to be similar, indeed it may make for a very much more robust system if the models are different (orthogonal) in nature. Walley [168] experimented with this idea, and showed that a combined system of a Bayesian model and a MLP provided the most reliable classification.

## 5.9 Summary

This section has demonstrated the utility of neural networks in a number of different applications relating to the interpretation and classification of benthic macroinvertebrate samples. The basic neural network model performed well. In tests based on the Severn-Trent data it proved to be relatively insensitive to the number of hidden units and regularisation procedures. In addition, when individual networks were combined to form committees of networks there was only insignificant improvements in performance. In contrast it was found that by using an additional input, which reflected the sample diversity, classification accuracy improved by 10%.

Following on from the development of networks for riffle biotopes, a modular architecture was used to develop networks for use on a mixed data set containing samples taken from different biotopes. These were trained and tested on the synthetic data and clearly demonstrate the value of such mod-

els. Another important topic was the reliability of the models when used on new and novel data. Like all regression models, neural networks are unreliable when used to extrapolate, so doubt must be cast on predictions made from data which, by some measure, is different from that which was used to train the model. By using a novelty detector prior to processing it is possible to determine when a normal or abnormal input vector has been presented to the network. This permits unusual or novel samples to be highlighted and flagged for special attention. The use of unsupervised learning procedures, based on Kohonen self-organising maps, demonstrated that the biological classification system used was reasonable, as the resulting maps could be easily interpreted by using the biological class labels.

# Chapter 6

# Selection of Key Indicator Taxa

## 6.1 Introduction

Taxa which are frequently used by biologists to identify specific ecological conditions are commonly referred to as 'indicator' or 'key' taxa. This dissertation modifies this loose definition to include the selection of indicator taxa for use in numerical classification and prediction models. In any well designed study the samples taken should provide sufficient discriminatory information about the subject of interest to permit reliable classification and/or prediction. Typically, the sample information is imperfect and limited by the practical confines of the study. For example, the microorganisms known as *Sewage fungus* are excellent indicators of fairly heavily polluted conditions [98], but their presence is often not reported because they are not benthic macroinvertebrates. A comprehensive study would utilise information of many faunal and floral groups [65, 139, 51], as the whole range of ecologists' expertise could then be drawn upon. No group would be excluded if it provided telling evidence as to the state of the river or its environment. Unfortunately, the assimilation of data from many groups happens only occasionally; so the best must be made of what is available, which is typically the macroinvertebrate data.

Section 6.2.1 briefly discusses what qualities biologists or ecologists consider when identifying indicator taxa, while from Section 6.3 onwards the emphasis is directed towards the selection of taxa which can be considered good indicators of quality class. Three different methods, of which one is a novel implementation based on a hybrid frequentist-Bayesian approach (referred to as the RMS-D method), are described and compared using the Severn-Trent sample

database of 292 classified samples (hereafter referred to as the 'Severn-Trent 292'). The indicator indices allow the change in information (or uncertainty) to be quantified for different levels of taxa identification and enumeration. A new method of encoding the input abundances, using an intermediate step from one of the selection methods, is tested and compared to the performance of the schemes used in Section 5.3.4. The chapter concludes with a summary of the findings.

## 6.2 Indicator Taxa

### 6.2.1 Biological and Ecological

In freshwater biomonitoring the term 'indicator species' is used to describe taxa that have well documented sensitivities to specific chemical and/or physical parameters. Changes in environment cause reactions in the community which can include changes in population, morphology, physiology or behaviour [75], and these need to be accurately defined if a taxon is to be considered suitable as an indicator. The environmental stresses that are reflected by changes in the community structure can be regarded as either abiotic (e.g. chemical, trace metal) or biotic (e.g. niche competition, predation) in origin.

Johnson et al. [75] and Hellawell [56] list the characteristics that should be attributed to a good indicator taxon. The taxon should:

*i.* be easily recognised and taxonomically sound,

*ii.* have a cosmopolitan distribution,

*iii.* occur frequently and in reasonable numbers,

*iv.* be suitable for field sampling and laboratory procedures,

*v.* be relatively sedentary and have a long life history, and

*vi.* have reasonably well defined ecological demands.

Experimental work has provided tolerance ranges for certain taxa, but these commonly refer to single toxicants or isolated conditions which are unlikely to

occur in the field. The transfer of laboratory work to field interpretation is also complicated by synergistic or antagonistic interactions among the sources of stress. Thus, in the laboratory the concentration of zinc which kills 50% of of individuals may be calculated, but this may change dramatically depending on the acidity of the waters, thus a single recommended safe concentration of zinc would be misleading. Taxa can also be used as sentinel organisms where their tissue is examined and its concentrations of any contaminants are extrapolated to infer the in-situ concentrations. This is becoming more common, but at present does not lend itself to routine monitoring [75].

### 6.2.2 Variables for Use in Computer Models

A common desire is to select the most appropriate indicators for use as predictors in the particular model under consideration. In neural network models, especially where the number of training patterns is small, it is important to remove the inputs that are either redundant or spurious to the problem at hand. For Bayesian knowledge-based systems, the identification of good indicators would help direct the emphasis of any direct knowledge elicitation (see Section 4.2.3), and reduce the time spent on elicitation and thus partly overcome the knowledge acquisition bottleneck. Also, by reducing the quantity of data needed within the system, the incidence of conflicting data will be reduced.

The selection of feature variables is common to most studies [154], and the criteria for selecting the most suitable subset ultimately depends on the underlying study [101]. An example is the difference between models which aim for a low error rate, allocatory, and those which maximise the separation between groups, separatory. McLachlan [101] demonstrates that the use of allocatory and separatory criteria can lead to different results in the reduction of the feature vector's dimensionality. When selecting a subset of variables from a large set it is important to note that a selection bias will be introduced [101]. Methods are available for the reduction of selection bias but involve a high computational penalty.

The necessity of identifying good indicators is particularly pertinent to the discipline of biological monitoring, because of the great number of possible sensors that are available to be used. For example, in the current database of

British Freshwater Animals [96][1] there are approximately 3000 species listed, with around 1500 species of benthic invertebrates. Not all of these, however, would be suitable as indicators, either due to taxonomic difficulties or low incidence in UK waters. In this study the data sets were supplied by outside bodies, thus the collecting strategy and level of identification were predetermined. Clearly, this constrained the scope of the study, in terms of the range and level of taxa under consideration. It also increased uncertainty in the data, due to problems of interpreting exactly what the data represented. Despite these constraints the results will clearly demonstrate the value of the methods developed and provide a sound basis on which to design a more comprehensive study.

The automation of the methods for the selection of indicator taxa would be particularly useful where expert knowledge is unavailable or does not exist. For example, the acidification of streams can be assessed by chemical means and the resulting pH range, say 4.0 to 8.0, can be banded into different classes. The conditional probabilities of the taxa and the class can be calculated, and from this the key taxa can be selected using one of the methods discussed in the following section. This does not utilise any domain knowledge, and so the following methods can be thought of as being problem independent.

## 6.3 Determination of Indicator Taxa

Three methods of quantifying the indicator values of taxa, and thus of identifying key taxa, are considered in the following sections. There are:

*i.* a novel implementation based upon a frequentist Bayesian approach (the RMS-D method),[2]

*ii.* a method taken from information theory, namely mutual information, and

*iii.* a stepwise discriminant procedure.

[1] Currently maintained by Dr M.T. Furse, at the IFE.
[2] W.J. Walley, personal communication

The desire is to be able to rank candidate input variables (i.e. taxa) into a order based on their usefulness for prediction and classification. This information could be used to reduce the dimensionality of the input vector in neural net models, and hence help to overcome over-fitting problems by reducing the number of weights needed in the networks.

The desirable properties of such an indicator index would be that it varies from zero to an upper bound value; with zero representing the poorest indicator possible (i.e. a taxon that provides no information as to the class of the river system) and the upper bound representing the perfect or absolute indicator. Also, there is a trade-off between the taxa occurring very frequently, which would imply poor discriminatory power between classes, and very infrequently, which would imply a restricted utility and a reliance on other taxa being present in order to provide routine classifications. This trade-off should be reflected in the indices.

### 6.3.1 The RMS-D Method

#### 6.3.1.1 Philosophy

The RMS-D (Root Mean Squares of the Deviations) method is based upon utilising the frequencies of the taxa derived from an analysis of the data set. It considers each taxon individually and also each state of abundance as being independent. For each taxon an expression is derived to quantify the information provided by each state, and these are combined to give an overall indicator value for each taxon. These indicator values provide a basis on which decisions can be made concerning which taxa to include in the model.

#### 6.3.1.2 Derivation

Using the distribution of *Asellus aquaticus* and *Gammarus pulex* from the Severn-Trent 292 database as an example, the first step is to derive the probability for each taxon state given the quality class, $P(e_{ik}|H_j)$, where $H_j$ is the $j$th class from B1a, B1b, ..., B4 and $e_{ik}$ the $k$th state of the $i$th taxon (e.g. *A. aquaticus* is *few*). The frequencies of these occurrences, expressed as probabilities normalised within each class, are shown in Table 6.1. However, it is the probability of the water class given the state of the taxon, $P(H_j|e_{ik})$,

*Asellus aquaticus*

| Class | Abs | Pres | Few | Com+ |
|---|---|---|---|---|
| B1a | 0.91 | 0.06 | 0.03 | 0.00 |
| B1b | 0.52 | 0.11 | 0.27 | 0.10 |
| B2 | 0.14 | 0.09 | 0.24 | 0.53 |
| B3 | 0.14 | 0.11 | 0.31 | 0.43 |
| B4 | 0.52 | 0.08 | 0.36 | 0.04 |

*Gammarus pulex*

| Class | Abs | Pres | Few | Com+ |
|---|---|---|---|---|
| B1a | 0.35 | 0.07 | 0.43 | 0.16 |
| B1b | 0.14 | 0.14 | 0.11 | 0.61 |
| B2 | 0.31 | 0.11 | 0.36 | 0.23 |
| B3 | 0.94 | 0.03 | 0.03 | 0.00 |
| B4 | 1.00 | 0.00 | 0.00 | 0.00 |

Table 6.1: Conditional probabilities, $P(e_{ik}|H_j)$ for *Asellus aquaticus* and *Gammarus pulex* derived from the Severn-Trent 292 data.

that we wish to use as the basis of the indicator values. It should be noted that $P(H_j|e_{ik})$ in this case is based on $P(H_j) = 0.2$ (i.e. so the prior probability of each class is equal, the Principle of Indifference is assumed to apply). In other words, the system is based on the evidence provided by the sample alone, not the sample plus prior knowledge of the frequencies of the classes $(H_j)$. We are starting with an open (unbiased) mind and judging the quality of the indicators by the evidence presented by the data.

From Table 6.1 the marginal probabilities $P(e_{ik})$ can be calculated from:

$$P(e_{ik}) = \sum_{j=1}^{5} P(e_{ik}|H_j)P(H_j) \tag{6.1}$$

where $P(H_j) = 0.2$ from the Principle of Indifference. Then using Bayes' formula the $P(H_j|e_{ik})$ values can be calculated from:

$$P(H_j|e_{ik}) = \frac{P(e_{ik}|H_j)P(H_j)}{P(e_{ik})} \tag{6.2}$$

with $P(e_{ik}|H_j)$ given in Table 6.1, $P(e_{ik})$ from Table 6.2 and $P(H_j) = 0.2$.

Using the $P(H_j|e_{ik})$ values from Table 6.3 the indicator value, $\mathcal{IV}_{ik}$ for a taxon $e_i$ in state $k$ can be defined as:

$$\mathcal{IV}_{ik} = \left[\sum_j \Big(P(H_j|e_{ik}) - P(H_j)\Big)^2\right]^{1/2} \tag{6.3}$$

This is shown graphically in Figure 6.1. The shaded area is the difference

| | *Asellus aquaticus* | | | |
|---|---|---|---|---|
| | Abs | Pres | Few | Com+ |
| $P(e_{ik})$ | 0.45 | 0.09 | 0.24 | 0.22 |

| | *Gammarus pulex* | | | |
|---|---|---|---|---|
| | Abs | Pres | Few | Com+ |
| $P(e_{ik})$ | 0.55 | 0.07 | 0.19 | 0.20 |

Table 6.2: Marginal probabilities, $P(e_{ik})$, for *Asellus aquaticus* and *Gammarus pulex* calculated from Equ. 6.1.

| | *Asellus aquaticus* | | | |
|---|---|---|---|---|
| Class | Abs | Pres | Few | Com+ |
| B1a | 0.41 | 0.12 | 0.03 | 0.00 |
| B1b | 0.23 | 0.25 | 0.22 | 0.09 |
| B2 | 0.06 | 0.19 | 0.20 | 0.49 |
| B3 | 0.06 | 0.26 | 0.26 | 0.39 |
| B4 | 0.23 | 0.18 | 0.30 | 0.04 |

| | *Gammarus pulex* | | | |
|---|---|---|---|---|
| Class | Abs | Pres | Few | Com+ |
| B1a | 0.12 | 0.20 | 0.46 | 0.16 |
| B1b | 0.05 | 0.41 | 0.12 | 0.61 |
| B2 | 0.11 | 0.30 | 0.38 | 0.23 |
| B3 | 0.35 | 0.08 | 0.03 | 0.00 |
| B4 | 0.37 | 0.00 | 0.00 | 0.00 |

Table 6.3: Conditional probabilities, $P(H_j|e_{ik})$, for *Asellus aquaticus* and *Gammarus pulex* derived from the Severn-Trent 292 data using Equ. 6.2.

between the probabilities for each state and the non-informative probabilities. The square of the distance is taken for all values, and the square root of the sum of these squares is calculated. These values are used later in Section 6.6 as an alternative input encoding format.

Equ. 6.3 gives the indicator value for each taxon ($i$) in each of its possible states of existence ($k$), including absence. The value of knowing that the state of a given taxon is *few* could be compared to that of any other given state via these values. But if we were are to rank taxa in terms of their overall value as indicators, we need to derive an overall indicator value for each taxon. Two methods are possible: one is to take the average value of the $\mathcal{IV}_{ik}$ values, the other is to weight each $\mathcal{IV}_{ik}$ by the frequency of each state. Thus the average is given by:

$$\mathcal{IV}_i^a = \frac{1}{K}\sum_k \mathcal{IV}_{ik} \tag{6.4}$$

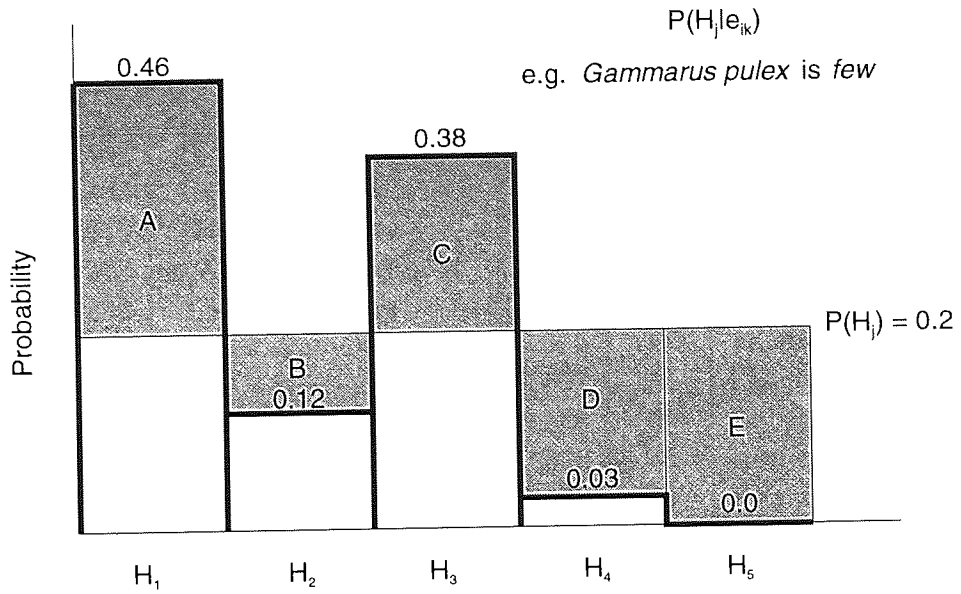



Figure 6.1: Graphical representation of the basis of RMS-D method. The shaded areas (A-E) are used in the calculation of the indicator value. If the taxon was always absent or always present the shaded area would be zero, and the indicator value would be zero. Probabilities shown are taken from Table 6.3.

| | $\mathcal{IV}^a$ | $\mathcal{IV}^w$ |
|---|---|---|
| *Asellus aquaticus* | 0.263 | 0.286 |
| *Gammarus pulex* | 0.386 | 0.359 |

Table 6.4: Indicator values for *Asellus aquaticus* and *Gammarus pulex*.

where there are $K$ states in total, and the weighted value by:

$$\mathcal{IV}_i^w = \sum_k P(e_{ik})\mathcal{IV}_{ik} \tag{6.5}$$

where $P(e_{ik})$ is given in Table 6.2. Table 6.4 shows both indicator values for the *A. aquaticus* and *G. pulex* data of Table 6.3.

The two methods of deriving the indicator values, Equ. 6.4 and Equ. 6.5, were compared using the Severn-Trent 292 sample database (see Section 4.3). Table 6.5 show the best 10 taxa selected by both methods, where the data

| | $\mathcal{IV}_i^w$ | | | $\mathcal{IV}_i^a$ | | |
|---|---|---|---|---|---|---|
| | Score | Freq. | Taxon | Score | Freq. | Taxon |
| 1 | 0.381 | 185 | Gammaridae | 0.673 | 3 | Brachycentridae |
| 2 | 0.363 | 153 | Baetidae | 0.648 | 18 | Chloroperlidae |
| 3 | 0.344 | 150 | Tipulidae | 0.639 | 47 | Leuctridae |
| 4 | 0.318 | 75 | Heptageniidae | 0.604 | 39 | Nemouridae |
| 5 | 0.296 | 179 | Asellidae | 0.595 | 57 | Perlodidae |
| 6 | 0.296 | 134 | Ancylidae | 0.575 | 17 | Taeniopterygidae |
| 7 | 0.286 | 69 | Rhyacophilidae | 0.552 | 20 | Piscicolidae |
| 8 | 0.282 | 57 | Perlodidae | 0.546 | 7 | Phryganeidae |
| 9 | 0.271 | 89 | Elminthidae | 0.538 | 75 | Heptageniidae |
| 10 | 0.266 | 87 | Limnephilidae | 0.531 | 69 | Rhyacophilidae |

Table 6.5: Comparison of two probabilistic indicator indices with the best 10 taxa listed in descending order. The was a total of 292 samples.

has been grouped to family level.[3] From Table 6.5 it is apparent that there is a considerable difference between the values produced by the two methods. There is a trade off between the frequency of the taxa and its value as an indicator. If the taxa is extremely frequent and is commonly present in all quality classes then its value as an indicator is very limited (however its absence may be important in determining the cause of pollution). The best indicators are the ones which appear only in one class, but these tend to have a low frequency within the data set. These are good indicators, but are of limited use because of their infrequency. Thus of the two indices, the averaging method ($\mathcal{IV}^a$, Equ. 6.4) identifies the best indicators without regard to frequency, while the weighted method ($\mathcal{IV}^w$, Equ. 6.5) takes frequency into account and provides the best overall measure of indicator value. The values of $\mathcal{IV}_i^w$ are used in preference to the $\mathcal{IV}_i^a$ values in Section 6.4 to compare the three indicator selection methods. Also, the weighted method shows a good correspondence to the taxa selected for use in the BERT system, and this is investigated further in Section 6.4.

[3]Note the derivation of the 2 probabilistic indicators were demonstrated using species level data on *A. aquaticus* and *G. pulex*.

## 6.3.2 An Information Theoretic Approach

### 6.3.2.1 Review

Henery [57] discusses information theoretic measures, and Jones [76] provides an introduction to information theory. Other related work includes Battiti [6], Kanaya and Nakogawa [80] and Karin [82]. Some of the diversity indices used in biomonitoring are in fact derived from information theory, for example the Shannon-Weiner Index (Equ. 2.3). The following sections provide details of the further development of these indices. The techniques discussed in this section are suitable for discrete attributes but can be adapted for continuous attributes.

#### Entropy

Entropy may be regarded as a measure of uncertainty, or randomness, of an attribute, and is defined as:

$$H(E) = -\sum_{k=1}^{n} p_k \log(p_k) \tag{6.6}$$

where $p_k$ is a probability of an event $P(E_k)$. If $p_k$ is zero, then zero is assigned to the indeterminate $p_k \log(p_k)$ expression. Also the entropy of a variable can never be negative. If there is complete certainty the entropy is zero, which is the lower bound, while the upper bound occurs when all events are equally probable (i.e. maximum uncertainty), and is equal to $\log n$ since $p_1 = p_2 = \cdots = p_n = 1/n$. Thus, if a variable has attributes that do not vary then it is not possible to discriminate between classes, and therefore the variable has no utility as a predictor.

### 6.3.2.2 Mutual Information of Class and Attribute

The amount of common information, or entropy, between two variables, $C$ and $E$ is given by:

$$M(C, E) = \sum_{ij} p_{ij} \log \left( \frac{p_{ij}}{\pi_i q_j} \right) \tag{6.7}$$

where $p_{ij}$ is the joint probability of observing class $C_i$ and the $j$th state of variable $E$, $\pi_i$ is the marginal probability of Class $C_i$ and $q_j$ the marginal probability for the state $E_j$. The mutual information, $M(C, E)$, will be zero if and only if $C$ and $E$ are independent [126, 42], while the maximum value for the mutual information of a system cannot exceed the sum of their separate entropies [76]. In the machine learning community the mutual information is useful as a splitting criteria for decision trees [57]. Henery [57] also discusses other information theoretic expressions for equivalent numbers of attributes, noisiness and the identification of irrelevant variables.

Pearl [126] also comments on the use of sensitivity matrices to assess the utility of information, but considers these to be too detailed a measure. The deficiency of the mutual information is that scale or ordering information is not considered in the values that a variable can take [126].

### 6.3.3 Classical Methods

Three methods are generally available in statistical packages to select predictor variables for parametric regression and discriminant analysis. The three methods are referred to as forward, backward and stepwise selection procedures. In forward selection, the model initially contains no predictor variables, for each step the variable that adds most to the discriminatory power of the model is selected and this continues until none of the unselected variables meet the selection criteria. For backward elimination the model starts with all possible predictor variables, and for each step the variable contributing least to the discriminatory power of the model is removed. Again, this is continued until all the remaining variables meet the criteria to stay in the model. Stepwise selection is similar to forward selection except that variables can be removed as well as added to the model. The model is examined at each step and selection is stopped when there is no unselected variable that meets the criterion to enter the model, and none of the selected variables are suitable to be eliminated. The selection criterion is based either on multiple correlation coefficients or the residual variance of the model, with the relevant statistical test being the Wilks' $\lambda$. Only one variable at a time can be added to the model, and no account is taken of the relationships between the unselected variables. Stepwise

discrimination does not always produce the best model, and Wilks' $\lambda$ may not be the most suitable test of discriminatory power [153].

## 6.4 Comparison of Indicator Expressions

### 6.4.1 Procedure

Using the 292 Severn-Trent data set all three of the above methods, namely the RMS-D weighted index, the mutual information and the stepwise analysis, were applied to the task of identifying the key indicators taxa from the 80 families occurring in the data. The data were collated to family level with the biological classification being used as the class attribute. The indicator values based on the mutual information and the RMS-D approach were evaluated using a spreadsheet, while the stepwise discriminant analysis was completed using the SAS procedure STEPDISC [153].

### 6.4.2 Results

Table 6.6 shows the ranking of the families according to each of the assessment procedures used. Column 1 gives the family, while Columns 2 and 3 give the rank and score for each family for the RMS-D approach. Likewise Columns 4 and 5 show the rank and scores for the mutual information. Column 6 shows those families which were selected by the stepwise discriminant procedure, while Columns 7 and 8 give the rank and number of occurrences of each family within the 292 samples. The last column shows the stage at which the family was incorporated into the BERT list of key taxa.[4] The double entries in the Heptageniidae and Hydropsychidae columns denote that taxa from within these families were included at different stages in the BERT project.

[4]This is a little simplified because of the mixed taxonomy of the BERT taxa, as certain families can be represented more than once within the BERT taxa. In addition, some of the BERT taxa were species which were considered to be much better indicators than their families (e.g. *Hydropsyche angustipennis* instead of Hydropsychidae).

Table 6.6: Ranking of taxa in terms of three indicator indices: RMS-D scores, mutual information and stepwise regression.

| Taxon | RMS-D $\mathcal{IV}_i^w$ | | Mutual Information | | Step wise | Number of Occurrences | | BERT Taxa |
|---|---|---|---|---|---|---|---|---|
| | Rank | Score | Rank | Score | | Rank | Total | |
| Gammaridae | 1 | 0.381 | 2 | 0.180 | x | 4 | 185 | 1 |
| Baetidae | 2 | 0.363 | 1 | 0.181 | x | 9 | 153 | 2 |
| Tipulidae | 3 | 0.344 | 3 | 0.164 | x | 10 | 150 | |
| Heptageniidae | 4 | 0.318 | 4 | 0.163 | x | 20 | 75 | 2/3 |
| Asellidae | 5 | 0.296 | 9 | 0.120 | x | 5 | 179 | 1 |
| Ancylidae | 6 | 0.296 | 8 | 0.123 | | 12 | 134 | 3 |
| Rhyacophilidae | 7 | 0.286 | 5 | 0.138 | | 21 | 69 | 1 |
| Perlodidae | 8 | 0.282 | 6 | 0.134 | x | 22 | 57 | 3 |
| Elminthidae | 9 | 0.271 | 10 | 0.116 | | 17 | 89 | 3 |
| Limnephilidae | 10 | 0.266 | 11 | 0.112 | | 18 | 87 | 3 |
| Hydropsychidae | 11 | 0.263 | 12 | 0.109 | x | 16 | 112 | 1/2 |
| Leuctridae | 12 | 0.262 | 7 | 0.126 | x | 29 | 47 | 1 |
| Sphaeriidae | 13 | 0.261 | 14 | 0.092 | x | 7 | 161 | 3 |
| Simuliidae | 14 | 0.252 | 13 | 0.093 | | 14 | 119 | 1 |
| Glossiphoniidae | 15 | 0.239 | 18 | 0.077 | | 11 | 149 | 2 |
| Hydrobiidae | 16 | 0.234 | 16 | 0.082 | x | 13 | 132 | 3 |
| Dytiscidae | 17 | 0.220 | 17 | 0.081 | x | 15 | 115 | 3 |
| Erpobdellidae | 18 | 0.216 | 20 | 0.060 | x | 8 | 155 | 1 |
| Tubificidae | 19 | 0.200 | 24 | 0.053 | x | 2 | 273 | 1 |
| Nemouridae | 20 | 0.194 | 15 | 0.086 | x | 33 | 39 | 3 |
| Hydracarina | 21 | 0.169 | 21 | 0.057 | | 19 | 85 | 3 |
| Leptophlebiidae | 22 | 0.157 | 19 | 0.062 | x | 31 | 43 | |
| Lymnaeidae | 23 | 0.157 | 35 | 0.035 | | 6 | 163 | 1 |
| Ephemeridae | 24 | 0.150 | 22 | 0.055 | | 25= | 50 | |
| Caenidae | 25 | 0.146 | 23 | 0.054 | | 25= | 50 | 2 |
| Oligochaeta | 26 | 0.146 | 33 | 0.036 | | 1 | 280 | |
| Chironomidae | 27 | 0.145 | 37 | 0.030 | x | 3 | 241 | 1 |
| Leptoceridae | 28 | 0.140 | 26 | 0.051 | | 23 | 56 | |
| Sericostomatidae | 29 | 0.133 | 25 | 0.053 | | 35= | 31 | |
| Planorbidae | 30 | 0.124 | 27 | 0.046 | | 24 | 54 | 3 |

Table 6.6 continued overleaf

Table 6.6: Ranking of taxa in terms of three indicator indices (cont'd).

| Taxon | RMS-D $\mathcal{IV}_i^w$ | | Mutual Information | | Step wise | Number of Occurrences | | BERT Taxa |
|---|---|---|---|---|---|---|---|---|
| | Rank | Score | Rank | Score | | Rank | Total | |
| Hydrophilidae | 31 | 0.107 | 34 | 0.036 | | 28 | 49 | |
| Rhagionidae | 32 | 0.107 | 29 | 0.041 | | 37= | 30 | |
| Physidae | 33 | 0.106 | 30 | 0.038 | x | 32 | 42 | |
| Chloroperlidae | 34 | 0.104 | 28 | 0.043 | | 46 | 18 | |
| Ephemerellidae | 35 | 0.102 | 32 | 0.036 | x | 35= | 31 | 3 |
| Haliplidae | 36 | 0.094 | 38 | 0.029 | | 25= | 50 | 2 |
| Planariidae | 37 | 0.093 | 36 | 0.032 | | 30 | 44 | 3 |
| Taeniopterygidae | 38 | 0.091 | 31 | 0.037 | | 47= | 17 | |
| Sialidae | 39 | 0.083 | 40 | 0.028 | | 34 | 32 | 2 |
| Polycentropidae | 40 | 0.082 | 39 | 0.029 | | 39= | 26 | 3 |
| Lumbriculidae | 41 | 0.066 | 43 | 0.024 | | 39= | 26 | 2 |
| Goeridae | 42 | 0.065 | 41 | 0.026 | | 44= | 20 | |
| Lumbricidae | 43 | 0.064 | 45 | 0.021 | | 37= | 30 | |
| Perlidae | 44 | 0.062 | 42 | 0.025 | x | 54= | 10 | |
| Valvatidae | 45 | 0.059 | 44 | 0.023 | | 43 | 23 | |
| Corixidae | 46 | 0.057 | 47 | 0.020 | | 42 | 25 | |
| Lepidostomatidae | 47 | 0.056 | 46 | 0.020 | | 47= | 17 | |
| Muscidae | 48 | 0.045 | 50 | 0.013 | | 39= | 26 | |
| Piscicolidae | 49 | 0.043 | 49 | 0.016 | | 44= | 20 | |
| Odontoceridae | 50 | 0.043 | 48 | 0.017 | | 57= | 7 | |
| Dendrocoelidae | 51 | 0.035 | 52 | 0.012 | | 49 | 16 | 3 |
| Psychomyiidae | 52 | 0.034 | 51 | 0.012 | | 50 | 15 | |
| Coenagriidae | 53 | 0.031 | 53 | 0.012 | | 51 | 12 | |
| Calopterygidae | 54 | 0.029 | 54 | 0.011 | | 52= | 11 | |
| Astacidae | 55 | 0.029 | 55 | 0.011 | | 54= | 10 | |
| Gyrinidae | 56 | 0.026 | 56 | 0.009 | x | 52= | 11 | |
| Scirtidae | 57 | 0.024 | 58 | 0.008 | | 57= | 7 | |
| Dixidae | 58 | 0.021 | 57 | 0.008 | | 62= | 5 | |
| Veliidae | 59 | 0.019 | 63 | 0.006 | | 57= | 7 | |
| Ceratopogonidae | 60 | 0.017 | 67 | 0.005 | | 56 | 9 | |

### 6.4.3 Discussion

From Table 6.6 it is apparent that there is a good correlation between ranking of the indicators selected by the RMS-D method and the mutual information. The rank correlation coefficient between the RMS-D and mutual information ranks (columns 2 and 4 from Table 6.6) is 0.9986 ($P < 0.0001$). It is also apparent from Table 6.6 that the taxa ranked as the best indicators corresponded closely to those identified as key taxa by the Expert during the knowledge elicitation exercise outlined in Section 4.2.3. This is an important result as it seems that the taxa which are considered key indicators by the Expert are also the ones identified from a mathematical analysis. This could prove useful in situations where such expertise is not available, since it implies that key indicators identified by analysis would most probably be identified as such by a field expert, if one was available. It should also be noted that two additional factors were considered by the Expert:

*i.* the ease of identification, which is subjective and therefore difficult to quantify in a numeric analysis, and

*ii.* the distribution of discriminatory power across the five quality classes.

The latter can, in fact, be treated as part of the selection procedure after the indicator value of the individual taxa has been determined.

The stepwise discriminant analysis yielded a set of indicators that differed appreciably from those highlighted by the other two algorithms. For example of the 20 taxa selected two were ranked numbers 44 and 56 according to the RMS-D method. There is no guarantee that the model generated by using the variables selected by the step-wise discriminant analysis would yield the most accurate model, it is likely that another subset could provide a more accurate model. It appears that the stepwise approach tries to select the best 'overall team' to discriminate between the classes (as per the task carried out by the Expert), whereas the other two methods identify good individual 'players' without attempting to form a team. However, the 'team' requirement seems to be particularly sensitive to the data as some of the taxa selected cannot be considered as good indicators.

There is some correlation between the frequency of occurrence of the taxa within the data set and the rank generated from the RMS-D analysis and the

mutual information, however the most frequent taxa in the data set are not necessarily the best indicators of water quality class and this is reflected by the results. For example, the most frequent taxa in the data were the Oligochaeta, which were ranked as number 26 and 33 by the RMS-D scores and mutual information scores, respectively. The Oligochaeta were so ubiquitous that the presence provided little information as to the quality of the river. Some taxa with a relatively low occurrence in the data were highlighted as good indicators, these generally being the good water quality indicators, such as the stone-fly and may-fly groups, for example Perlodidae and Heptageniidae.

Considering the relationship between the ranks from the RMS-D method and the mutual information with the BERT taxa, it can be seen that all of the first ten taxa selected for use in the BERT project (those denoted by '1' in Column 9) occur within the top half of the list. Elsewhere the second ten and the last 21 are more randomly distributed. It is important to remember that the Expert used additional criteria, such as the spread across the classes and the ease of identification, when selecting the taxa for the BERT lists. One anomaly is Tipulidae, which is ranked number 3 by both the indices, but this is due to the problem of the sampling strategy used and the assumptions made in the elicitation of the BERT taxa. Another source of error is the difference in taxonomic levels between the family groupings (due to the inconsistencies in identification in the Severn-Trent data and the mixed levels of the BERT taxa). This was unavoidable with the available data.

There are two drawbacks to using these new methods. Firstly the spread of the classifications covered by the selected taxa is not taken into consideration. This means that it would be possible for the best 10 indicators selected to only discriminate between two classes, say B1a and B1b, while providing no discriminatory power for the classes B2, B3 and B4. It would be possible to generate *ad hoc* rules to ensure complete coverage of the target classes, but this would make the implementation untidy and inefficient. The easiest method of overcoming this would be to examine a model's confusion matrices and highlight any class which is poorly classified. Then examine the indicator scores and the frequencies of the taxa, and select the taxa which is not presently used and has the highest ranking score. Secondly, the use of probabilities based on the 'binning' of the data can cause problems with the density of

the occurrences within the bins. For example, in the above derivation with 5 classes and 4 states the total number of bins is 20. However, if the number of classes or states were to increase the number of bins would increase also, and this may lead to problems with the adequate coverage of all the bins, and any sampling error would increase. Thus, if the number of classes and/or states were large with respect to the number of samples then the reliability of the methods would be questionable. Also, if the underlying conditional independence assumption is removed, then the situation is made much worse, with the number of bins being given by the number of classes to the power of the number of taxa.

At present the effect of correlations between the variables have not been considered. It may be advantageous to remove variables that have a high correlation with one another, but it is important to take into account the classifications as well. It may be that two species have a high negative correlation e.g. *Gammarus pulex* and *Asellus aquaticus* due to competition [54, 176], for example, but are indicative of different classes by their presence.

In most studies one group of animals are considered in isolation (e.g. the macroinvertebrates or diatoms or algae) and the taxonomy is taken to equivalent levels across the whole range of the taxa. It is generally the case that taxa are identified to species level or family level so the selection of indicator taxa may be limited in scope. Ideally suitable flora or fauna should be selected before a study is started, and should comprehensively cover all the available indicators. If a single group has been selected then the taxonomic soundness of the group becomes important. It has been suggested that analysis to family level is sufficient for most studies (this applies more to community structure prediction) and that nothing extra is to be gained by taking the identification to species level. This approach may be suitable for community structure work, but it can result in substantial information loss.

It was found in this study that a few individual species are considered by the Expert to be unrepresentative of the rest of the family (or genus), and that knowledge of presence or absence of these species does provide a substantial gain in information over that provided by group data only. The knowledge elicitation exercise highlighted three clear examples of species having significantly different characteristics from the other members of its family, these were

*Hydropsyche angustipennis*, *Chironomus riparius* and *Tubifex tubifex*. In the BERT system the indicator taxa used included *Hydropsychidae angustipennis* and 'Other Hydropsychidae' since the Expert considered that the tolerance of *H. angustipennis* was higher than that of the other species of Hydropsychidae. The Expert emphasised that classification would be hindered if the distinction was not made between between Hydropsychidae and *H. angustipennis*. A similar argument also applies to *Chironomus riparius* (or more generally red chironomids) and Chironomidae, *C. riparius* being more tolerant of organic enrichment than the other members of its family. Some groups are considered difficult to identify to species level (i.e. the Nematodes, Porifera or Hydracarina) and provide little extra information anyway, so it is not desirable to take the identification to a higher level. *Tubifex tubifex* was another example, but none of the data identified it and we were forced to use Tubificidae on practical grounds.

### 6.4.4 Absent/Present and Information Loss

Using the indices of the preceding section a comparison was made between the data for two scenarios, firstly when the taxonomic information was available in four abundance levels (*absent, present, few* and *com+*), and secondly with the taxa being recorded as either absent or present. This was carried out to identify any quantifiable gain in the value of the indices (i.e. a gain in the total information that is available) when going from absent/present to four levels of abundance. Table 6.7 records the results, giving $\mathcal{IV}_i^w$, mutual information and rank for both the 4-level abundance (taken from Table 6.6) and the absent/present data. The final three columns record the percentage gain in the value of the indices moving from absent/present to the 4-level abundance, and the change in mean rank position.

Considering the present/absent data it can again be observed that there is a good correlation between the $\mathcal{IV}_i^w$ and mutual information indices, and generally the best absent/present indicators are the same as those identified when considered in the four levels of abundance. Comparing the weighted indices of Table 6.7 with those of Table 6.6 it can be noted that there is a reduction in all of the indicator values for all of the taxa, but the magnitude

Table 6.7: Comparison of RMS-D scores and mutual information indices for Absent/Present and 4-Level Abundance of the Severn-Trent NRA 292 family data.

| Taxon | 4-Level Abundance | | | Absent/Present | | | Gain | Gain | +/− |
|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{IV}_i^w$ | M.I. | Rank | $\mathcal{IV}_i^w$ | M.I. | Rank | $\mathcal{IV}_i^w$ | M.I. | Rank |
| Gammaridae | 0.381 | 0.180 | 1.5 | 0.336 | 0.151 | 2.5 | 13.2% | 18.9% | 1.0 |
| Baetidae | 0.363 | 0.181 | 1.5 | 0.338 | 0.159 | 1 | 7.3% | 13.3% | -0.5 |
| Tipulidae | 0.344 | 0.164 | 3 | 0.330 | 0.151 | 3.5 | 4.3% | 8.9% | 0.5 |
| Heptageniidae | 0.318 | 0.163 | 4 | 0.313 | 0.158 | 3 | 1.8% | 3.6% | -1.0 |
| Asellidae | 0.296 | 0.120 | 7 | 0.269 | 0.091 | 10 | 10.2% | 31.6% | 3.0 |
| Ancylidae | 0.296 | 0.123 | 7 | 0.279 | 0.112 | 8 | 6.1% | 10.0% | 1.0 |
| Rhyacophilidae | 0.286 | 0.138 | 6 | 0.283 | 0.135 | 5 | 1.1% | 1.8% | -1.0 |
| Perlodidae | 0.282 | 0.134 | 7 | 0.282 | 0.134 | 6 | 0.0% | 0.1% | -1.0 |
| Elminthidae | 0.271 | 0.116 | 9.5 | 0.268 | 0.115 | 8.5 | 1.1% | 1.3% | -1.0 |
| Limnephilidae | 0.266 | 0.112 | 10.5 | 0.260 | 0.109 | 11 | 2.1% | 3.2% | 0.5 |
| Hydropsychidae | 0.263 | 0.109 | 11.5 | 0.261 | 0.107 | 11 | 1.1% | 1.5% | -0.5 |
| Leuctridae | 0.262 | 0.126 | 9.5 | 0.261 | 0.122 | 8.5 | 0.3% | 3.5% | -1.0 |
| Sphaeriidae | 0.261 | 0.092 | 13.5 | 0.245 | 0.076 | 14.5 | 6.3% | 22.1% | 1.0 |
| Simuliidae | 0.252 | 0.093 | 13.5 | 0.232 | 0.080 | 14 | 8.6% | 16.1% | 0.5 |
| Glossiphoniidae | 0.239 | 0.077 | 16.5 | 0.227 | 0.063 | 16.5 | 5.1% | 22.8% | 0.0 |
| Hydrobiidae | 0.234 | 0.082 | 16 | 0.227 | 0.071 | 16.5 | 3.1% | 14.7% | 0.5 |
| Dytiscidae | 0.220 | 0.081 | 17 | 0.218 | 0.078 | 16 | 1.2% | 4.1% | -1.0 |
| Erpobdellidae | 0.216 | 0.060 | 19 | 0.210 | 0.051 | 20 | 2.9% | 16.0% | 1.0 |
| Tubificidae | 0.200 | 0.053 | 21.5 | 0.048 | 0.013 | 46 | 316.7% | 310.6% | 24.5 |
| Nemouridae | 0.194 | 0.086 | 17.5 | 0.192 | 0.080 | 16 | 1.2% | 6.9% | -1.5 |
| Hydracarina | 0.169 | 0.057 | 21 | 0.164 | 0.050 | 21.5 | 3.0% | 12.3% | 0.5 |
| Leptophlebiidae | 0.157 | 0.062 | 20.5 | 0.157 | 0.060 | 20 | 0.2% | 3.4% | -0.5 |
| Lymnaeidae | 0.157 | 0.035 | 29 | 0.144 | 0.023 | 30 | 8.6% | 50.3% | 1.0 |
| Ephemeridae | 0.150 | 0.055 | 23 | 0.149 | 0.055 | 21 | 0.8% | 1.6% | -2.0 |
| Caenidae | 0.146 | 0.054 | 24 | 0.137 | 0.047 | 24 | 6.9% | 14.2% | 0.0 |
| Oligochaeta | 0.146 | 0.036 | 29.5 | 0.046 | 0.016 | 46 | 216.8% | 127.8% | 16.5 |
| Chironomidae | 0.145 | 0.030 | 32 | 0.083 | 0.012 | 42.5 | 74.5% | 159.0% | 10.5 |
| Leptoceridae | 0.140 | 0.051 | 27 | 0.135 | 0.047 | 25 | 3.9% | 9.6% | -2.0 |
| Sericostomatidae | 0.133 | 0.053 | 27 | 0.132 | 0.052 | 23.5 | 0.9% | 1.9% | -3.5 |
| Planorbidae | 0.124 | 0.046 | 28.5 | 0.118 | 0.037 | 27 | 5.7% | 22.3% | -1.5 |

Table 6.7 continued overleaf

Table 6.7: Comparison of RMS-D scores and mutual information (cont'd).

| Taxon | 4-Level Abundance | | | Absent/Present | | | Gain | Gain | +/− |
|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{IV}_i^w$ | M.I. | Rank | $\mathcal{IV}_i^w$ | M.I. | Rank | $\mathcal{IV}_i^w$ | M.I. | Rank |
| Hydrophilidae | 0.107 | 0.036 | 32.5 | 0.096 | 0.025 | 34 | 11.7% | 43.8% | 1.5 |
| Rhagionidae | 0.107 | 0.041 | 30.5 | 0.101 | 0.035 | 29 | 5.8% | 18.2% | -1.5 |
| Physidae | 0.106 | 0.038 | 31.5 | 0.104 | 0.034 | 29.5 | 2.2% | 10.5% | -2.0 |
| Chloroperlidae | 0.104 | 0.043 | 31 | 0.104 | 0.041 | 27 | 0.1% | 3.5% | -4.0 |
| Ephemerellidae | 0.102 | 0.036 | 33.5 | 0.100 | 0.034 | 31 | 2.0% | 7.8% | -2.5 |
| Haliplidae | 0.094 | 0.029 | 37 | 0.091 | 0.026 | 33.5 | 3.2% | 15.5% | -3.5 |
| Planariidae | 0.093 | 0.032 | 36.5 | 0.085 | 0.023 | 36.5 | 9.4% | 41.4% | 0.0 |
| Taeniopterygidae | 0.091 | 0.037 | 34.5 | 0.091 | 0.035 | 31.5 | 0.9% | 5.5% | -3.0 |
| Sialidae | 0.083 | 0.028 | 39.5 | 0.080 | 0.026 | 35 | 3.9% | 9.2% | -4.5 |
| Polycentropidae | 0.082 | 0.029 | 39.5 | 0.080 | 0.027 | 35 | 2.5% | 8.1% | -4.5 |
| Lumbriculidae | 0.066 | 0.024 | 42 | 0.055 | 0.013 | 45.5 | 20.0% | 85.3% | 3.5 |
| Goeridae | 0.065 | 0.026 | 41.5 | 0.063 | 0.022 | 39 | 3.8% | 16.9% | -2.5 |
| Lumbricidae | 0.064 | 0.021 | 44 | 0.062 | 0.020 | 41 | 3.2% | 7.1% | -3.0 |
| Perlidae | 0.062 | 0.025 | 43 | 0.062 | 0.025 | 37.5 | 0.00% | 0.00% | -5.5 |
| Valvatidae | 0.059 | 0.023 | 44.5 | 0.058 | 0.022 | 41 | 2.8% | 3.8% | -3.5 |
| Corixidae | 0.057 | 0.020 | 46.5 | 0.057 | 0.018 | 43 | 1.1% | 9.0% | -3.5 |
| Lepidostomatidae | 0.056 | 0.020 | 46.5 | 0.054 | 0.019 | 43.5 | 3.0% | 8.9% | -3.0 |
| Muscidae | 0.045 | 0.013 | 49 | 0.043 | 0.010 | 50.5 | 4.4% | 26.4% | 1.5 |
| Piscicolidae | 0.043 | 0.016 | 49 | 0.038 | 0.012 | 49 | 13.5% | 35.4% | 0.0 |
| Odontoceridae | 0.043 | 0.017 | 49 | 0.043 | 0.017 | 46.5 | 0.00% | 0.00% | -2.5 |
| Dendrocoelidae | 0.035 | 0.012 | 51.5 | 0.035 | 0.011 | 51 | 2.1% | 13.3% | -0.5 |
| Psychomyiidae | 0.034 | 0.012 | 51.5 | 0.030 | 0.009 | 53 | 11.8% | 33.3% | 1.5 |
| Coenagriidae | 0.031 | 0.012 | 53 | 0.030 | 0.011 | 51.5 | 3.5% | 9.2% | -1.5 |
| Calopterygidae | 0.029 | 0.011 | 54 | 0.028 | 0.011 | 53 | 3.3% | 6.8% | -1.0 |
| Astacidae | 0.029 | 0.011 | 55 | 0.028 | 0.008 | 55 | 5.2% | 25.4% | 0.0 |
| Gyrinidae | 0.026 | 0.009 | 56 | 0.024 | 0.007 | 56.5 | 8.3% | 28.8% | 0.5 |
| Scirtidae | 0.024 | 0.008 | 57.5 | 0.024 | 0.007 | 56.5 | 0.4% | 6.0% | -1.0 |
| Dixidae | 0.021 | 0.008 | 57.5 | 0.019 | 0.006 | 58 | 10.5% | 47.1% | 0.5 |
| Veliidae | 0.019 | 0.006 | 59 | 0.018 | 0.005 | 59 | 1.5% | 12.5% | 0.0 |
| Ceratopogonidae | 0.017 | 0.005 | 60 | 0.016 | 0.004 | 60 | 4.1% | 28.3% | 0.0 |

of the decrease varies. For example, Gammaridae falls from 0.381 to 0.336 (0.045) while Leuctridae falls from 0.262 to 0.261 (0.001). There is a small reordering of the top 12 taxa while ranks 13 to 20 are unaltered. For the mutual information scores there is a reduction again in value from the 4 state abundance to the absent/present scenario, and this can be directly interpreted as a loss of information.

The most interesting columns in Table 6.7 are those giving the percentage gain in calculating the indices for 4-level abundance when compared to absent/present. In all cases there is information lost when the taxa are recorded as absent/present (except for a few rare taxa which only occurred in two states in the full data set). It is worth noting that there is a very large range of percentage gains ranging from a few percent to over 300% for Tubificidae. Most of the information gains are in the region of 0-10%, but for some of the more common taxa the information gain is very much higher than this. This reveals that for a proportion of the taxa the additional time taken to provide a simple assessment of the order of abundance is worthwhile, while for others simple absent/present status is almost as useful as a more detailed recording. It is possible to categorise and rank the terms of the information gains, and this may be considered in the design of sampling program. For example, consider the following (the numbers represent the mutual information):

$$\underset{0.016}{\text{Oligochaete (a/p)}} \longrightarrow \underset{0.036}{\text{Oligochaete (+abund.)}} \longrightarrow \underset{0.053}{\text{Tubificidae (+abund.)}}$$

$$\underbrace{\qquad\qquad}_{\text{Abundance gain}} \qquad \underbrace{\qquad\qquad\qquad}_{\text{Identification gain}}$$

which clearly demonstrates that there is an increase in information progressing from lower effort processing, recording subclass and absence/presence, to a more reasonable, in this case,[5] family with abundance.

An additional consideration in the analysis is that the numerical levels which differentiate the states of *present*, *few* and *com+* are the same for all taxa. Also due to the scale of the Severn-Trent bandings very few taxa occurred as *com+*, those that did are the ones which have the larger information gains

[5]Considering the appreciable taxonomic difficulties, where, for Oligochaeta, identification to anything below family level is difficult.

in Table 6.7. Ideally, each taxon would have it own individual definition of *present*, *few* and *com+* (this was adopted for the BERT system, see Table 4.1). The abundance level adopted was too coarse for the taxa which occur in small numbers, this implies that the information gains of Table 6.7 are on the low side. If the ideal analysis was completed then the information gains would be more uniform across all taxa, but with the common taxa again benefitting most from the increased level of enumeration.

## 6.5 Model Performance Using Increasing Numbers of Indicators

### 6.5.1 Procedure

This experiment was conducted to investigate the relationship between the number of indicators used in the model and the resulting performance of the model. The 80 families were ranked according to their weighted scores given in Section 6.3.1. For the first experiment the best 5 indicator taxa were selected and a series of MLP's, linear and quadratic discriminant functions were trained using these taxa. The next best 5 were then added to the original set and another series of models were trained, this was continued until the best 50 indicators had been used. Note that for the input scaling *absent* was taken to be 0.0, *present* 0.33, *few* 0.66 and *com+* 1.00. Eight hidden units were used in the hidden layer, and each configuration was run 20 times using different initial weights, while the discriminant networks were processed only once.

### 6.5.2 Results

Figure 6.2 shows the training and testing results for the linear, quadratic and MLP networks. The MLP results are the average of 20 runs. From Figure 6.2 it is apparent that the difference in the performance between the linear, quadratic and MLP networks on the test data was not significant, although the three models all exhibited slightly different trends. The quadratic classifier performed at about the same level (65-70%) for all the numbers of input taxa used, but it was the worst performer of the three models. The classification

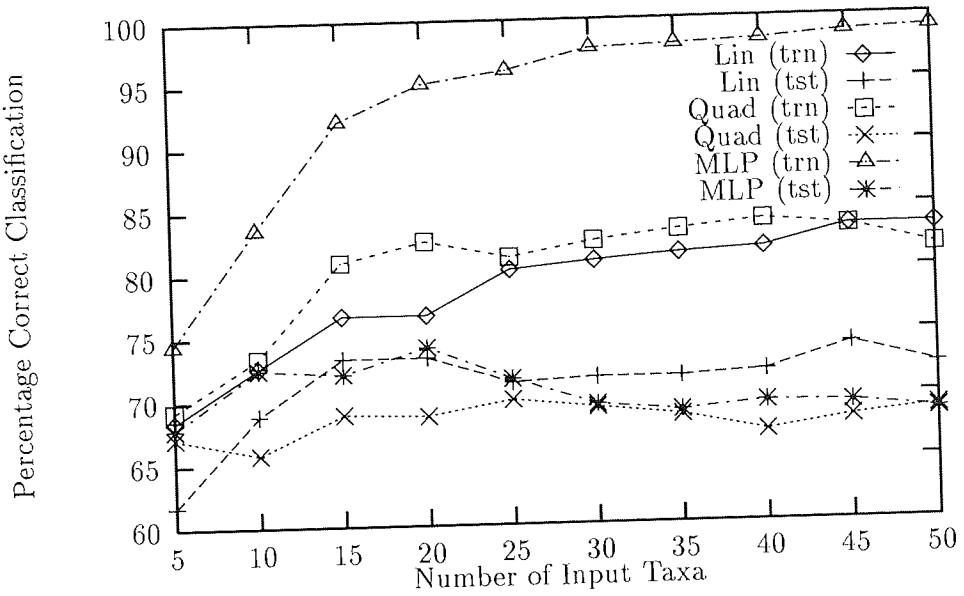



Figure 6.2: Performance of Linear, Quadratic and MLP models using different totals of indicator taxa, selected using the RMS-D method of Section 6.3.1. All models trained using four-fold cross validation and plain input coding.

rate of the linear classifier improved from 5 to 15 inputs, and from this point onward the classification rate degraded fractionally and then remained fairly constant. On the other hand, the MLP networks performance improved from 10 to 20 inputs and then started to slowly degrade. This was an expected result since it merely reflects the effects of overfitting.

The performance on the training set yielded a trend that was repeated in all three methods. With the fewest indicators the performance was at the lowest level, then with 10 and 15 indicators there was a notable improvement in the classification rate, the rate of increase tailed off with further additions to the set. The more indicators that were used the greater the discrimination between the classes, and hence the higher classification rates. The MLP had the best performance on the training set over all numbers of inputs.

### 6.5.3 Discussion

This experiment has shown that by careful selection the number of taxa required can be substantially reduced whilst at the same time improving the

predictive performance of the model. A reasonable level of classification was reached by using just five indicators (Gammaridae, Baetidae, Tipulidae, Heptageniidae and Asellidae), and a simple classification system could be implemented using just these five taxa. But this result may be highly dependent on the 292 sample data and should not be inferred as generally applicable for all rivers and regions.

By increasing the number of taxa the training performances of the three models all improved, but this was not accompanied by a corresponding increase in the performance on the test set. In fact the performance on this set exhibited the hallmarks of an increased model complexity (i.e. increasing dimension of input space) resulting in a fall in predictive performance. This is apparent from a comparison of the training and testing results. The most capable model, the MLP, had by far the best training performance but on the test data its performance was only marginally better than that of the linear network, which had the poorest performance on the training data. This reflects its greater freedom to fit the data, in fact over-fit the data, as is indicated by its performance on the test set.

## 6.6 Input Encoding Using RMS-D Values

### 6.6.1 Procedure

All the models considered in this dissertation, see Section 5.3.4, use the same scale of input encoding for all the taxa which are used as predictor variables. Thus, even though it is known that one taxon may be more significant than another, the same input weight would be used for both. Ideally, a method that uses a different scaling for different taxa would be able to encode prior information into the network. A possible mechanism for achieving was the indicator values, $\mathcal{IV}_{ik}$, from Section 6.3.1. The effects of the scaling is shown in Figure 6.3. At the top of the figure both Asellidae and Ephemerellidae have common codings presented to the network. Thus, the absence of each is given equal weighting. This may not be realistic since Ephemerellidae occurs far less frequently than Asellidae and provides less information with regard to classification. When the $\mathcal{IV}_{ik}$ are used as inputs, the weighting changes

between the taxa, as does the scaling between their discrete states of presence. In the bottom of Figure 6.3 the absence of Asellidae has an absolute value of 0.315, while that of Ephemerellidae is much smaller at 0.055. Thus during the back-propagation, the change of the weights will be much stronger for the absence of Asellidae than Ephemerellidae, thus the networks will learn that the absence of Asellidae has a greater impact on the classification than that of Ephemerellidae.

The experiment described in the previous section was repeated, but using the $\mathcal{IV}_{ik}$ indicator values as the input scaling rather than the uniform scale 0.0, 0.33, 0.66 and 1.0. The networks were trained for 10 starts using four-fold cross validation.

### 6.6.2 Results

The classification rates for the three models are given in Table 6.8. The linear network did not perform as well as before, its highest classification rate being 69.5% compared to 74.0% when using uniform scaling. However, the results from the quadratic model show a marked improvement from a maximum of 69.9% to 74.3%. The MLP produces a similar range of values except for the 15 indicators where the classification rate reached its highest value of 77.2%. Of the 10 networks trained, the best individual one achieved a classification rate of 81.2%, which was the highest classification rate achieved by any of the methods, using an equivalent training scheme. The confusion matrix for this network is given in Table 6.9. The error rates, Table 6.10, are highest for the B1b and B3 classes, with a large number of B1b classes being classified as B2's.

### 6.6.3 Discussion

The use of input scalings based on the information value of the inputs appears to have improved the classification rate achieved by the models. For the more complex (capable) models the classification rate increased, while that of the linear model decreased. The use of this revised input scaling provides a means of implicitly encoding *a priori* information into the network.

An unexpected feature of the method is that some of the weightings for the *present*, *few* and *com+* do not follow the same order. For example, it may

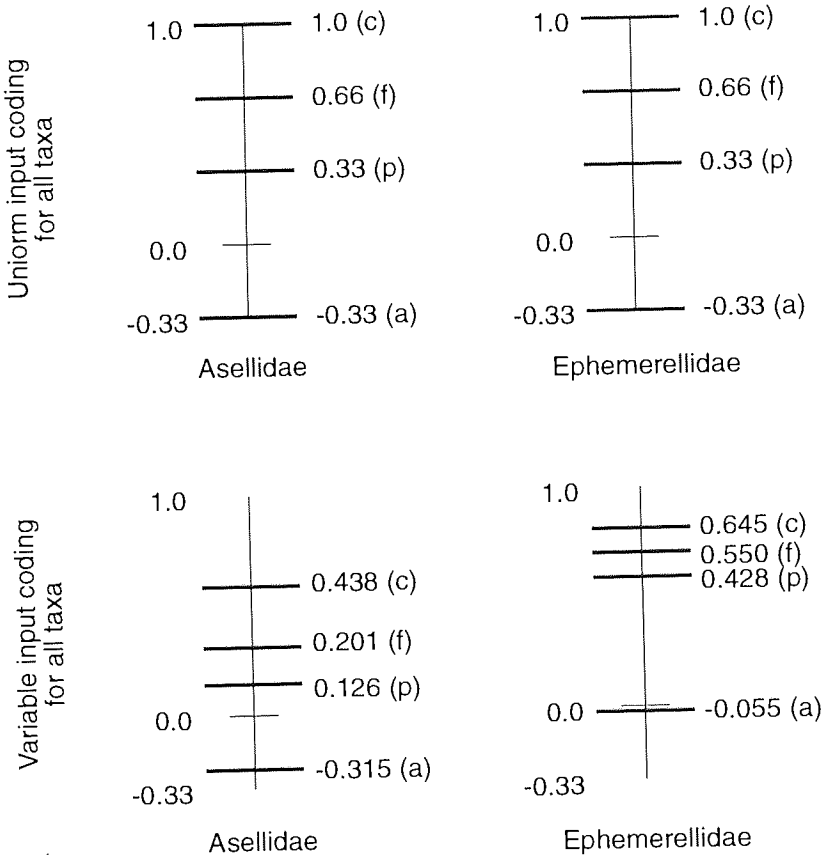



Figure 6.3: The effect of scaling the input encoding using the $\mathcal{IV}_{ik}$ indicator values. The top two scales represent the uniform method of using the same values for all the taxa, while the lower two use the intermediate probabilities. The letters in brackets denote the state of the taxon: a-absent; p-present; f-few; c-com+.

| Best $n$ Indicators | Linear Test | Quadratic Test | MLP Test |
|---|---|---|---|
| 5 | 67.5% | 66.4% | 68.1% |
| 10 | 67.8% | 68.8% | 71.6% |
| 15 | 69.2% | 74.3% | 77.2% |
| 20 | 69.5% | 72.6% | 73.9% |
| 25 | 68.8% | 74.3% | 74.6% |
| 30 | 68.5% | 72.3% | 70.6% |
| 35 | 66.4% | 72.9% | 69.3% |
| 40 | 67.1% | 71.9% | 69.4% |
| 45 | 68.5% | 71.6% | 70.7% |
| 50 | 67.8% | 71.2% | 70.6% |

Table 6.8: Classification rates of linear, quadratic and MLP models using different totals of indicator taxa, identified from the RMS-D scores of Section 6.3.1. All models trained using four-fold cross validation and $\mathcal{IV}_{ik}$ values as input.

| | | Network Output | | | | |
|---|---|---|---|---|---|---|
| | | B1a | B1b | B2 | B3 | B4 |
| Target Output | B1a | 55 | 3 | 0 | 0 | 0 |
| | B1b | 5 | 50 | 15 | 0 | 0 |
| | B2 | 0 | 8 | 89 | 6 | 0 |
| | B3 | 0 | 0 | 5 | 22 | 8 |
| | B4 | 0 | 0 | 0 | 3 | 22 |

Classification rate = 80.2%

Table 6.9: Confusion matrix for the network with the lowest error rate trained using the 'best' 15 indicators and probability values as input. Note the columns corresponding to the classes output by the network, while the rows correspond to the desired classes.

| Error per class | | | | |
|---|---|---|---|---|
| B1a(58) | B1b(71) | B2(103) | B3(35) | B4(25) |
| 5.2% | 28.2% | 13.6% | 37.1% | 12.0% |

Table 6.10: Error per class for the network from Table 6.9. The overall error rate was 18.8%, numbers in brackets indicate number of samples in that class.

be possible that *few* has a higher absolute value than *com+*. The ordering of states may differ from taxon to taxon, see for example Ancylidae (Fig. 6.4), and was probably due to sampling error. This did not appear to effect the performance of the model, and as such it was not explicitly coded in that *present* < *few* < *com+*. Also this method would not be suitable for use with continuous variables, unless the variable was idealised to discrete variables on a fairly coarse scale. Another difficulty associated with the use of the indicator indices is the 'binning' of the data. If the classification was taken to thirteen classes (e.g. see Section 4.3) then the estimated probabilities would become very uneven and unrepresentative of the true distribution.

Figure 6.4 shows the histograms of the values for *absent* through to *com+* for the best 10 indicators. The graphs show a clear ordering for Gammaridae, Asellidae, Heptageniidae and Limnephilidae, while the other six have similar values for *present*, *few* and *com+*. The histogram for Ancylidae shows the mis-ordering of the inputs that can occur. The similarity between the values of the histograms for *present* through to *com+* reveals a binary split into absent/present, which was unexpected. Where there is a difference in the histograms for *present*, *few* and *com+* there was an appreciable drop in the indicator indices for these taxa when considered as absent/present (see Section 6.4 and Table 6.7). There is also a fall off in the magnitude of the *absent* bar, which implies that absence of the lower ranked indicators plays a smaller part in the classification than that of the top indicators.

To summarise, this is a novel method of applying a variable scaling across the networks input. It is particularly useful with small data sets because it helps to extract more utility out of the data, easing the model order selection problems and helping to reduce the problem of network overfitting.

## 6.7 Summary

This chapter has described three methods of selecting indicator variables for use in computer models. Of the three, the RMS-D probability method and the mutual information approaches produced similar results with a good correlation between them. Also, the taxa selected by these two methods corresponded well with those elicited from the Expert.

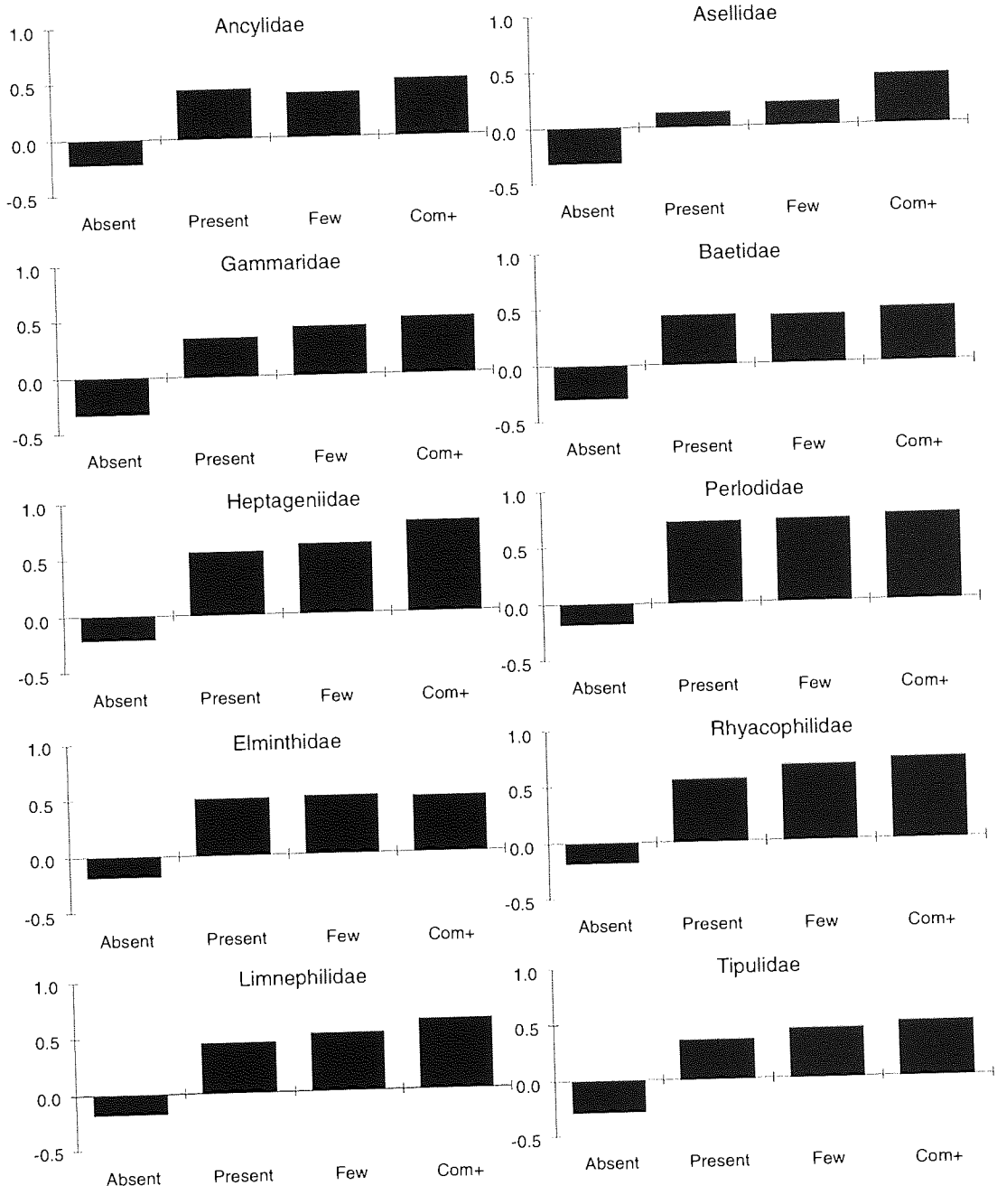



Figure 6.4: Input values, $\mathcal{IV}_{ik}$, for the best 10 indicator taxa from Table 6.6.

The change in information was quantified for different levels of taxa identification and enumeration. As would be expected the identification to species and four levels of abundance had a higher information content than that for family level and absent/present data. The change in information was only significant for some taxa, and it was suggested that this could highlight strategies for optimising the trade-off between the time taken to process a sample and the amount of useful information produced.

By ranking the taxa according to indicator value the number of indicators required for classification was investigated. It was found that the number of indicators used in a model does have an effect on the model's performance, but performance does not continuously increase with the number of indicators used. Indeed peak performance was achieved using the top 15 indicator taxa only, any further indicators appeared only to be increasing the 'noise' in the system. The use of a variable input encoding, using the RMS-D scores, was also demonstrated to be of benefit.

# Chapter 7

# Great Lakes

## 7.1 Introduction

This chapter aims to demonstrate the utility of neural networks in another area of freshwater biomonitoring. Using data from the Laurentian Great Lakes, the classification and prediction of the benthic community structure and toxicity groups, derived from an ordination analysis, are investigated using the environmental variables as predictors. The chapter compares the use of neural networks to the more commonly used discriminant analyses, and investigates alternative strategies for classifying community structure type. This work was completed during a three month visit to the National Water Research Institute, Burlington, Ontario. It forms a small part of a larger project which is investigating the development of sediment guidelines for the remediation of contaminated sediments in the Laurentian Great Lakes [137, 140].

The chapter first describes, briefly, the fundamental aspects and the basic design of the project sampling programme. Section 7.3 reviews the reference site data, and includes the results of ordination on the three available data matrices. The next two sections, Sections 7.4 and 7.5, report on the classification of community and bioassay group type from the environmental variables, which is the most fundamental part of the project. The following two sections discuss alternative ideas that center upon predictive capabilities of the neural network models, with the final section summarising the findings of the chapter.

## 7.2 The Development of Sediment Guidelines

### 7.2.1 Description of Study

The contamination of sediments occurs in freshwater and marine systems throughout the world. Chemical methods of classifying the level of contamination in sediments cannot entirely take account of the biological stress caused by the contaminants, or whether this stress will continue after the primary sources of pollution have been controlled. Concern for the degree of environmental protection afforded by chemical guidelines, together with a lack of uniform international criteria and a failure to introduce plans to remediate degraded areas in the Laurentian Great Lakes [66, 67], prompted scientists at the National Water Research Institute to investigate more comprehensive methods of sediment assessment and evaluation criteria. In 1990, they proposed a study into the development of biological sediment guidelines using two approaches: sediment toxicity tests and benthic invertebrate community structure [137, 140].

The aim of the study is to develop numeric criteria for the biological assessment of sediment contamination at sites in the Laurentian Great Lakes. The overall objectives are to:

*i.* develop a classification system for unpolluted nearshore sites based on the benthic community structure and selected bioassay endpoints,

*ii.* determine the degree to which the site classification can be predicted from physio-chemical variables,

*iii.* establish the relationship between the community structure and bioassay assessments,

*iv.* develop procedures for the prediction of key elements of the fauna expected at a site from its environmental features not affected by human activity,

*v.* select key species and toxicity tests that show the most robust predictive response for the purpose of developing guidelines,

*vi.* establish the sensitivity of selected guidelines at a range of impacted sites.

The biological objectives are to be based upon benthic invertebrate community structure and the bioassay response of four benthic invertebrates to samples of field sediment. The fundamental assumption underlying the development of sediment guidelines is that it is possible to predict the community structure and bioassay responses at unpolluted sites from a few physio-chemical variables [181, 74]. Comparison of the observed community structure, or bioassay endpoint, with the predicted values determines whether or not the site specific guidelines have been met. The purpose of the guidelines is to indicate whether remediation of the site is necessary because of sediment contamination. In this chapter contamination refers to the presence of chemical species that are present in the sediment from anthropogenic processes, while toxicity refers to the response of the flora and fauna to the contamination. Thus a low level contamination may be extremely toxic, depending on the specific chemical involved.

The prediction of benthic macroinvertebrate community structure for setting biological objectives was originally explored in a series of papers from the FBA [4, 181, 41] (see Section 2.3.1), and has been detailed by Reynoldson et al. [136] in terms of freshwater lakes. Other work on lentic community structure includes that by Johnson and Weiderholm [74]. The difficulties in using community structure are its inherent spatial and temporal variability, and the changes in composition that occur through natural fluctuations. The approach based on use of benthic community structure is to predict the assemblage of fauna that would be expected at a site if it were unpolluted, and then compare this with the observed community structure. Any degradation of 'actual' below 'predicted' is assumed to be the result of stress, which for lakes implies that sediment contamination is the probable cause. The first step, in this approach, is to establish a database of uncontaminated (clean) reference sites, containing the community structure data and bioassay data.

### 7.2.2 Reference Sites

The selection of reference sites was based upon the eco-regions and eco-districts of the Great Lakes, and a requirement that 'unpolluted' sites be located 10km 'upstream' of known discharges and within 2km of the shoreline, have a depth of less then 30m (except for Lake Michigan) and be known or suspected to have a fine-grained substrate [136]. A total of 250 sites were identified as reference locations, Figure 7.1, and the sampling of these was to take place over 3 years (91–93). Data from 1991 (50 sites) and 1992 (43 sites) are used in this dissertation. All the sites have been sampled once in late summer or early autumn, with 10% being sampled in each of the three field years. Also four sites have been sampled monthly to determine seasonal and annual variation. Sections 7.3.2 to 7.3.4 give details of the data analyses.

### 7.2.3 Field Procedures

The establishment of a reliable reference database requires the adoption of standard sampling methodologies, a full description of those used in this study can be found in Reynoldson et al. [136]. At each site samples were taken for sediment, water and pore-water chemistry. The water chemistry was determined from samples taken at 0.5m above the sediment-water interface. A mini-box core was used to collect the pore-water chemistry samples, and either this or a mini-ponar grab was used to collect samples of sediment. Samples of the benthic invertebrate community structure were taken from a mini-box core, from which five smaller cores (10cm by 5.5cm) were sub-sampled. The sediment from the smaller cores was sieved using $250\mu m$ mesh, and stored. The five replicate samples were sorted and identified to species level where possible. The bioassay tests were performed on sediments from five replicate samples taken using a mini-ponar grab.

### 7.2.4 Data Analysis

The end products of the field and laboratory work were three matrices; the first containing the environmental (physio-chemical) variables, the second the community structure and the third the results of the bioassay endpoints. The

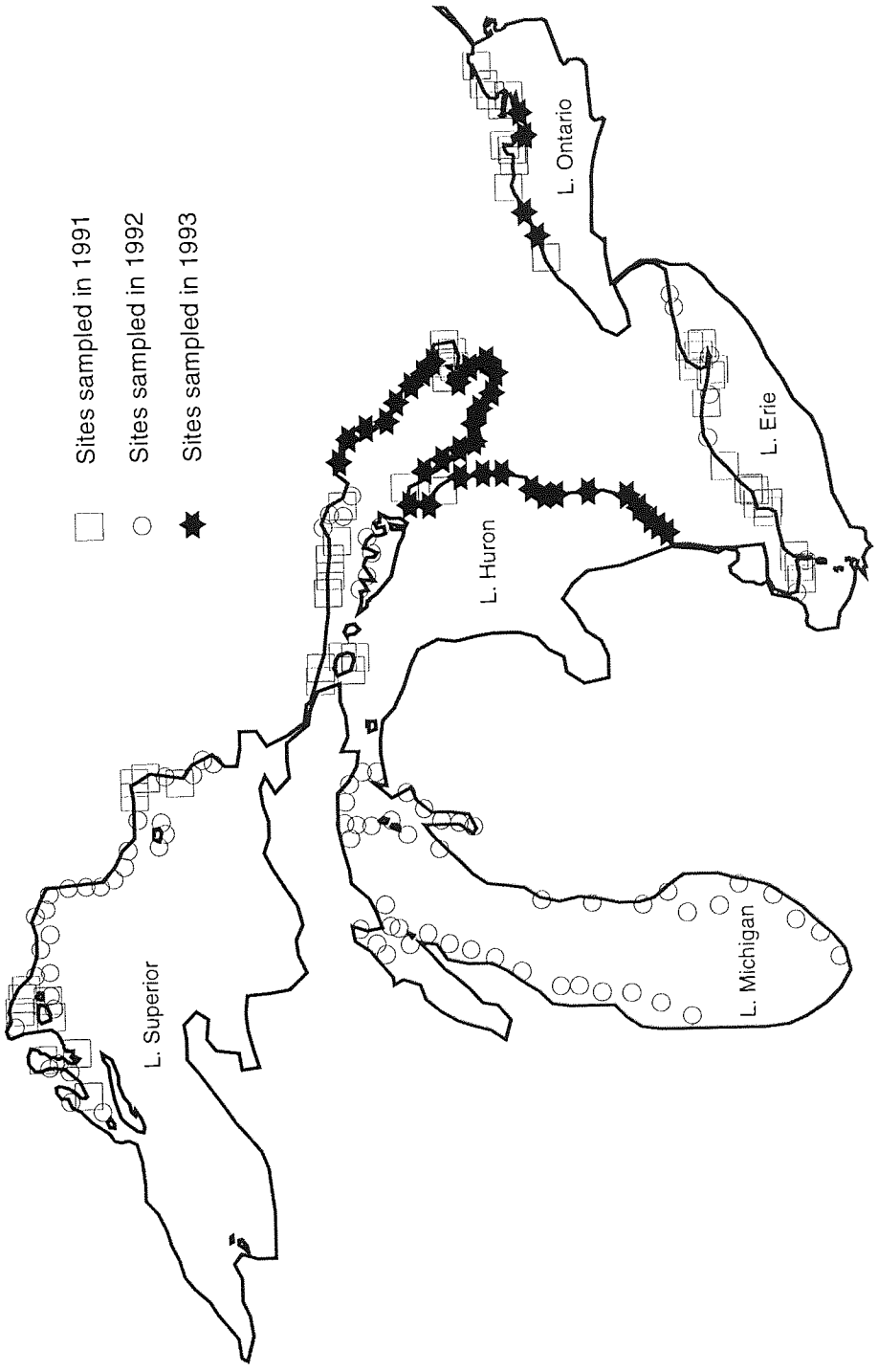



Figure 7.1: Map showing reference site locations in the Great Lakes.

mean results of the community structure and bioassay endpoints were used, as opposed to each individual replicate sample. The standard analysis followed a similar method to that of Wright et al. [181], except that different numerical algorithms were used to cluster the data. The environmental data are used as predictor variables for the classification of the community structure and bioassay data and two classification methods are considered, namely MLPs and multiple discriminant analysis (MDA).

Figure 7.2 summarises the relationship between the data sets from Section 7.3 and the experiments of Sections 7.4 to 7.7. The experimental work can be broken down into three distinct elements:

*i.* the classification of ordination groups from the environmental data,

*ii.* the prediction of abundances of key taxa from environmental variables,

*iii.* the prediction of ordination vectors from environmental variables.

The ability to perform the classification of site groupings from the environmental variables, item *i*, is the fundamental procedure within the project structure. The group predicted from the environmental variables dictates the expected composition of the community structure, or bioassay endpoints. The magnitude of the difference between expected and observed communities will form the basis of the decision making process. Item *ii* provides an alternative mechanism for the determination of the key species, while *iii* provides an alternative technique for site classification.

In Figure 7.2:

- Path $A$ depicts item *ii*. This has previously been described by Ruck et al. [149] using a small data set and is reported again here in Section 7.7 using a larger set.

- Path $B$ represents item *iii*. This had not previously been attempted on this data and the results are reported in Section 7.6.

- Paths $C$ and $D$ cover item $i$ for the community structure; $C$ is the experimental work in Section 7.4, while $D$ is the work reported in Reynoldson et al. [136] using MDA. There is a difference in the discriminant analyses

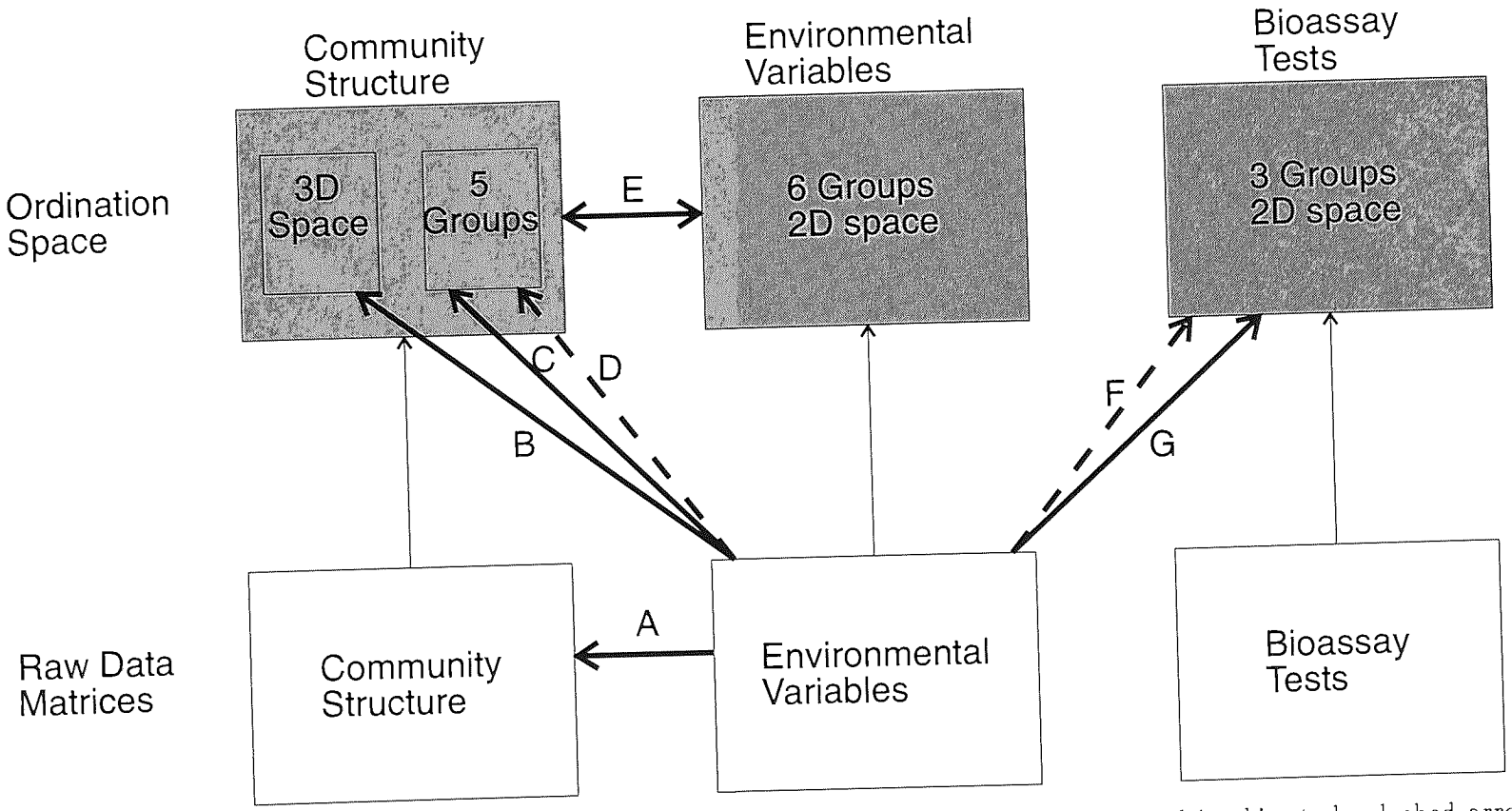



Figure 7.2: Graphical relationship between experiments. Bold arrows are experiments conducted in this study, dashed arrows are reported in Reynoldson et al.[136]. See text for full description of labels.

conducted in this study and that of Reynoldson et al., the performance of all the models in this study are reported in terms of leave-one-out cross validation to allow direct comparison between the neural networks and discriminant analyses, whereas Reynoldson et al. used a less rigorous validation procedure for their discriminant model.

- Path $E$ represents the investigation into the disparity between the structure of the ordination of the environmental data and the ordination of the community structure data. This disparity may account for some of the model's misclassifications, and this is reported in Section 7.4.3.

- Paths $F$ and $G$ depict item $i$ for the toxicity test data; with $F$ being reported by Reynoldson et al. and $G$ being reported in Section 7.5. The above comments concerning the differences of the analyses apply to these experiments as well.

From each of the three ordinations, clusters of data were identified and grouped. Ideally, there should be good agreement between the groups derived from different ordinations. The number of groups and the dimension of the ordination space for each of the three data matrices are summarised in Figure 7.2. In the classification exercises, Sections 7.4 and 7.5, the groups represent five unique classes of site based upon their community structure and three classes of toxicity test results respectively. As a site can only belong to a single class the classification problem can be considered as a *one*-from-$N$ classification.

In the following sections the terms biological group and community structure group refer to the same thing; the groups identified from the ordination of the community structure, likewise bioassay group and toxicity-test group are equivalent terms for the groups from the bioassay ordination. The environmental groups are taken from the ordination of the environmental variables.

### 7.2.5 Numerical Guidelines

These are the decision making criteria that are used to determine if a site is severely impacted (that is the sediments are highly toxic), and whether remediation is necessary. The guidelines are based upon the results of the statistical analyses referred to above, but will also have to consider external factors, such

as cost and the practicality of remediation. The guidelines to be implemented will not be selected until the performance of the statistical systems has been fully researched. At present there are a number of statistical mechanisms being considered [9], and the form of the guidelines will be dependent on the level of performance of the statistical systems, and thus will only be finalised when the full data set is available and has been analysed.

## 7.3 Reference Site Data

### 7.3.1 Ordination

As described in Section 7.2.4 three data sets were available for analysis. These were derived from the field and laboratory work, and comprised of:

*i.* the environmental variables,

*ii.* the benthic invertebrate community structure, and

*iii.* the bioassay test responses.

An initial statistical analysis, followed by ordination, was completed for each data set. The ordination method used in the study is discussed in the following paragraphs and the results of the analyses are given in Sections 7.3.2 to 7.3.4.

Ordination is a multivariate statistical technique used to summarise the underlying patterns of a data set. The site or species data are represented in a low-dimensional (usually 2 or 3) Euclidean space such that separation in the ordination space is based on the dissimilarity of the sites in terms of their composition [36]. It is commonly used to interpret community structure, but it can be applied to any two-dimensional matrix. Generally a two step procedure is adopted, the first step being a site-by-site comparison or a species by species comparison using an association measure, with the second step reducing the dimensionality of the resulting association matrix.

Ordination methods are frequently viewed as 'objective' [63, 69], but different choices of association measure and scaling algorithm can lead to very discordant solutions [69], and thus to different interpretations and conclusions. However, in this study only one technique was used for the association measure

and scaling, these being the Bray-Curtis [171], and a hybrid multi-dimensional scaling algorithm [35] respectively.

The Bray-Curtis dissimilarity index is given by:

$$D_{ij} = \frac{\sum_k |x_{ik} - x_{jk}|}{\sum_k |x_{ik} + x_{jk}|} \tag{7.1}$$

where $x_{ik}$ is the number of species $k$ in sample $i$, $x_{jk}$ is the number of species $k$ in sample $j$ and $D_{ij}$ is the dissimilarity index for samples $i$ and $j$. This index has been widely used in ecological studies and has been shown to produce results that can be intuitively interpreted [37, 35, 69, 136]. The semi-strong hybrid multi-dimensional scaling algorithm was developed for ecological data, [8], and it maximises the distances between different clusters while compacting individual groups. Jackson [69] reviews and experiments with a number of scaling mechanisms, but does not consider this particular algorithm. Reynoldson et al. [136] and Faith et al. [35] have shown it to be effective in recent studies. The scores from the scaling algorithm were clustered by an agglomerative hierarchical fusion method using Unweighted Pair Group Mean Average (UPGMA) [8]. The selection of the groups was based on information from the dendrogram produced by the clustering and on the visual (spatial) separation of the groups in ordination space. The subjectivity in this procedure was noted, but the scope of the study did not permit an investigation into the comparative worth of other algorithms. The software package PATN [8], developed by the Australian CSIRO, was used for the ordination.

### 7.3.2 Ordination of Environmental Variables

The environmental variables used in the study are given in Table 7.1. Of the 43 variables listed only 27 (in italics) were considered as the base set of predictor variables for the experimental work. The base matrix contained data on the 27 variables (columns) for the 93 available sites (rows). Site numbers 5602 and 5802 had 4 and 15 missing data values respectively, the missing data being replaced with the mean of the remaining cases. Summary statistics for the full data set are given in the last two columns of Table 7.4.

| **Geo-physical** | **Water** |
|---|---|
| *Water Depth (DPTH)** | *Alkalinity mg/l (ALK)** |
| *Bottom temp. (TMP)* | *Total phosphorus mg/l (TPW)* |
| Latitude | *Kjeldahl nitrogen mg/l (TKN)** |
| Longitude | *Nitrate-nitrite mg/l (NO)** |
| *pH** | Ammonia mg/l |
| *Oxygen mg/l (OXY)* | |

**Sediment Chemistry ($\mu$ g/g drt wt )**

| | | |
|---|---|---|
| *Silica (SI)** | *%Sand (SN)** | Nickel |
| Titanium | *%Silt (SL)** | Copper |
| *Aluminium* (AL)* | *%Clay (CL)* | Zinc |
| *Iron (FE)* | *%Gravel (GR)* | Arsenic |
| *Manganese (MN)* | Selenium | Strontium |
| *Magnesium (MG)* | *Vanadium (V)** | Yttrium |
| *Calcium (CA)** | *Chromium(CR)* | Molybdenum |
| *Sodium (NA)** | *Cobalt (CO)* | Silver |
| *Potassium (K)* | *T. Org. Carbon (TOC)* | Cadmium |
| *Total phosphorus (TP)* | *Loss on Ignition (LOI)** | Tin |
| *Total nitrogen (TN)* | | Lead |

Table 7.1: Summary of measured environmental variables, *italics* denote set of 27 variables used as predictors, '*' denotes subset of thirteen.

Further data sets were derived from the base data for training the neural network models, the key to these sets is shown in Table 7.2. The extra data sets were generated by standardisation and principal component analysis (PCA). The data was standardised by subtracting the class mean and dividing through by the standard deviation, thus producing $z$ scores. The principal component analysis produced 7 eigenvalues greater than unity, and the corresponding eigenvectors, following a varimax rotation, were used as the third data set. The first 7 factors accounted for 78.8% of the variance.

From a preliminary step-wise discriminant analysis using the environmental variables as predictors and groupings derived from the ordination of the community structure (see Section 7.3.2), 13 variables were highlighted as a good set of predictor variables, these are denoted with '*' in Table 7.1. These were also found to have good correlation with the ordination scores. A further three

| Description of matrix | Rows | Columns | Key |
|---|---|---|---|
| Full raw | 93 | 27 | E27raw |
| Full standardised | 93 | 27 | E27std* |
| PCA of standardised | 93 | 7 | E27pca7* |
| Subset of raw | 93 | 13 | E13raw |
| Subset standardised | 93 | 13 | E13std* |
| PCA of standardised subset | 93 | 4 | E13pca4* |

Table 7.2: Key to data sets used for training neural net models ('*' denotes data sets used in Sections 7.4 to 7.7).

matrices were developed by standardisation and PCA (see Table 7.2 for key). The PCA reduced the matrix from 13 to 4 columns, with the 4 dimensions accounting for 83.2% of the variance.

The E13std data were ordinated using the strategy outlined in Section 7.3.1. A three dimensional ordination analysis suggested that a reduction to two dimensions may be possible, and two dimensions were chosen after the analysis gave a stress value of 0.11. The stress value is used to rank separate ordination analyses, and a value of below 0.15 indicates an acceptable result.

Six groups were identified from the ordination analysis. Their position in ordination space is shown in Figure 7.3 and their geographic distribution is given in Table 7.3. Figure 7.3 shows that there is good discrimination between the groups, and that the Lake Michigan sites (mainly Groups 3 and 6) are well clustered to the right of the diagram. Table 7.4 shows the mean and standard deviations of the 13 variables for the six individual groups, and this shows that the sites in Groups 3 and 6 are characterised by being deeper and having a greater alkalinity, low Kjeldahl Nitrogen and sodium concentrations and high Nitrate/Nitrite concentrations. Sites from Lake Erie (Groups 1 and 4) and four sites from Lake Ontario (Group 1) are also well defined, lying on the leading diagonal of the figure. These sites are relatively shallow, have high calcium concentrations and pH values and high Kjeldahl Nitrogen concentrations. Finally sites from Georgian Bay, Lake Superior and the North Channel (L. Huron) form the majority of Groups 2 and 5, to the top and left of the plot. These sites are characterised by high aluminium and vanadium concentrations, and low calcium and alkalinity.

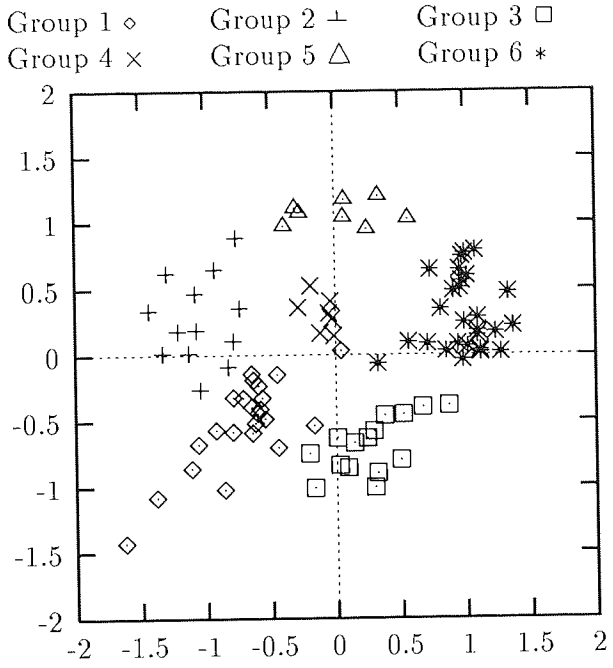



Figure 7.3: Location of environmental groups in ordination space.

| Lake | Environmental Group | | | | | |
|---|---|---|---|---|---|---|
| | 1 (n=23) | 2 (n=13) | 3 (n=15) | 4 (n=7) | 5 (n=8) | 6 (n=27) |
| L.Erie (25) | 16 | 1 | 1 | 7 | 0 | 0 |
| L.Ontario (5) | 4 | 0 | 0 | 0 | 1 | 0 |
| L.Michigan (43) | 2 | 0 | 14 | 0 | 0 | 27 |
| Georgian Bay (9) | 1 | 6 | 0 | 0 | 2 | 0 |
| L.Superior (5) | 0 | 4 | 0 | 0 | 1 | 0 |
| North Channel (6) | 0 | 2 | 0 | 0 | 4 | 0 |

Table 7.3: Geographic distribution of sites of 6 groups derived from ordination of environmental variables.

| Environmental Variable | Group 1 (n=23) | | Group 2 (n=13) | | Group 3 (n=15) | | Group 4 (n=7) | | Group 5 (n=8) | | Group 6 (n=27) | | All Groups (n=93) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| SI | 49·34 | 11·57 | 58·54 | 5·08 | 54·31 | 7·02 | 57·36 | 3·21 | 70·58 | 8·06 | 75·68 | 11·25 | 61·50 | 14·08 |
| AL | 8·90 | 1·81 | 12·02 | 1·28 | 8·50 | 2·19 | 9·30 | 1·41 | 10·56 | 2·01 | 4·7 | 1·27 | 8·22 | 3·00 |
| CA | 11·74 | 7·05 | 4·01 | 2·87 | 7·84 | 2·89 | 8·97 | 3·04 | 2·80 | 1·87 | 4·11 | 3·06 | 6·84 | 5·39 |
| NA | 1·22 | ·37 | 2·17 | ·39 | 0·84 | ·29 | 1·65 | ·39 | 2·26 | ·67 | 0·69 | ·19 | 1·26 | 0·68 |
| LOI | 18·18 | 7·74 | 10·88 | 3·40 | 16·30 | 3·42 | 10·89 | ·97 | 4·91 | 2·94 | 7·38 | 4·94 | 12·03 | 6·98 |
| SN | 9·85 | 9·71 | 11·01 | 10·62 | 19·91 | 16·95 | 50·56 | 7·73 | 71·49 | 22·39 | 88·46 | 12·53 | 42·82 | 36·81 |
| SL | 58·31 | 10·25 | 44·39 | 20·54 | 35·84 | 19·21 | 22·02 | 9·28 | 9·63 | 9·19 | 1·10 | 4·10 | 29·21 | 25·95 |
| V | 22·52 | 6·20 | 47·54 | 14·33 | 25·57 | 7·93 | 24·57 | 10·15 | 34·75 | 17·82 | 14·96 | 10·79 | 25·52 | 14·84 |
| DPTH | 9·42 | 7·71 | 6·92 | 3·32 | 68·73 | 19·57 | 10·38 | 7·00 | 8·19 | 3·47 | 44·36 | 23·02 | 28·75 | 27·98 |
| PH | 8·08 | ·26 | 7·18 | ·28 | 7·75 | ·23 | 8·18 | ·10 | 7·17 | ·38 | 7·99 | ·64 | 7·81 | 0·54 |
| ALK | 89·10 | 9·95 | 63·4 | 15·04 | 111·4 | 4·67 | 88·90 | 5·52 | 58·91 | 10·07 | 110·5 | 4·52 | 92·71 | 21·14 |
| TKN | ·218 | ·058 | ·211 | ·126 | ·123 | ·021 | ·206 | ·041 | ·198 | ·109 | ·130 | ·060 | 0·17 | 0·08 |
| NO | ·147 | ·106 | ·159 | ·118 | ·355 | ·050 | ·156 | ·100 | ·147 | ·104 | ·283 | ·092 | 0·22 | 0·12 |

Table 7.4: Mean and standard deviation of the environmental variables; grouped on the basis of the ordination of the environmental data.

### 7.3.3 Ordination of Community Structure

The matrix for the community structure data consisted of the available 93 samples representing the rows, with the abundances of the taxa as the attributes (columns). For all taxa the number of occurrences were recorded, not just absence/presence. The list of taxa consisted of 103 species and 44 genera plus some higher level groups where the taxonomy was only taken to class, order or family levels (e.g. Porifera, Platyhelminthes and Nematodes). Of all the taxonomic groups identified, the Chironomids were the most diverse with 42 genera and the Oligochaeta next with 37 species. For each site there were 5 replicate samples from which the mean value of the abundance of each taxon was evaluated and used in the analyses. To make the ordination procedure tractable the number of taxa was reduced to 55, on the basis that each taxon had to have an abundance of over 0.05% within the data set. The new data contained 17 Chironomids, 5 Naidids, 7 Tubificid taxa and 6 species of the Sphaeridae family. Also to be found were other taxa including Valvatidae, Sabellidae, Amphipoda and Asellidae. See Appendix A3 for the complete species list.

For the ordination of the community structure a three dimensional space was found to be the most suitable, the location of the reference sites being shown in Figure 7.4. Five groups were identified from the clustering procedure, and Table 7.5 shows the geographic distribution of these groups. The L. Michigan sites form a large group, which are all deep, have a low water temperature and high alkalinity, and have a substrate which is more sandy and less silty than the other groups, Table 7.6. Sites from L. Erie which form the majority of Group 1 are deeper than the sites which cluster in Groups 2 and 4, with the shallow sites of Group 4 being from Long Point Bay, L.Erie and Presque'Ile Bay Lake Ontario. There is a strong correlation between the Vector 1 scores and the depth of the sites, with high (positive) scores on Vector 1 corresponding to shallow sites, and low (negative) scores the deeper ones. Positive scores on Vector 2 are indicative of oligotrophic conditions, with negative values (coupled with positive Vector 1 scores) corresponding to eutrophic waters.

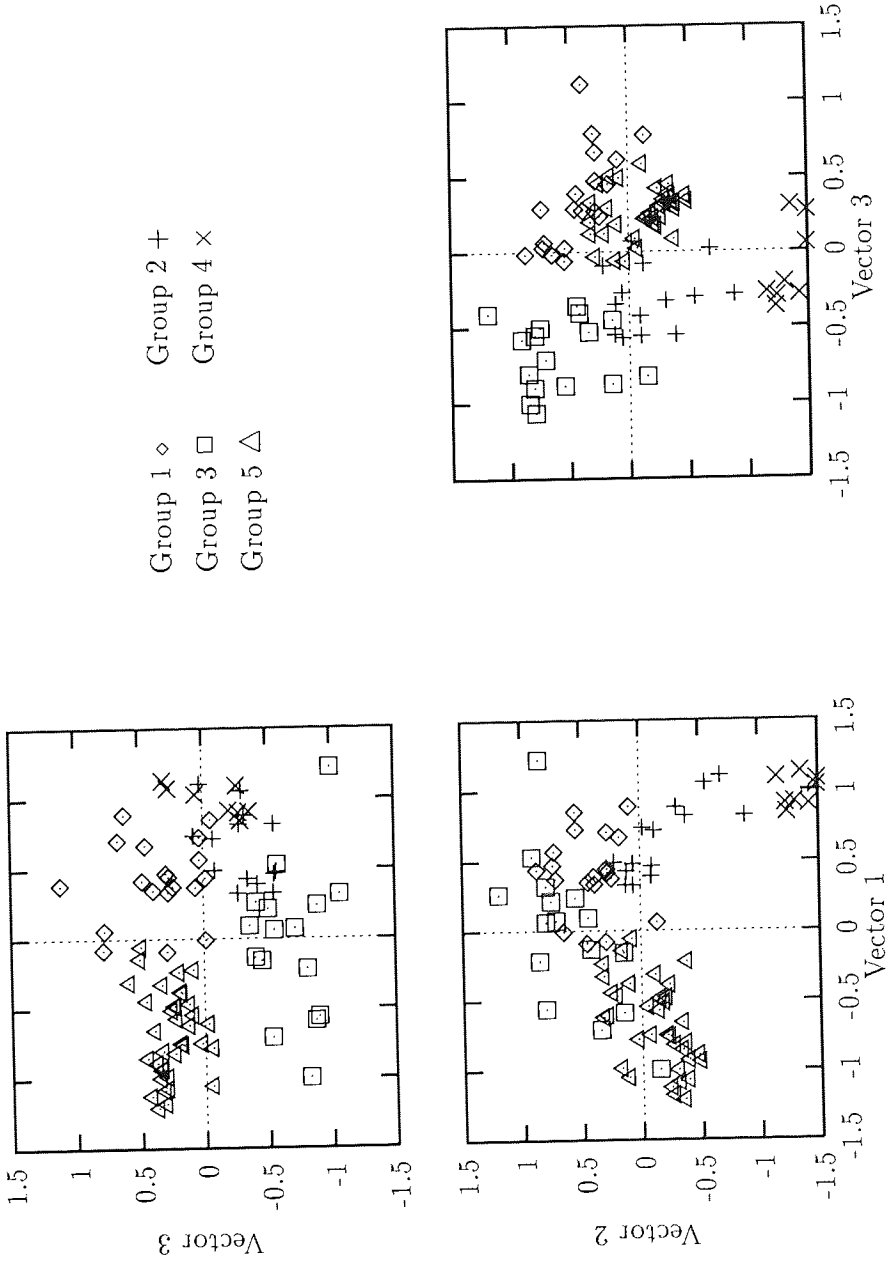



Figure 7.4: Ordination of benthic community structure.

| Lake | Biological Group | | | | |
|---|---|---|---|---|---|
| | 1 (n=19) | 2 (n=14) | 3 (n=16) | 4 (n=8) | 5 (n=36) |
| L.Erie (25) | 16 | 4 | 0 | 5 | 0 |
| L.Ontario (5) | 0 | 2 | 0 | 3 | 0 |
| L.Michigan (43) | 3 | 0 | 4 | 0 | 36 |
| Georgian Bay (9) | 0 | 8 | 1 | 0 | 0 |
| L.Superior (5) | 0 | 0 | 5 | 0 | 0 |
| North Channel (6) | 0 | 0 | 6 | 0 | 0 |

Table 7.5: Geographic distribution of sites of 5 groups derived from the community structure ordination.

Table 7.7 shows the species (in descending order of occurrence) that occur in at least 50% of sites in a particular group. The sites of Groups 3 and 5 have a composition of species that is usually associated with oligotrophic conditions (e.g. *Diporeia hoyi*). The fauna in Groups 1 and 3 are associated with more organic waters, while the shallow sites of Group 4 are dominated by the Porifera (sponges).

The confusion matrix showing the site classification by biological and environmental ordination, Table 7.8, gives an indication of the disagreement between the two methods. Ideally there should be good agreement between the two ordinations, with a small amount of scatter or overlap between the groups. From Table 7.8, it can be seen that the L. Michigan sites, which constitute the biological group 5, are situated in two environmental classes, 3 and 6. The sites in the biological groups 1 and 4 are concentrated in the environmental groups 1 and 4, while the biological groups 2 and 3 have some spread over the environmental classes. Thus, in the prediction experiments it likely that these sites in the biological groups 2 and 3 will be the most difficult to classify, as they are poorly correlated with the underlying structure of the environmental data.

### 7.3.4 Ordination of Bioassay Data

The endpoints used in the toxicity tests are shown in Table 7.9, along with the test duration. For complete details of the sediment bioassay experimental

| Environmental Variable | Group 1 (n=19) | | Group 2 (n=14) | | Group 3 (n=16) | | Group 4 (n=8) | | Group 5 (n=36) | | All Groups (n=93) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| SI | 57·00 | 8·37 | 54·67 | 9·14 | 69·21 | 11·56 | 45·05 | 16·69 | 66·77 | 13·94 | 61·50 | 14·08 |
| AL | 9·48 | 2·24 | 10·94 | 2·10 | 9·00 | 3·42 | 8·02 | 2·15 | 6·21 | 2·27 | 8·22 | 3·00 |
| FE | 4·11 | 1·66 | 4·81 | 1·98 | 3·54 | 1·86 | 2·76 | 0·64 | 3·02 | 2·53 | 3·58 | 2·14 |
| MN | 0·13 | 0·16 | 0·11 | 0·07 | 0·07 | 0·04 | 0·07 | 0·02 | 0·15 | 0·25 | 0·12 | 0·17 |
| MG | 3·32 | 0·68 | 2·76 | 1·51 | 2·23 | 1·33 | 1·99 | 0·73 | 3·43 | 2·00 | 2·98 | 1·60 |
| CA | 8·10 | 3·35 | 6·79 | 5·62 | 3·74 | 2·92 | 15·80 | 9·36 | 5·57 | 3·50 | 6·84 | 5·39 |
| NA | 1·26 | 0·39 | 1·79 | 0·60 | 1·79 | 0·84 | 1·51 | 0·82 | 0·76 | 0·24 | 1·26 | 0·68 |
| K | 2·49 | 0·50 | 2·36 | 0·48 | 1·87 | 0·36 | 1·66 | 0·57 | 2·14 | 0·46 | 2·16 | 0·52 |
| TP | 0·17 | 0·07 | 0·21 | 0·09 | 0·11 | 0·05 | 0·17 | 0·05 | 0·43 | 1·81 | 0·27 | 1·13 |
| LOI | 12·37 | 2·81 | 13·53 | 6·78 | 7·42 | 4·23 | 21·92 | 11·41 | 11·12 | 6·27 | 12·03 | 6·98 |
| TN | 1433·26 | 655·22 | 3152·86 | 2130·11 | 1588·63 | 1445·47 | 4132·63 | 3154·51 | 1271·04 | 1022·92 | 1888·26 | 1763·47 |
| TOC | 1·50 | 0·70 | 2·89 | 1·89 | 1·76 | 1·96 | 4·54 | 3·25 | 1·47 | 1·20 | 2·00 | 1·85 |
| GR | 0·70 | 3·01 | 0·10 | 0·26 | 0·05 | 0·20 | 0·12 | 0·35 | 0·10 | 0·54 | 0·22 | 1·40 |
| SN | 24·90 | 26·89 | 25·36 | 33·07 | 47·54 | 35·81 | 23·38 | 34·15 | 61·30 | 35·45 | 42·82 | 36·81 |
| SI | 40·27 | 19·77 | 42·87 | 27·06 | 25·97 | 24·67 | 50·71 | 24·59 | 14·73 | 21·19 | 29·21 | 25·95 |
| CL | 34·13 | 16·94 | 31·67 | 22·98 | 26·44 | 18·16 | 25·78 | 12·96 | 23·89 | 21·55 | 27·76 | 19·78 |
| V | 25·63 | 8·91 | 34·79 | 18·46 | 34·19 | 20·05 | 17·75 | 5·26 | 19·74 | 10·81 | 25·52 | 14·84 |
| CR | 21·68 | 10·17 | 34·86 | 21·06 | 33·25 | 20·39 | 15·50 | 10·30 | 21·59 | 15·50 | 25·09 | 17·07 |
| CO | 6·63 | 3·53 | 10·14 | 6·59 | 11·00 | 9·36 | 3·06 | 2·95 | 7·01 | 4·40 | 7·75 | 6·00 |
| DPTH | 15·95 | 9·28 | 8·34 | 3·99 | 8·59 | 7·07 | 1·86 | 0·65 | 58·38 | 21·85 | 28·75 | 27·98 |
| OXY | 9·92 | 1·02 | 8·40 | 1·65 | 10·04 | 1·47 | 9·39 | 1·04 | 10·39 | 1·39 | 9·85 | 1·49 |
| PH | 8·10 | 0·19 | 7·59 | 0·54 | 7·31 | 0·49 | 8·16 | 0·14 | 7·87 | 0·57 | 7·81 | 0·54 |
| TMP | 11·73 | 1·80 | 17·23 | 4·33 | 16·78 | 2·62 | 15·23 | 1·11 | 6·47 | 2·92 | 11·69 | 5·35 |
| ALK | 92·54 | 7·32 | 82·16 | 11·19 | 67·45 | 27·59 | 78·65 | 10·39 | 111·24 | 4·26 | 92·71 | 21·14 |
| TPW | 0·01 | 0·01 | 0·02 | 0·03 | 0·05 | 0·13 | 0·02 | 0·01 | 0·01 | 0·00 | 0·02 | 0·05 |
| TKN | 0·19 | 0·03 | 0·20 | 0·03 | 0·19 | 0·12 | 0·31 | 0·08 | 0·12 | 0·03 | 0·17 | 0·08 |
| NO | 0·22 | 0·08 | 0·10 | 0·08 | 0·20 | 0·08 | 0·01 | 0·00 | 0·33 | 0·07 | 0·22 | 0·12 |

Table 7.6: Mean and standard deviation of the environmental variables; groups from the community structure ordination.

| Group 1 | Group 2 | Group 3 |
|---|---|---|
| **Tubificidae (co hr)** | **Porifera** | **Tubificidae (co hr)** |
| **Tubificidae (c hr)** | **Tubificidae (co hr)** | Tubificidae (c hr) |
| **Procladius spp** | **Procladius spp** | Procladius spp |
| Pisidium spp | **Cryptochironomus spp** | Pisidium casertanum |
| Pisidium casertanum | **Chironomus spp** | Diporeia hoyi |
| Porifera | **Pisidium casertanum** | Micropsectra spp |
| Spirosperma ferox | **Tubificidae (c hr)** | |
| Dreissena polymorpha | **Tanytarsus spp** | |
| Platyhelminthes | Valvata tricarinata | |
| Limnodrilus hoffmeisteri | Aulodrilus pigueti | |
| Chironomus spp | Platyhelminthes | |
| Aulodrilus pigueti | Pisidium spp | |
| | Polypedium spp | |

| Group 4 | Group 5 |
|---|---|
| **Porifera** | **Diporeia hoyi** |
| **Procladius spp** | **Stylodrilus herringlanus** |
| **Chironomus spp** | **Pisidium spp** |
| **Dicrotendipes spp** | **Vejdovskyella intermedia** |
| Microtendipes spp | Platyhelminthes |
| Cryptochironomus spp | Heterotrissocladius spp |
| Tubificidae (co hr) | Pisidium casertanum |
| Physella spp | Tubificidae (co hr) |
| Polypedium spp | Tubificidae (c hr) |
| Pisidium casertanum | |
| P. nitidum | |
| Endochironomus spp | |
| Pseudochironomus spp | |

Table 7.7: Species which occur in at least 50% of sites in a group in descending frequency of occurrence, **bold** > 70% occurrence (after Reynoldson et al. [136]).

| Environmental Group | Biological Group | | | | |
|---|---|---|---|---|---|
| | 1 (n=19) | 2 (n=14) | 3 (n=16) | 4 (n=8) | 5 (n=36) |
| 1 (n=23) | 11 | 6 | 0 | 6 | 0 |
| 2 (n=13) | 1 | 5 | 7 | 0 | 0 |
| 3 (n=15) | 1 | 0 | 0 | 0 | 14 |
| 4 (n=7) | 5 | 1 | 0 | 1 | 0 |
| 5 (n=8) | 0 | 2 | 5 | 1 | 0 |
| 6 (n=27) | 1 | 0 | 4 | 0 | 22 |

Table 7.8: Confusion matrix showing classification of sites from ordination of biological community structure and environmental variables.

| Taxon | Length of test (days). | Endpoints |
|---|---|---|
| *Hyalella azteca* | 28 | survival and growth |
| *Chironomus riparius* | 10 | survival and growth |
| *Hexagenia spp.* | 21 | survival and growth |
| *Tubifex tubifex* | 28 | production of cocoons and young |

Table 7.9: Taxa, test duration and endpoints of toxicity tests.

procedure see Reynoldson et al. [136]. For each of the four taxa two endpoints were used, although other endpoints were available which could have been considered in the analysis. But in order to keep an even weighting, and to avoid the possibility of the analysis becoming too dependent on any particular test species, equal numbers of endpoints were used for each. A two-dimensional ordination space was used in the analysis with 3 distinct groups being identified from the clustering procedure. Figure 7.5 shows the position and grouping of the sites in ordination space, and Table 7.10 summarises the means and standard deviations of the toxicity tests for each of the 3 groups.

From Figure 7.5 and Table 7.10 it can be noted that the Group 3 sites (5600, 5708, 5804 and 5805) represent samples that may be contaminated, with the survival and growth of *H. azteca* being much lower than that for Groups 1 and 2. Group 1 can be separated from Group 2 by the lower reproduction of *T.*

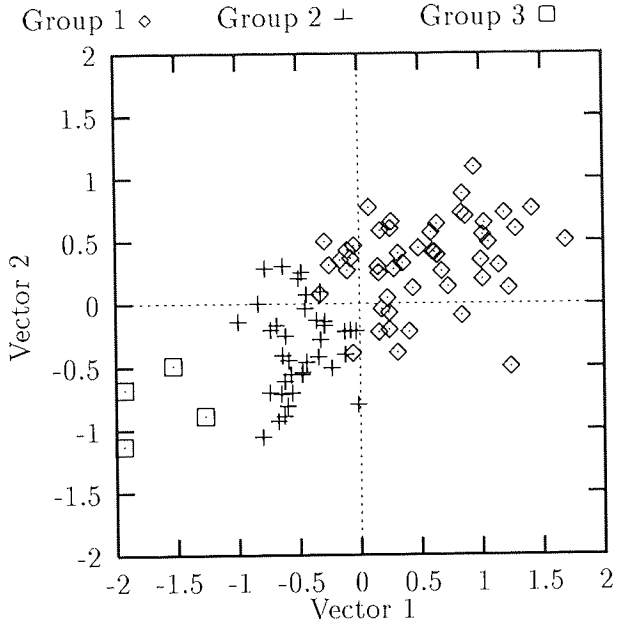



Figure 7.5: Location of bioassay groups in ordination space.

| Endpoints | Group 1 (n=53) | | Group 2 (n=37) | | Group 3 (n=3) | | All Groups (n=93) | |
|---|---|---|---|---|---|---|---|---|
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| *C. riparius* | | | | | | | | |
| % survival | 80·2 | 7·8 | 86·6 | 8·6 | 82·9 | 6·2 | 82·9 | 8·6 |
| growth∗ | 0·33 | 0·08 | 0·37 | 0·08 | 0·30 | 0·04 | 0·34 | ·08 |
| *H. azteca* | | | | | | | | |
| % survival | 88·7 | 11·8 | 84·7 | 12·9 | 24·5 | 22·4 | 83·7 | 19·0 |
| growth∗ | 0·51 | 0·11 | 0·50 | 0·17 | 0·29 | 0·11 | 0·50 | 0·14 |
| *Hexagenia sp.* | | | | | | | | |
| % survival | 97·4 | 3·4 | 96·4 | 4·8 | 95·5 | 4·6 | 96·9 | 4·1 |
| growth∗ | 3·07 | 2·56 | 4·43 | 4·64 | 1·49 | 0·31 | 3·5 | 3·6 |
| *T. tubifex* | | | | | | | | |
| cocoons | 32·8 | 6·1 | 37·6 | 4·2 | 35·8 | 2·8 | 34·9 | 5·8 |
| young | 58·5 | 23·5 | 124·0 | 23·0 | 122·2 | 24·9 | 87·7 | 40·0 |

Table 7.10: Means and standard deviations of environmental variables for toxicity test endpoints (∗ mg dry wt); groups based on ordination of bioassay endpoint data.

*tubifex*, but apart from this there is little difference between the endpoints of these groups. The small overall variation of the toxicity test results is to be expected, as one of the criteria for including a site in the reference database is that it represents 'uncontaminated' conditions. If a sample did have a large or unusual deviation from the average, then doubt would be cast on the validity of the sample representing 'clean' conditions, and it would be excluded from the database.

## 7.4 Classification of Biological Groups

### 7.4.1 Objectives

The classification of the environmental variables into ordination groups, as described in Fig. 7.2 (Section 7.2.4), is the key to the success of the overall study. It is a *one*-from-$N$ classification which in this study can be accomplished by two broad alternative methods: multiple discriminant analysis and multi-layer perceptrons. There are also countless other techniques which have not been considered in this dissertation. Michie et al. [104], reporting on the StatLog project, and Ripley [141] compare a number of different learning algorithms.

### 7.4.2 Preliminary Experiments

#### 7.4.2.1 Procedure

The basic network configuration (see Section 3.2) with a single hidden layer having 5, 7 and 9 hidden units formed the basic model used in this study. The four data sets from Section 7.3.2 were used as input to the network. The target data were the biological groups identified from the ordination (Section 7.3.3); five nodes being used in the output layer to represent the five different types of biological group. For each network four values of the weight decay parameter were also used: 0.0, 0.0001, 0.001 and 0.01. This gave a total of 48 different configurations for each experiment. In addition ten random starts of the weight parameters were run for each procedure, with the weights being initialised between [-0.3:0.3]. Since the available data set was small, containing only 93 samples, a leave-one-out cross validation method was used to gain a measure

of the generalisation ability of the networks. A conjugate gradient method was used as the minimisation procedure, with the weights being updated after each pass of the training data set.

#### 7.4.2.2 Results

Table 7.11 shows the average and standard deviation of the classification rate, the figures are derived from the average of the 10 random starts. To summarise the table: the critical factor on networks performance was the training data set, the E13pca4 data gave the poorest performance overall and was generally 10% worse when compared to the best set E13std. It is interesting to note that both data sets derived from the principal component analysis had a poorer performance than the original full data set.

Considering all the results, the number of hidden units had little effect on the performance of the networks, except that there was a marginal benefit in using more hidden units when the weight decay parameter was not equal to zero. An interpretation of this may be that as the weight decay parameter inhibits the complexity of the network, the mapping requires the extra degrees of freedom provided by the additional hidden units. Of the three non-zero weight decay $\lambda$'s the 0.01 value could be judged to be of least benefit. It can be concluded that the weight-decay term is useful, and the method is robust over a range of values.

### 7.4.3 Analysis of Results of Preliminary Experiments

#### 7.4.3.1 Procedure

In this section a single network is considered, having $\lambda = 0.001$, 9 hidden nodes and the training set E13std configuration. This network provided the highest average classification of all the configurations used in Section 7.4.2. The results from this simulation are analysed with particular emphasis being placed on the misclassifications, and whether there is an underlying explanation for the errors.

The misclassifications from the 10 individual simulations were collated and sites that were consistently predicted in error were identified. The structure of the training set was examined using the ordination analysis of Section 7.3.2

| Weight Decay $\lambda$ and number of hidden nodes | Training data set (see Section 7.3.2) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | E27std | | E27pca7 | | E13std | | E13pca4 | |
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| $\lambda = 0.0$ | | | | | | | | |
| h=5 | 75.3% | 3.1 | 71.7% | 1.6 | 77.7% | 2.7 | 70.1% | 4.4 |
| h=7 | 77.1% | 2.6 | 71.8% | 2.7 | 78.8% | 2.7 | 70.7% | 2.4 |
| h=9 | 76.2% | 2.4 | 71.3 % | 3.5 | 79.9 % | 1.5 | 69.6 % | 2.4 |
| $\lambda = 0.0001$ | | | | | | | | |
| h=5 | 76.1 % | 1.5 | 71.4 % | 2.3 | 78.5 % | 2.4 | 69.4 % | 2.1 |
| h=7 | 77.0 % | 2.4 | 71.4 % | 2.0 | 80.3 % | 3.0 | 63.4 % | 2.0 |
| h=9 | 79.7 % | 2.5 | 72.5 % | 3.0 | 79.2 % | 2.4 | 70.4 % | 3.9 |
| $\lambda = 0.001$ | | | | | | | | |
| h=5 | 77.2 % | 3.5 | 70.2 % | 3.0 | 79.0 % | 3.1 | 70.1 % | 4.2 |
| h=7 | 75.5 % | 2.3 | 72.5 % | 2.1 | 81.2 % | 2.4 | 70.4 % | 1.9 |
| h=9 | 76.5 % | 2.3 | 72.3 % | 3.6 | 81.5 % | 2.3 | 70.8 % | 2.8 |
| $\lambda = 0.01$ | | | | | | | | |
| h=5 | 75.3 % | 3.0 | 68.3 % | 1.5 | 76.3 % | 3.2 | 50.9 % | 3.2 |
| h=7 | 77.4 % | 2.8 | 74.2 % | 3.7 | 76.7 % | 2.7 | 71.0 % | 2.6 |
| h=9 | 78.3 % | 3.7 | 73.9 % | 2.3 | 78.6 % | 1.5 | 71.2 % | 2.8 |

Table 7.11: Classification rates for 93 sites to Biological Group using environmental variables.

| Network | Percentage Correct | Incorrect Classifications |
|---|---|---|
| 1 | 83.9% | 15 |
| 2 | 83.9% | 15 |
| 3 | 82.8% | 16 |
| 4 | 82.8% | 16 |
| 5 | 82.8% | 16 |
| 6 | 81.7% | 17 |
| 7 | 81.7% | 17 |
| 8 | 79.6% | 19 |
| 9 | 78.5% | 20 |
| 10 | 77.4% | 21 |

Table 7.12: Rank classification order for networks considered in Section 7.4.3.

to determine if there was any correlation between the misclassifications and groupings or general features within the set.

#### 7.4.3.2 Results

The 10 networks were ranked in ascending order of cross-validation error rate, as given in Table 7.12. In total, for all 10 networks, there were 172 misclassified samples, with 34 different samples being misclassified at least once. Thirteen of these 34 were misclassified 5 or more times, details of which are given in Table 7.13. These 13 sites accounted for a total of 124 (72%) of the misclassifications, and the dendrogram in Figure 7.6 shows the position of the misclassified sites from the ordination of the environmental data, as well as the biological and environmental classes of each site.

Examination of the dendrogram indicates that the majority of the sites consistently misclassified occurred where there was some disagreement between its biological group and that of its near neighbours. For example, sites 5800 and 5803 (biological groups 1 and 3 respectively) occur in the middle of a large cluster of group 5 sites. Of the thirteen sites 8 of the misclassifications can be interpreted in this manner. This suggests that the network is placing a site into a class associated with the closest region of the input space, as sites in close proximity on the dendrogram would also be in close proximity

Figure 7.6: Dendrogram of environmental ordination showing misclassified sites.

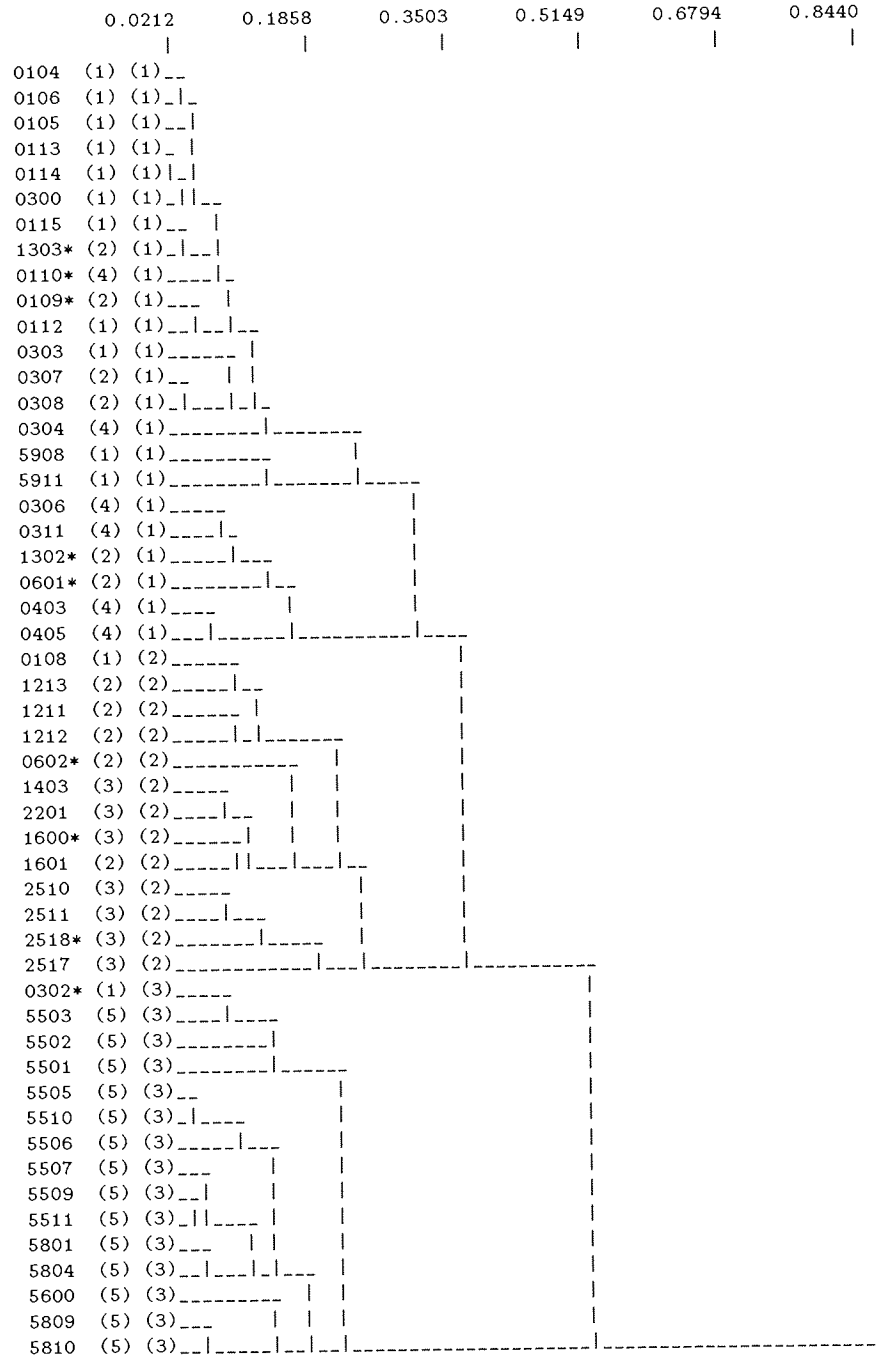
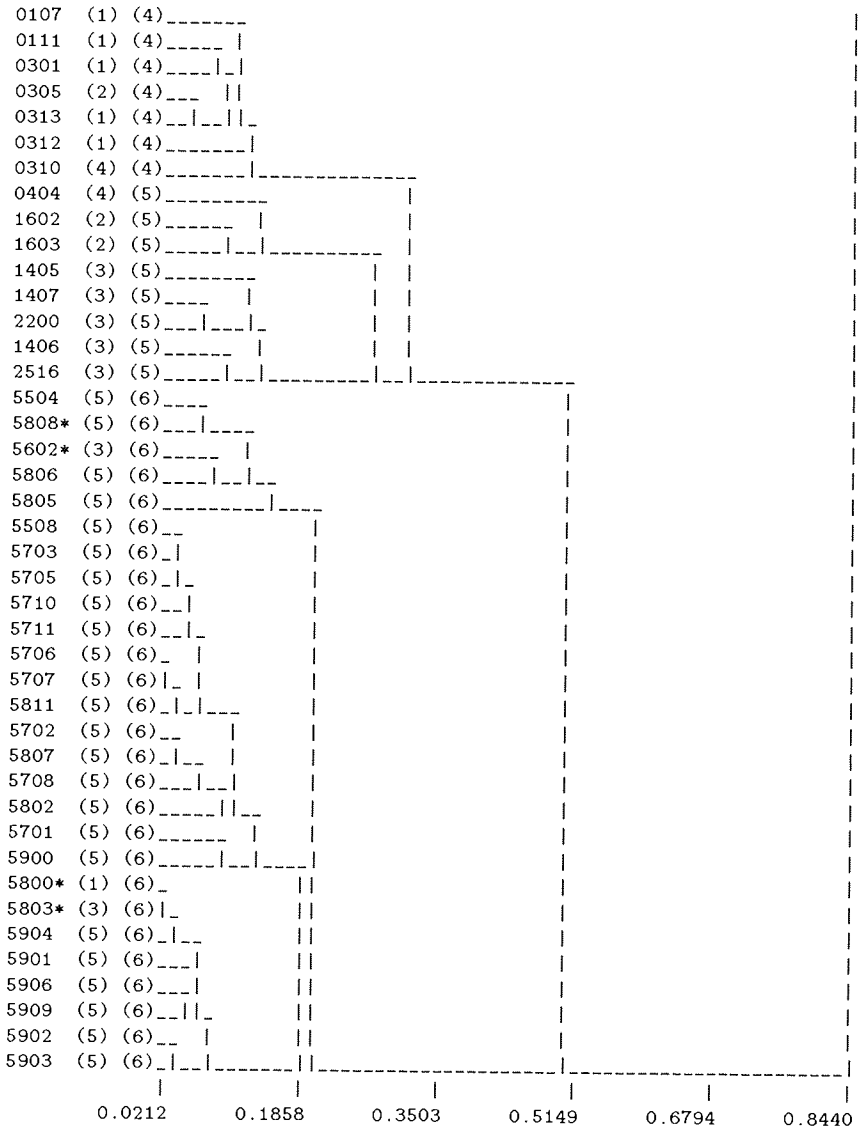



Figure 7.6 continued overleaf

Figure 7.6: Dendrogram of environmental ordination (cont'd).




*Sites denoted by * are consistently misclassified by the neural networks. The site reference number is given, with the first number in brackets being the biological group derived from the cluster analysis of the community structure ordination (Section 7.3.3), while the second gives the class from the ordination of the environmental variables (Section 7.3.2).*

| Site Ref. | Incorrect classifications | Target group | Predicted groups. Brackets indicate frequency |
|---|---|---|---|
| 0109 | 10 | 2 | 1(10) |
| 0110 | 8 | 4 | 2(8) |
| 0302 | 10 | 1 | 5(10) |
| 0601 | 9 | 2 | 3(7),4(2) |
| 0602 | 9 | 2 | 3(9) |
| 1302 | 9 | 2 | 4(9) |
| 1303 | 10 | 2 | 1(10) |
| 1600 | 10 | 3 | 2(9),4(1) |
| 2518 | 9 | 3 | 1(7),2(3) |
| 5602 | 10 | 3 | 5(9) |
| 5800 | 10 | 1 | 3(8),2(2) |
| 5803 | 10 | 3 | 1(9),5(1) |
| 5808 | 10 | 5 | 1(10) |

Table 7.13: Sites misclassified more than 5 times by networks in Table 7.12.

in the input space of the network. The above is what would be expected, and indicates that the grouping of the community structure is conflicting with the natural grouping of the environmental variables. A reason for this is probably the small number of samples under consideration, with groups being under-represented and thus not able to form distinct clusters. As a larger data set will be used for the final system the present groups will probably change, with a likely improvement in the definition of the clustering.

### 7.4.4 Committees of Networks

#### 7.4.4.1 Procedure

The 10 networks used in the preceding section were ranked in order of performance and committees formed using the simple average and product average methods, described in Section 5.4. Nine committees were formed for each method: the first committee comprised the two networks with the lowest error-rate on the cross-validation data, whilst the second committee comprised the three lowest, and so forth until all 10 networks were combined.

| Network | Percentage Correct | Product Average | Simple Average |
|---|---|---|---|
| 1 | 83.9% | — | — |
| 2 | 83.9% | 83.9% | 86.0% |
| 3 | 82.8% | 83.9% | 86.0% |
| 4 | 82.8% | 82.8% | 85.0% |
| 5 | 82.8% | 81.7% | 86.0% |
| 6 | 81.7% | 82.8% | 83.9% |
| 7 | 81.7% | 83.9% | 85.0% |
| 8 | 79.6% | 83.9% | 83.9% |
| 9 | 78.5% | 83.9% | 82.5% |
| 10 | 77.4% | 83.9% | 83.9% |

Table 7.14: Classification rates of biological group from environmental variables for committees of networks.

#### 7.4.4.2 Results

The results are shown in Table 7.14. The collective performance of the individuals combined using the product average was no better than the best individual, while there was a slight improvement when the simple averaging procedure was adopted (86%). The performance is similar to that obtained by Reynoldson et al. [136] using MDA, which gave 90% under less exacting validation conditions. It is interesting to note that the error rate of 14% indicates that 13 samples out of the 93 were misclassified, and that the thirteen sites were the same as those reported in Section 7.4.3.2. Thus the simple averaging procedure has removed the random misclassifications of the individual networks, which is what would be expected by such a technique.

### 7.4.5 Discriminant Analysis

#### 7.4.5.1 Procedure

The most commonly used classification tool in community structure studies is probably multiple discriminant analysis (MDA). For example, the studies by Wright et al. [181] and Reynoldson et al. [136] both used this technique. Mitchell [106] discusses discriminant analysis and gives details of the algorithms, while McLachlan [101] covers the subjects of discriminant analysis and

| Method | Input data | | | |
|---|---|---|---|---|
| | E27std | E27pca7 | E13std | E13pca4 |
| Linear | 77.4% | 77.4% | 77.4% | 75.3% |
| Quadratic | 79.6% | 77.4% | 82.8% | 74.2% |

Table 7.15: Linear and quadratic discriminant analysis, classification rates of biological groups from environmental variables.

statistical pattern recognition in greater depth.

The four data sets: E27std, E27pca7, E13std and E13 pca4 (see Table 7.2), as discussed in the proceeding sections of this chapter, were used as the model's predictor variables, with leave-one-out cross validation used to assess the generalisation of the technique. Each of data sets were run against the linear and quadratic models, so a series of eight experiments was completed. The software used was Public Domain code, written by Henery (see [104]), and was the discriminant analysis software used in the StatLog project, Michie et al. [104].

#### 7.4.5.2 Results

Table 7.15 gives the results of the classification of the benthic data into community structure groups using linear and quadratic discriminant analysis. From Table 7.15 it can be noted that the performance of the discriminant classifiers is not significantly different from those of the individual neural network classifiers reported in Table 7.11, as the results for the discriminant analysis (Table 7.15) are within a standard deviation of the majority of the results of the neural networks (Table 7.11). The quadratic model produced a better classification rate than the linear model for two data sets (both of which were standardised data). In Reynoldson et al. [136] a linear discriminant model was used, with the reported classification rate over the 5 validation runs being 90%. It can be seen that this was an over-estimate, as the more rigorous validation procedure used in this study generated a lower classification rate.

### 7.4.6 Discussion

The results presented in this section indicate that the classification of the community structure from environmental variables can be performed by a simple MLP model, to a reasonable level of performance. The performance was insensitive to the number of hidden units used in the networks, while the weight decay regulariser was only slightly beneficial. The biggest factor determining the classification rate was the training data used for the analysis. Both of the training sets derived from a principal component analysis produced poorer classifiers than the original unfactored data. The dimensionality reduction of the principal component analysis seems to remove some of the information necessary to discriminate between the sites. The smaller subset of environmental variables proved to be the most reliable set of predictor variables.

Comparison of the MLP with the linear and quadratic discriminant models demonstrated that the results between the parametric models and the individual neural network classifiers were not significantly different. However, using a combination of back-propagation classifiers a better classification rate was achieved. But it should be noted that a greater effort was expended on the neural net models than for the discriminant ones.

## 7.5 Classification of Toxicity Groups

### 7.5.1 Preliminary Experiments

#### 7.5.1.1 Procedure

A similar experimental procedure was adopted for the classification of bioassay test groups (see Section 7.3.4), as that used in Section 7.4 for the classification of biological groups. Only one training set E13std, see Section 7.3.2, was tested since preliminary experiments showed little difference in performance between the four sets used in Section 7.4. The network topology was 13 input nodes and 3 output nodes, one for each of the 3 groups identified during the ordination phase (Section 7.3.4). Once again, networks comprising of 5, 7 and 9 hidden units were tested over a series of 10 random starts, using four different weight decay (regularisation) terms.

| Weight decay $\lambda$ and number of hidden units | Training data E13std | |
|---|---|---|
| | Mean | s.d. |
| $\lambda$=0.0 | | |
| h=5 | 69.6% | 2.9 |
| h=7 | 69.8% | 1.9 |
| h=9 | 69.3% | 2.5 |
| $\lambda$=0.0001 | | |
| h=5 | 68.6% | 2.8 |
| h=7 | 68.8% | 2.3 |
| h=9 | 69.4% | 3.0 |
| $\lambda$=0.001 | | |
| h=5 | 68.3% | 2.3 |
| h=7 | 67.20% | 2.2 |
| h=9 | 68.0% | 3.5 |
| $\lambda$=0.01 | | |
| h=5 | 68.6% | 1.7 |
| h=7 | 68.8% | 2.4 |
| h=9 | 69.4% | 2.5 |

Table 7.16: Classification of 93 sites to bioassay group using the environmental variables as predictors.

#### 7.5.1.2 Results

Table 7.16 shows the results of the prediction of bioassay group using the environmental variable data set E13std as predictors. The performance is notably worse than the preceding experiments, with an error rate of over 30% percent compared to the 15% previously obtained in the biological group experiments. This can be explained by the poorer separation of groups 1 and 2, as noted in the ordination, and that the group which represented slightly toxic conditions accounted for less than 5% of the data. Relative performance differences between the number of hidden units were negligible, while the networks which had a zero weight decay term had a slightly better performance.

## 7.5.2 Committees of Networks

### 7.5.2.1 Procedure

The results from the 10 trained networks with 5 hidden units and a zero weight decay term were used in the committee experiments. Again two different methods of combining the results were tried, the product and simple averaging procedures used previously (Section 5.4). The ten networks were ranked and combined in the following manner. Firstly all 10 were combined, with the result for this is given in the row labelled '10' and the column headed by the relevant combining method, Table 7.17, after this the best 9 were combined, with the results being in row 9, and so on.

### 7.5.2.2 Results

It is apparent from Table 7.17 that the performance of the classifier is improved by combining the results from a series of networks. The product and averaging procedures produced equivalent results with the best classifier obtaining an error rate of 21.5%, (classification rate of 78.5%). This compares well to the MDA analysis carried out in Reynoldson et al. [136], where the error rate in the validation experiments was 32%. (The validation experiments carried out in this report are based on leave-one-out cross validation, which is more rigorous than the validation procedure adopted in Reynoldson et al. [136]).

From Table 7.18 it can be observed that Group 3 samples were all incorrectly classified. The most probable reason for this is the low relative frequency of occurrence of these sites within the data set. The classification rates of Group 1 and Group 2 sites, 84.6% and 78.4% respectively, compare favourably to Reynoldson et al. [136], 76% and 55%.

## 7.5.3 Discriminant Analysis for Toxicity Tests

### 7.5.3.1 Procedure

This section reports on the classification using linear and quadratic discriminant models of the toxicity test groups identified from the ordination of Section 7.3.4. The four data sets: E27std, E27pca7, E13std and E13pca4 (see Table 7.2) were used as input to both linear and discriminant models, with a

| Model Rank | Classification Rate | Product Average | Simple Average |
|---|---|---|---|
| 1 | 73.1% | — | — |
| 2 | 72.0% | 78.5% | 78.5% |
| 3 | 72.0% | 74.2% | 74.2% |
| 4 | 72.0% | 74.2% | 74.2% |
| 5 | 69.9% | 74.2% | 75.3% |
| 6 | 68.8% | 74.2% | 73.1% |
| 7 | 68.8% | 74.2% | 74.2% |
| 8 | 67.7% | 73.1% | 73.1% |
| 9 | 67.7% | 73.1% | 72.0% |
| 10 | 63.4% | 73.1% | 73.1% |

Table 7.17: Ranking and classification of networks from experiment reported in Table 7.16, $\lambda = 0.0$ and h = 5, and classification rate based upon product and averaging procedures.

| Expected Groups | Predicted Groups | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 44 *(86.6%)* | 0 | 0 |
| 2 | 8 | 29 *(78.4%)* | 0 |
| 3 | 2 | 2 | 0 *(0.0%)* |

Table 7.18: Confusion table showing predicted toxicity groups and the within group classification rate. Classifier was based on the average of 2 networks.

| Method | Training data | | | |
|---|---|---|---|---|
| | E27std | E27pca | E13std | E13pca4 |
| Linear | 62.3% | 62.5% | 64.5 % | 66.7% |
| Quadratic | 61.3% | 63.4% | 64.5% | 69.9% |

Table 7.19: Classification rate for toxicity test using discriminant analysis.

leave-one-out cross validation technique being implemented.

#### 7.5.3.2 Results

Table 7.19 shows the classification rate for the bioassay groups for the linear and quadratic discriminant models. The difference between the linear and quadratic models was minimal (i.e. one misclassified site for the data sets E27std, E27pca7 and E13std), with the quadratic model using the lowest dimensional input data, E13pca4, producing the highest classification rate. All the neural networks models of Section 7.5.1 had higher classification rates than the linear models, while their performance was comparable to the best performing quadratic discriminant model. The committees of MLPs produced noticeably higher classification rates.

### 7.5.4 Discussion

The neural network models applied in this section perform well when compared to the MDA systems used by Reynoldson et al. [136]. However, the input variables used in this study were different to those used by Reynoldson et al. [136], in that a smaller standardised set of 13 variables were used as opposed to a larger set of 19 variables. The discrimination between groups was better overall, but Group 3 caused problems due to the low frequency in the data set. The sites that composed Group 3 have a low survival of *Hyalella azteca* and low growth rate of *H. azteca* and *Hexagenia spp.*, and may be construed as slightly toxic or sensitive to a particular combination of environmental variables, however the classifier should be able to discriminate between these sites and the remaining data, which the neural classifiers are not doing at present.

The implementation of guidelines would be based on deviations in the endpoint data from the reference database, and it may be that ordination of the bioassay data is an unnecessary step. At present the difference between the endpoints of the groups representing the uncontaminated sites is small, and it may be better just to define a good bioassay result, indicating uncontaminated conditions, on the averages of the clean site database, while specifying toxicity as some measure of deviation from the good assay results. Also, as the bioassay endpoints provide more direct evidence of possible toxicity problems it appears that the convoluted route of analysis through ordination is unnecessary, as all the sites, bar those at present in Group 3, constitute one large group representing an uncontaminated condition. The derivation of toxicity groups based on the physio-chemical variables, at this stage, seems inappropriate as the link between toxicity and environmental variables is less well defined than that of community structure and environmental data. However, when the full reference database is analysed the differences in bioassay response between 'clean' sites may become more apparent, and the ordination stage would then be necessary.

### 7.5.5 Conclusion

A back-propagation based network was applied to the classification of toxicity groups using the environmental variables for predictors. A classifier based on a committee of two networks was found to provide the best performance, and regularisation, in the form of a weight decay term, was found unnecessary. The performance of the neural net classifier compared well (a decrease in error rate of 5%) with the MDA systems previously used (Section 7.5.3 & Reynoldson et al. [136]), but problems relating to the classification of unusual sites were noted and two possible methods of overcoming these were suggested.

## 7.6 Prediction of Ordination Vectors

### 7.6.1 Motivation

The motivation behind this experiment is to predict the position of the sites in the community structure ordination space. Although this is not necessary for the determination of group membership, it may provide alternative means for the classification and prediction of expected fauna, without having to average community structure over the whole group. That is, it provides a different starting point in ordination space from which to predict the expected community structure at a site. In essence it may provide an extra degree of confidence in the classification process.

### 7.6.2 Preliminary Experiments

#### 7.6.2.1 Procedure

For the preliminary tests the output set of the three-dimensional ordination vectors were split into three scalar values, with different networks being used for each scalar value, thus one network predicted the Vector 1 scores, another Vector 2 and another Vector 3. It would be possible to use one network for all three scores, but it was decided to train separate networks for each vector component to avoid over-parameterising the models. The data sets referred to in Table 7.2 were used as input to the network. The number of hidden units were 5, 7 and 9 as before. As the problem is one of prediction, *tanh* transfer functions were used for the hidden units, while the output activation function was linear. Four values of weight decay parameter were considered: 0.0, 0.0001, 0.001 and 0.1. The performance of each network was averaged over 10 random starts, with the weights being randomly initialised between values of -0.1 and 0.1.

#### 7.6.2.2 Results

*First Ordination Vector*

The results of all the trials are given in Table 7.20. This shows the correlation coefficient between the values predicted by the network and the target values

from the ordination scores. To summarise, the E13pca4 input set produced the best results, based on the correlation coefficient of the four input sets considered. Two of the weight decay terms (0.0001 and 0.001) seemed to inhibit performance when compared to zero weight decay, but the 0.01 weight decay produced comparable, if not improved, results relative to zero weight decay. The number of hidden units used had little effect on the results.

*Second Ordination Vector*

Again the best performance was obtained by using the E13pca4 data set, Table 7.21. The networks with zero weight-decay performed marginally better than before, but again the number of hidden units used made little difference to the performance.

*Third Ordination Vector*

Overall performance was poorer in this case than the other two, the best performance being from the E13std data, closely followed by the E27pca7 and the E13pca4 data sets (Table 7.22). The weight-decay terms 0.0001 and 0.001 again produced networks that had poorer results than the others. The networks with zero weight-decay produced the best overall performance. It is worth noting that the E27std data set produced the worst results in all cases.

#### 7.6.2.3 Discussion

It is apparent from the results that the performances of the predictions of Vectors 1, 2 and 3 decreased in order from Vector 1 to Vector 3. This was only to be to be expected when one considers the mechanism of the ordination algorithm. The first vector of the ordination contains most of the discriminatory power and generally has the best correlation with the predictor variables. The other vectors tend to have decreasing correlation with the predictor variables, and there is less discrimination between the sites. There appeared to be some benefit from using the principal component data (E13pca4); in essence it appears that for the regression analyses the smaller the dimension of the input data the better the resulting correlations of the model's predictions.

| Weight Decay $\lambda$ and number of hidden nodes | Training data set (see Section 7.3.2) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | E27std | | E27pca7 | | E13std | | E13pca4 | |
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| $\lambda = 0.0$ | | | | | | | | |
| h=5 | 0.762 | 0.026 | 0.780 | 0.028 | 0.795 | 0.062 | 0.856 | 0.012 |
| h=7 | 0.763 | 0.047 | 0.744 | 0.057 | 0.830 | 0.028 | 0.838 | 0.005 |
| h=9 | 0.775 | 0.024 | 0.741 | 0.058 | 0.851 | 0.012 | 0.828 | 0.012 |
| $\lambda = 0.0001$ | | | | | | | | |
| h=5 | 0.713 | 0.032 | 0.772 | 0.038 | 0.804 | 0.078 | 0.836 | 0.026 |
| h=7 | 0.721 | 0.055 | 0.720 | 0.050 | 0.776 | 0.063 | 0.813 | 0.025 |
| h=9 | 0.745 | 0.037 | 0.691 | 0.043 | 0.792 | 0.062 | 0.790 | 0.024 |
| $\lambda = 0.001$ | | | | | | | | |
| h=5 | 0.714 | 0.029 | 0.760 | 0.039 | 0.807 | 0.060 | 0.836 | 0.018 |
| h=7 | 0.719 | 0.060 | 0.705 | 0.050 | 0.796 | 0.066 | 0.815 | 0.022 |
| h=9 | 0.746 | 0.033 | 0.698 | 0.027 | 0.804 | 0.040 | 0.799 | 0.021 |
| $\lambda = 0.01$ | | | | | | | | |
| h=5 | 0.751 | 0.048 | 0.756 | 0.035 | 0.825 | 0.032 | 0.860 | 0.012 |
| h=7 | 0.735 | 0.048 | 0.764 | 0.046 | 0.828 | 0.029 | 0.829 | 0.021 |
| h=9 | 0.757 | 0.032 | 0.712 | 0.069 | 0.837 | 0.026 | 0.809 | 0.026 |

Table 7.20: Prediction of first vector ordination scores using environmental variables as predictors. Pearson $r$ correlation coefficient and standard deviation.

| Weight Decay $\lambda$ and number of hidden nodes | Training data set (see Section 7.3.2) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | E27std | | E27pca7 | | E13std | | E13pca4 | |
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| $\lambda = 0.0$ | | | | | | | | |
| $\lambda = 0.0$ | | | | | | | | |
| h=5 | 0.444 | 0.074 | 0.511 | 0.057 | 0.588 | 0.047 | 0.699 | 0.049 |
| h=7 | 0.399 | 0.088 | 0.558 | 0.045 | 0.610 | 0.041 | 0.688 | 0.045 |
| h=9 | 0.443 | 0.060 | 0.578 | 0.057 | 0.627 | 0.046 | 0.689 | 0.037 |
| $\lambda = 0.0001$ | | | | | | | | |
| h=5 | 0.430 | 0.087 | 0.506 | 0.067 | 0.560 | 0.068 | 0.652 | 0.044 |
| h=7 | 0.502 | 0.066 | 0.524 | 0.075 | 0.567 | 0.052 | 0.621 | 0.046 |
| h=9 | 0.409 | 0.074 | 0.525 | 0.048 | 0.575 | 0.065 | 0.624 | 0.044 |
| $\lambda = 0.001$ | | | | | | | | |
| h=5 | 0.431 | 0.100 | 0.526 | 0.054 | 0.567 | 0.059 | 0.657 | 0.056 |
| h=7 | 0.480 | 0.056 | 0.550 | 0.061 | 0.546 | 0.088 | 0.649 | 0.061 |
| h=9 | 0.431 | 0.025 | 0.475 | 0.162 | 0.579 | 0.051 | 0.608 | 0.042 |
| $\lambda = 0.01$ | | | | | | | | |
| h=5 | 0.474 | 0.073 | 0.474 | 0.073 | 0.585 | 0.056 | 0.675 | 0.042 |
| h=7 | 0.426 | 0.039 | 0.531 | 0.037 | 0.567 | 0.078 | 0.654 | 0.047 |
| h=9 | 0.384 | 0.056 | 0.544 | 0.086 | 0.585 | 0.066 | 0.627 | 0.032 |

Table 7.21: Prediction of second vector ordination scores using environmental variables as predictors. Pearson $r$ correlation coefficient and standard deviation.

| Weight Decay $\lambda$ and number of hidden nodes | Training data set (see Section 7.3.2) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | E27std | | E27pca7 | | E13std | | E13pca4 | |
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| $\lambda = 0.0$ | | | | | | | | |
| $\lambda = 0.0$ | | | | | | | | |
| h=5 | 0.451 | 0.071 | 0.605 | 0.056 | 0.607 | 0.055 | 0.572 | 0.029 |
| h=7 | 0.479 | 0.043 | 0.599 | 0.063 | 0.581 | 0.070 | 0.568 | 0.037 |
| h=9 | 0.501 | 0.054 | 0.576 | 0.053 | 0.582 | 0.057 | 0.542 | 0.045 |
| $\lambda = 0.0001$ | | | | | | | | |
| h=5 | 0.447 | 0.054 | 0.520 | 0.073 | 0.561 | 0.084 | 0.528 | 0.063 |
| h=7 | 0.448 | 0.040 | 0.541 | 0.065 | 0.530 | 0.066 | 0.483 | 0.062 |
| h=9 | 0.475 | 0.062 | 0.463 | 0.063 | 0.521 | 0.088 | 0.465 | 0.046 |
| $\lambda = 0.001$ | | | | | | | | |
| h=5 | 0.451 | 0.077 | 0.517 | 0.062 | 0.546 | 0.073 | 0.511 | 0.071 |
| h=7 | 0.450 | 0.049 | 0.545 | 0.059 | 0.537 | 0.063 | 0.516 | 0.049 |
| h=9 | 0.464 | .0510 | 0.500 | 0.047 | 0.501 | 0.057 | 0.508 | 0.051 |
| $\lambda = 0.01$ | | | | | | | | |
| h=5 | 0.459 | 0.068 | 0.564 | 0.053 | 0.571 | 0.069 | 0.554 | 0.050 |
| h=7 | 0.451 | 0.056 | 0.569 | 0.063 | 0.581 | 0.058 | 0.561 | 0.045 |
| h=9 | 0.473 | 0.055 | 0.515 | 0.056 | 0.568 | 0.062 | 0.568 | 0.056 |

Table 7.22: Prediction of third vector ordination scores using environmental variables as predictors. Pearson $r$ correlation coefficient and standard deviation.

### 7.6.3 Committees of Networks

#### 7.6.3.1 Procedure

Using the results of the proceeding section the network with 5 hidden nodes, weight decay $\lambda = 0.1$ and training data E13pca4 was selected as the basis of the experiments in this section. The averaging procedure was the only method considered for the formation of the committees, as the product averaging method is not applicable to regression analyses. The 10 networks were ranked and the committees formed using the best 2, then the best 3 and so forth.

#### 7.6.3.2 Results

Table 7.23 shows the ranked performance of the 10 networks along with the committees' performance. The optimum combination was obtained by using the best 2 networks, and the results from this network are displayed graphically in Figure 7.7. The correlation coefficients between the committee's performance and the target ordination vectors scores are given in Table 7.24, broken down by individual groups.

Considering the complete data set it can be observed that the networks were best at predicting Vector 1, then Vector 2 and finally Vector 3. This is the same as the findings in Section 7.6.2.2. It seems that the number of samples within the cluster affected the correlation coefficient, as Groups 1, 3 and 5, those with the greatest frequency, produced significant correlations Table 7.24. The graphs in Figure 7.7 show the position of the predicted vectors in ordination space against the target vectors, and clearly show good clustering of the Group 5 (L. Michigan) sites. If the classes are viewed individually then it is apparent that the network achieves good discrimination between clusters, but a poorer performance occurs on the intra-cluster scale.

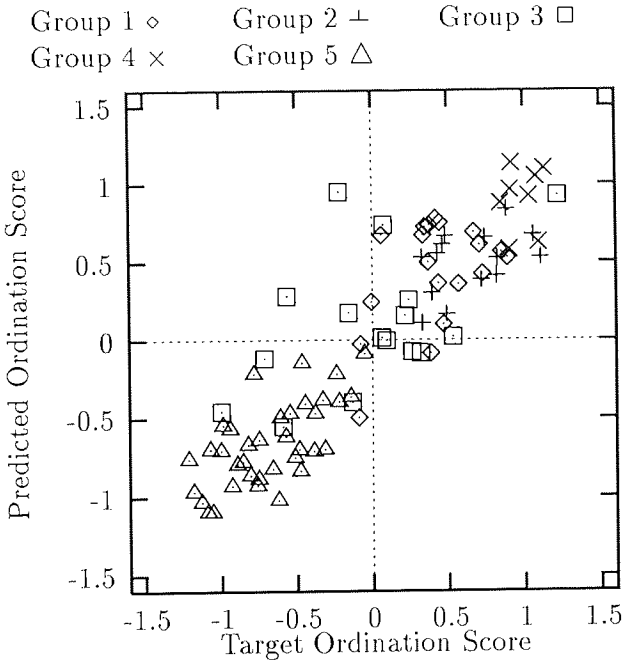
| Rank of Networks | Vector 1 | | Vector 2 | | Vector 3 | |
|---|---|---|---|---|---|---|
| | Corr. | Average | Corr. | Average | Corr. | Average |
| 1 | 0.880 | — | 0.717 | — | 0.626 | — |
| 2 | 0.877 | 0.885 | 0.717 | 0.757 | 0.606 | 0.656 |
| 3 | 0.865 | 0.882 | 0.701 | 0.764 | 0.583 | 0.651 |
| 4 | 0.863 | 0.880 | 0.700 | 0.764 | 0.564 | 0.654 |
| 5 | 0.860 | 0.879 | 0.700 | 0.761 | 0.561 | 0.653 |
| 6 | 0.859 | 0.878 | 0.678 | 0.754 | 0.557 | 0.648 |
| 7 | 0.858 | 0.877 | 0.658 | 0.752 | 0.548 | 0.647 |
| 8 | 0.851 | 0.877 | 0.654 | 0.751 | 0.546 | 0.645 |
| 9 | 0.849 | 0.876 | 0.637 | 0.745 | 0.491 | 0.638 |
| 10 | 0.841 | 0.875 | 0.583 | 0.739 | 0.455 | 0.630 |

Table 7.23: Ranking and classification of networks from experiment reported in Table 7.17, $\lambda = 0.1$ and $h = 5$, and correlation coefficient $r$ based upon averaging.

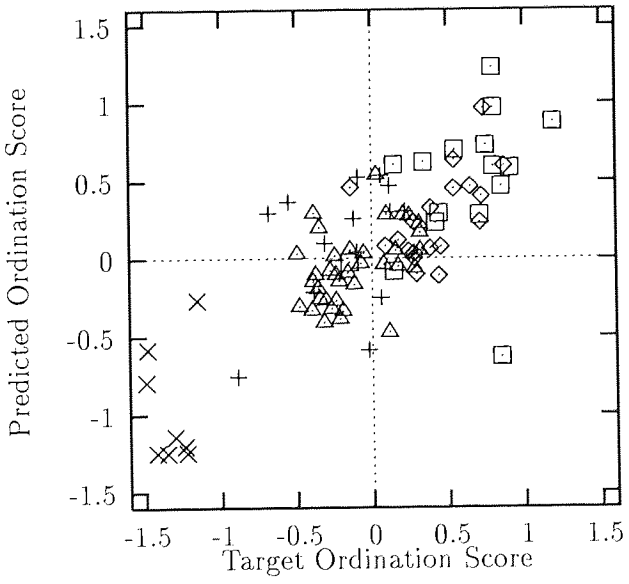
| | Gp 1(19) | Gp 2(14) | Gp 3(16) | Gp 4(8) | Gp 5(36) | All Gp's(93) |
|---|---|---|---|---|---|---|
| Vector 1 | *0.472* | 0.454 | *0.504* | 0.150 | *0.651* | *0.855* |
| Vector 2 | *0.508* | 0.386 | 0.371 | -0.135 | *0.450* | *0.757* |
| Vector 3 | 0.218 | 0.390 | 0.243 | -0.098 | -0.070 | *0.656* |

Table 7.24: Correlation coefficient $r$ between actual and predicted ordination vectors for each biological group. (Numbers in *italics* P>0.05).

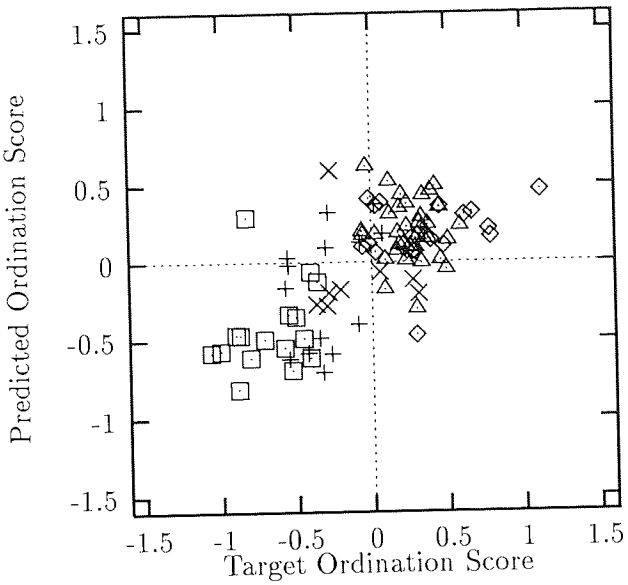
Figure 7.7: Regression of neural network predictions against vector scores from ordination of community structure.




(a) Vector 1 Scores




(b) Vector 2 Scores




(c) Vector 3 Scores

### 7.6.4 Discussion

The use of separate networks to predict each component of the position vector enabled the location of a site in ordination space to be predicted with some confidence. It was found that the best performance was achieved using a data set based on a principal component analysis, and that the weight decay parameter was of little benefit when used with these networks of low input dimensionality. The groups which occurred most frequently in the data set were predicted with greater accuracy than the less frequent groups.

The method may prove useful in identifying unusual environmental conditions, as the predicted position of a novel data set would be unlikely to fall within the bounds of the ordination groups, however this may be difficult to express in quantitative terms.

| Family | Freq. | Composition |
|---|---|---|
| Sphaeriidae | 86.0% | 14 species |
| Chironomidae | 84.9% | 42 genera |
| Naididae | 65.6% | 15 species |
| Tubificidae | 82.8% | 19 species |
| Haustoriidae | 51.6% | 1 species |

Table 7.25: Summary of the frequency of occurrence and composition of the five most common families.

## 7.7 Prediction of the Abundance of Taxa

### 7.7.1 Procedure

As described in Ruck et al. [149] it is possible to predict the abundances of key elements of the benthic community structure directly from the environmental variables. This section describes a series of experiments using the database of 93 samples referred to in Section 7.3. Using an earlier database of just 53 samples, Ruck et al. [149] found that the technique's ability to predict abundance varied from taxon to taxon, with some being predicted more reliably than others. However, some success was evident with a good correlation between predicted and target abundances being achieved.

The five families (excluding the Porifera) with the highest frequency of occurrence were identified and used as the basis for the experiments. The five families were Sphaeriidae, Chironomidae, Naididae, Tubificidae and Haustoriidae. Table 7.25 shows the percentage of samples in which each of the five families were present, and also the composition of the families in terms of the taxa listed in Appendix A3. For each family the number of individuals in each sample was used as the target value for the network, with the environmental data as input. The abundance of the family was calculated by simply summing all the occurrences of the taxa that make up the family.

Again, after preliminary tests, one data set was chosen for the experimental work. This was the subset of 13 standardised environmental variables, E13std (see Table 7.2) as in previous cases. For each of the five families ten networks were trained using 5, 7 and 9 hidden units. Only two values of the weight-decay

term, $\lambda$, were tested, these being 0.0 and 0.001.

### 7.7.2 Results

Table 7.26 records the mean, standard deviation and maximum values of the correlation coefficient of the 10 starts. From this is can be seen that the model predicted Haustoriidae most accurately, followed by the Chironomidae, Sphaeriidae, Tubificidae and finally Naididae. To achieve 95% significance, assuming a one-tailed test, a correlation coefficient of greater than 0.171 would be needed, for 99% significance the value rises to 0.241 [81]. These values indicate that the models are achieving statistically significant predictions for Haustoriidae, Chironomidae, Sphaeriidae and Tubificidae. However, the correlations are poor when compared to those achieved by Ruck et al. [149], and are not good enough for the model to be considered useful in a predictive system. Thus, the expansion of the data set has lead to a decrease in the quality of the predictions, which is contrary to what would be expected.

### 7.7.3 Discussion

The reliable prediction of individual invertebrate species directly from the environmental variables would be a valuable aid, especially when target communities are compared to observed communities. However, there are problems with this approach, and these are essentially to do with the species being treated as individual units (i.e. taking no account of the other species that make up the community), and the fact that the same set of environmental variables can support (slightly) different communities, due to the natural fluctuations associated with community structure. Taking a species as independent from the others does not necessarily lead to poor results, but in practice this is probably not the best policy. The second problem is the property of a 'normal' relationship between community structure and its environment, but for prediction the problem is compounded by the possible absence of the family. The absent cases tend to hinder the prediction of the abundances. A further possibility is that an important environmental parameter that influenced the abundances may have been absent from the available set of environmental variables.

| **Sphaeriidae** | mean | s.d. | max |
|---|---|---|---|
| $\lambda$=0.0 | | | |
| h=5 | 0.197 | *0.104* | 0.398 |
| h=7 | 0.258 | *0.089* | 0.450 |
| h=9 | 0.217 | *0.095* | 0.327 |
| $\lambda$ = 0.001 | | | |
| h=5 | 0.198 | *0.116* | 0.405 |
| h=7 | 0.104 | *0.131* | 0.438 |
| h=9 | 0.235 | *0.052* | 0.340 |

| **Chironomidae** | mean | s.d. | max |
|---|---|---|---|
| $\lambda$=0.0 | | | |
| h=5 | 0.431 | *0.062* | 0.568 |
| h=7 | 0.440 | *0.030* | 0.481 |
| h=9 | 0.499 | *0.073* | 0.578 |
| $\lambda$ = 0.001 | | | |
| h=5 | 0.405 | *0.058* | 0.470 |
| h=7 | 0.421 | *0.069* | 0.521 |
| h=9 | 0.392 | *0.063* | 0.459 |

| **Naididae** | mean | s.d. | max |
|---|---|---|---|
| $\lambda$ = 0.0 | | | |
| h=5 | 0.022 | *0.052* | 0.093 |
| h=7 | 0.048 | *0.026* | 0.082 |
| h=9 | 0.064 | *0.040* | 0.155 |
| $\lambda$ = 0.001 | | | |
| h=5 | 0.033 | *0.041* | 0.091 |
| h=7 | 0.048 | *0.045* | 0.121 |
| h=9 | 0.081 | *0.054* | 0.196 |

| **Tubificidae** | mean | s.d. | max |
|---|---|---|---|
| $\lambda$ = 0.0 | | | |
| h=5 | 0.181 | *0.123* | 0.371 |
| h=7 | 0.177 | *0.142* | 0.374 |
| h=9 | 0.235 | *0.093* | 0.390 |
| $\lambda$ = 0.001 | | | |
| h=5 | 0.182 | *0.138* | 0.420 |
| h=7 | 0.257 | *0.111* | 0.412 |
| h=9 | 0.248 | *0.107* | 0.411 |

| **Haustoriidae** | mean | s.d. | max |
|---|---|---|---|
| $\lambda$ = 0.0 | | | |
| h=5 | 0.669 | *0.035* | 0.705 |
| h=7 | 0.682 | *0.052* | 0.798 |
| h=9 | 0.644 | *0.057* | 0.719 |
| $\lambda$ = 0.001 | | | |
| h=5 | 0.651 | *0.050* | 0.715 |
| h=7 | 0.625 | *0.050* | 0.720 |
| h=9 | 0.624 | *0.048* | 0.705 |

Table 7.26: Mean, standard deviation and maximum values of the correlation coefficient of model's predictions against target abundances, averaged over 10 networks.

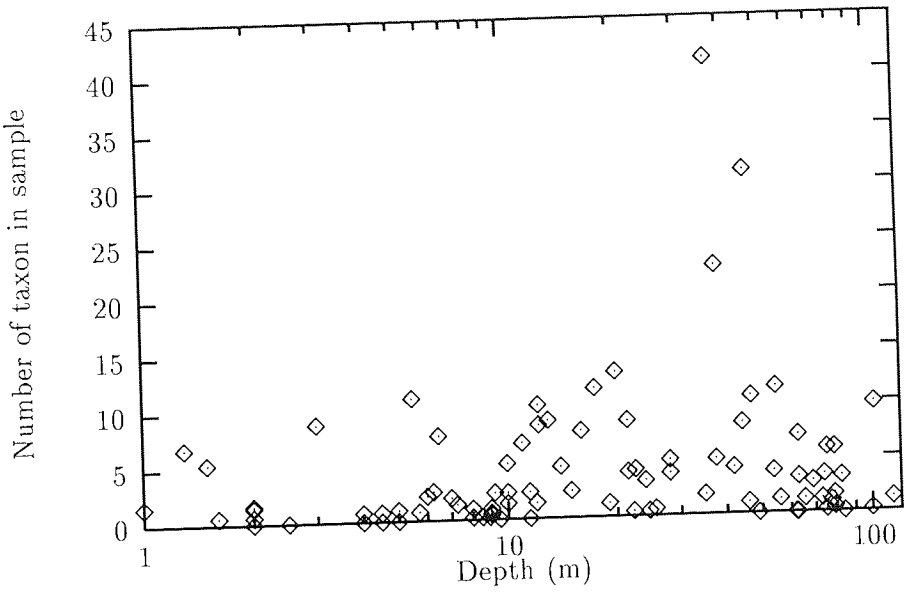



Figure 7.8: Plot of Sphaeriidae abundance against sample depth.

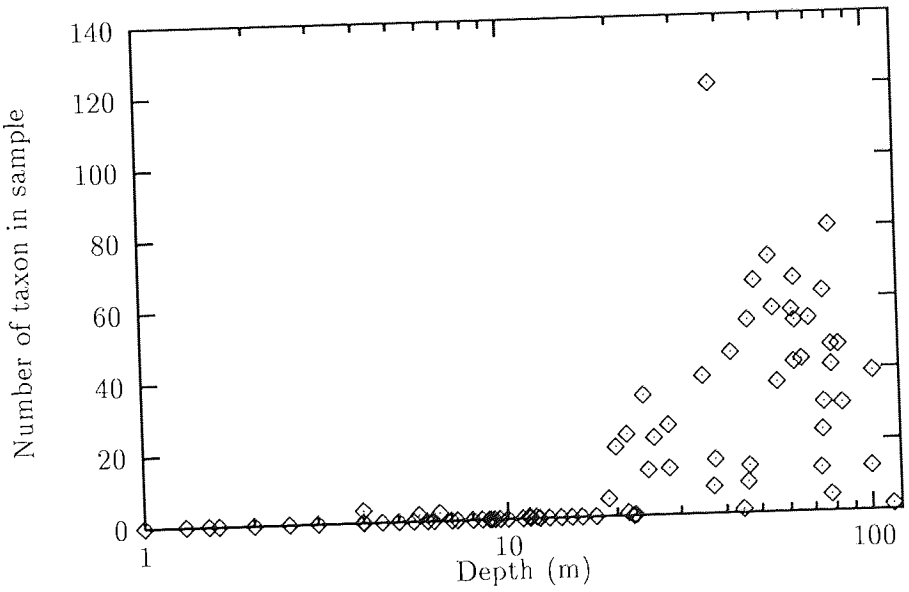



Figure 7.9: Plot of Haustoriidae abundance against sample depth.

Considering the Sphaeriidae and Haustoriidae, Figures 7.8 and 7.9 show how the abundance of the each family varies with depth. The Sphaeriidae graph shows that there is a weak relationship between the abundance of the family and depth, while the Haustoriidae show a distinct change around 18-19m depth. Although only one variable, depth, is considered extrapolation to the extra dimensions of the input data is likely to lead to an even more complex relationship between the abundances and the predictor variables. In view of the Haustoriidae plot, it may have been better to have used a two stage process: first predict simple presence or absence, then predict the abundance of the 'present' taxa. For example, from Figure 7.9 it is possible to say with a good deal of confidence that Haustoriidae will be present in waters with a depth greater than 18m. This would help to eliminate the interference of the absent cases when predicting the abundance.

### 7.7.4 Conclusion

The abundances of 5 families of taxa were predicted using a series of networks. The best predictions were achieved with the Haustoriidae, which gave the highest correlation for any single model of 0.798. However, the overall performance of the models was disappointing. This was due to the complexity of the data, and the nature of family abundances with respect to the environmental variables.

## 7.8 Summary

Despite the small size of the data set, the results of this chapter demonstrate that back-propagation neural networks have considerable potential for use as classifiers in environmental monitoring. Their ability to classify has be shown to compare well with that achieved by the discriminant analyses carried out in this dissertation and in a previous study by Reynoldson et al. [136]. Indeed, the networks slightly out-perform discriminant analysis, but owing to the limited size of the data set this result cannot be taken as conclusive.

On the specific tasks to which they were applied, the networks performed best when classifying the benthic community group to be expected at a site

from a knowledge of the site's environmental variables. The few misclassifications which did occur (i.e. approximately 14%) mainly corresponded to sites having conflicting locations in two ordination spaces, namely the community structure and environmental ordination spaces. Thus they related to what might be considered difficult or problem cases.

The classification of expected bioassay groups from environmental variables was less successful than the classification of community structure. This was mainly due to two factors: the relatively small separation in ordination space between the 'clean' sites (i.e. the vast majority) and sites exhibiting slight toxicity; and the low frequency of the slightly toxic sites.

The prediction of a site's location in the ordination space of community structure was attempted and found to produce good correlations on the first and second dimensions of the ordination vectors. The groups which contained the highest number of samples were the most reliable to predict, with the performance tailing off with group size. In all cases, the use of committees of networks was found to result in improved performance over that of individual networks, even the best individual networks.

# Chapter 8

# Discussion and Conclusion

## 8.1 Introduction

This chapter has three principal objectives. The first is to reiterate the principal contributions made by this dissertation and to summarise the preceding chapters. The second is to discuss the experimental work in a wider context; commenting on the practical application and implementation of the various models described, and the possible directions for future research. The third is to conclude the dissertation.

## 8.2 Discussion

### 8.2.1 Contributions and Summary

The true potential of freshwater biomonitoring has not yet been fully realised. This has lead to a reliance on chemically based methods for most monitoring purposes. The original motivation of this research project was to redress this imbalance between biological and chemical monitoring by demonstrating the potential of applying AI techniques, particularly artificial neural networks, to freshwater biomonitoring. It has accomplished this goal and more, via its principal contributions:

*i.* A thorough and principled investigation of the application of neural networks to the direct interpretation of freshwater benthic invertebrate communities.

*ii.* An extension of the term 'indicator taxa' to encompass computer modelling, and the quantification of information loss associated with different levels of identification and enumeration of invertebrate taxa.

*iii.* A demonstration of the utility of neural networks within a community structure based approach to the assessment of sediment toxicity in the Great Lakes.

After reviewing the present methods of freshwater biomonitoring and the pertinent areas of neural network research, a full description and analysis of the available river data was given (Chapter 4). The development of a new biological classification system (B1a, B1b, ..., B4) for water quality, based on the present NWC classification, was introduced. As the classes have a biological basis, there is a clear monotonic relationship between organic pollution and quality classes. The classification can be applied to different biotopes but has presently only been developed and tested using riffles. There is no conceptual difficulty extending it to other biotopes. A possible criticism of the biological classification is that it is based on the subjective assessment of a single expert, but this also applies to virtually all of the biotic indices presently used, where an individual or, more typically, a panel of experts have agreed upon some particular measure or ranking. Any system designed to interpret a biological system will be influenced by some degree of subjectivity. There is a maxim to the effect that the 'best available technology' should be used, and one conclusion of this project is that the Expert (and no doubt any other experienced river ecologist) was much more reliable than any of the other available classification systems, especially the biotic indices.

A data set consisting of 292 samples from the Upper Trent catchment of the NRA Severn-Trent Region was constructed from a database of samples taken for routine monitoring purposes. This data was classified to biological class by the Expert, and was used to demonstrate the relationship between the biological classes and BMWP score, ASPT and TBI indices. This data was used extensively for the neural network experimentation. The generation of a large database of synthetic data, based on elicited domain knowledge was described, and was subsequently used for the neural network experimentation. Using the NRA national database it was shown that the variation of BMWP score and

ASPT between the regions was large and that the spatial distributions of seven common invertebrate families varied substantially. These facts highlight the difficulties of designing a uniform national monitoring programme.

In Chapter 5 the major neural network results were presented. After considering the experimental methodology and some preliminary work, an investigation into the direct interpretation of a subset of taxa from the invertebrate community into biological water quality class was investigated. It was demonstrated that it was possible to classify this subset of invertebrate taxa into the biological classification with a fairly high degree of accuracy. A classification rate of over 80% for test data was achieved, with the majority of misclassification occurring between the high variance B1b and B2 classes.

A method for identifying unusual samples was presented. Taking synthetic riffle data as typical it was possible to identify samples that were drawn from a different distribution, pool biotopes in this case. The method should also work for other scenarios, for example heavy metal pollution or acidification. A more complex neural network model, based on a mixture of experts, was adopted as a means of classifying data which drawn from distinct distributions, in this case different biotopes. Prior knowledge, in the form of synthetic samples, was used to 'prime' a MLP network, but this was found to hinder the model's performance. An unsupervised learning method, namely Kohonen's self-organised maps, was investigated, and it was demonstrated that the biological classification offered a good explanation of the resulting map.

In Chapter 6 the identification of indicator taxa was considered. After redefining the term 'indicator taxa', a quantitative analysis was undertaken using the Severn-Trent data. The development of three methods for selecting good indicators of quality class was described. It was demonstrated that there is an information gain going from family (absent/present) to family (four abundance levels) to species (four abundance levels). An important result was the strong correlation between the Expert's list of indicator taxa (from the BERT system) and the corresponding lists of taxa derived from the two selection methods. It was shown that the selection of inputs, both with regard to number and coding, is an important factor affecting model performance.

In Chapter 7 the dissertation concentrated on the use of neural networks for classification and prediction of community structure for use in the assess-

ment of sediment toxicity in the Great Lakes. The MLP models out performed the more traditional discriminant analysis for the classification of community structure and bioassay groups. The community structure misclassifications could be explained by a conflict in the ordination spaces of the community structure and the environmental variables. The prediction of a taxon's abundance was found to be unreliable.

### 8.2.2 Practical Application and Implementation

Two frequently encountered words associated with neural network applications are 'validation' and 'verification'. These words refer to the expected performance and robustness of (*i*) the model and (*ii*) the complete system in which the model is embedded. Verification is concerned only with the model, while validation involves the whole system, but these terms are not used with any consistency at the present time. The validation of the whole system is important, but is not of such importance in this study as, for example, in an area like safety critical control.

Verification would typically entail calculating confidence intervals on a model prediction and/or the flagging of unusual samples (see, for example, Section 5.5). For the applications considered in this dissertation, there would always be some degree of doubt in the modelling process as there is no 'hard' embedded autecological information. Ideally, the best solution would to build a numerical equivalent to the physical model (i.e. model the whole of the benthic system, including chemical, biological and physical relationships), but this is too complex and is the reason why models like neural networks are used in practice.

With the advent of large scale recording of benthic samples on computers, the integration of the models into a monitoring programme would be relatively simple. Once the data is in machine readable form, simple pre-processing is all that is required in order to prepare the data for input to the network, and this is easily achievable with the present database technology. The model's predictions or classifications would provide additional information that could be easily added to the existing sample record and then be manipulated at will.

Another possible use for the models which was considered in this dissertation, apart from classification, was that of smart-pointers to identify samples of concern. For routine monitoring programs it would be possible to flag samples which are novel or unusual (i.e. depart from normality in some respect), and save the ecologist time and effort by avoiding the need to interpret the many samples which show no cause for concern. The ultimate aim is not to replace the expert biologists or ecologists, but to make their work more efficient and effective by targeting their effort on those samples that require expert interpretation.

AI tools could prove useful as teaching aids for trainee river biologists and ecologists. The computer models could be embedded in a hypermedia system, which could be questioned interactively. For example, a benthic sample could be shown and the student would see the effects upon the classification of altering the abundance levels or absence/presence state of certain taxa. As they have little power of explanation, the neural network models would be limited in this role, unlike, for example, knowledge-based systems. The models could be combined with graphics and other information pertaining to the ecology of the taxa, and would constitute a powerful learning environment.

Another area that is of increasing interest is rapid assessment techniques. Typically, the sample is assessed in terms of a simple index which can be readily calculated and interpreted on the river bank. With the increasing use of portable computers the assessment could utilise trimmed down versions of the networks described in this dissertation. The thing that is rapid in 'rapid assessment' is the quantification of the sample information. This does not imply the use of simplistic interpretation techniques or models.

### 8.2.3 Future Research

Within this section some possible directions for neural network research are suggested, and then a more holistic view is taken, which incorporates other methods from AI and elsewhere.

Aside from the direct interpretation of invertebrate samples, three ideas introduced in this dissertation warrant special attention. The first was a detection method for novel (or unusual) samples. A simple method was used in

this dissertation, but this idea could be developed further to take account of, for example, effects due to metal pollution or acidification. The second idea was the use of mixtures of experts. This idea is attractive for a number of reasons. Geography is an important factor affecting the composition of the benthic community, as well as biotope. The use of mixtures of expert models accommodate different regions and/or biotopes would result in a more robust and rational system. Also, interference effects between different regions/biotopes would be minimised. The mixture of experts model [120] has been the one used in this dissertation, but there several other models that could have been implemented [77]. Both of these ideas were successfully tested on a large synthetic set of data, and it seems natural to experiment further on real data, to see if the success is carried over. The third is the mathematical formulation of what constitutes an 'indicator taxon'. The results demonstrated that it is possible to quantify the information loss associated different levels of identification and enumeration. Using this kind of information it would be possible to design monitoring programs so that a reasonable trade-off is reached between the useful information provided by a sample and the effort put into sample sorting and identification.

The area of biological monitoring research that is likely to be of most benefit is the further improvement of methods of interpreting benthic samples, especially with regard to the sourcing of pollution problems. A frequent criticism of biological monitoring is that even though it is apparent that a community is 'stressed',[1] the specific cause remains unclear. This is true to a certain extent, as it is unlikely to be able to categorically name the pollutant, but with an expert interpretation the most probable qualitative cause can generally be identified. From a spatial viewpoint the various communities provide excellent evidence as to the extent and source of the pollution problem. It must be noted that many one-off pollution problems are identified because of a change in colour or odour of the water course, so the potential for biological monitoring contributing to this kind of detection is restricted.

Allied to interpretation is the power of explanation, which, if considered important, the most promising tools to achieve this appear to be casual belief

[1] It should be noted that a community cannot be stressed, it is the individuals within the community that experience stress.

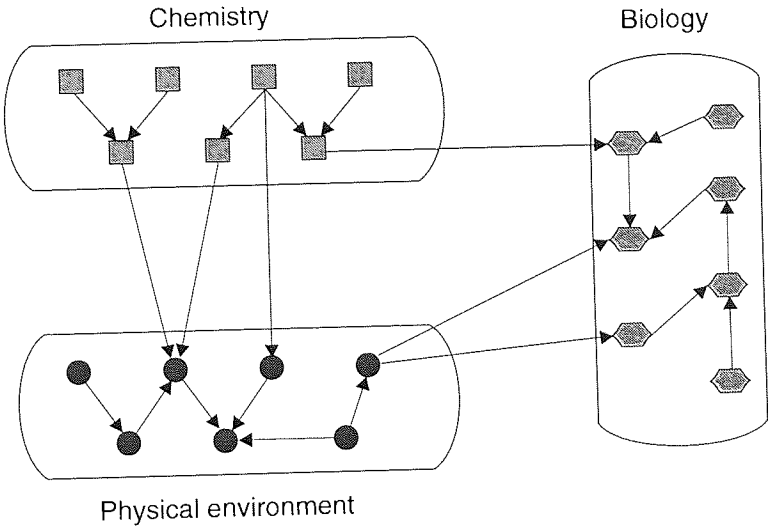



Figure 8.1: Schematic illustration of an example causal belief network for biological monitoring. The lines connecting the various nodes represent casual links. The effect of changing a single variable propagates throughout the system, and information on the whole system is available at any particularly instant.

networks and some of the machine learning techniques [27]. A causal belief network is an extension of probabilistic reasoning systems in which dependencies, or more specifically causal links, are represented in the form a network (Figure 8.1). Their exposition has been detailed by Pearl [126], Neapolitan [118] and Spiegelhalter et al. [160]. As they are a knowledge-based systems an appreciable amount of elicitation (either subjective or objective) must be completed before they can be made functional. Although, it is almost always a productive exercise to elicit probabilities from a domain expert or experts, in large complex systems the expertise may simply not be available, which would be a difficult obstacle to overcome. If successful, this would be a significant advance and would belief networks an advantage over other systems.

A further possibility for future research is the incorporation of the artificial intelligence models into a Geographic Information System (GIS). One such example is the RAISON (Regional Analysis by Intelligent Systems on a Microcomputer) system, which integrates an expert system shell into a GIS, and can predict the effect of pollution emissions on water quality [87]. The ability of a GIS to handle and store data from many sources (e.g. hydrologic, geographic

and topographic) would permit the use of a complex spatial interpretation system. The access to historic data would also be beneficial, as this would allow for temporal trends to be detected. Another benefit of using a GIS is the ease of visual presentation of the data and the results, hence an improvement could be expected in the communication and dissemination of the results to managers, scientists, politicians and the general public.

## 8.3 Conclusions

The new method of classifying river water quality, using a subset of benthic invertebrate taxa, compared favourably to the existing biotic systems that form the current basis of biological assessment in the UK. The use of five categories for the classification seemed appropriate. It was neither too coarse to be meaningless nor too fine so that it was difficult to apply. The Expert was capable of dividing the five categories into a finer classification (based on the original five classes), so the system is capable of finer discrimination if need be.

Although the neural network models can adequately classify data from the Severn-Trent region, there is no reason to have confidence that these models could also classify data from any other NRA region in the UK. This would also apply, to a certain extent, to any domain expert as well, as their expertise is generally limited to specific geographic regions. Further data acquisition would be necessary before the models could be applied on a national basis.

From the experimental work, it is apparent that a heirarchical approach to modelling (i.e. one that sub-divides the problems into layers or hierarchies) may be a particularly useful idea. Two good examples of this are classification within different biotopes and the identification of novel data. Both of these tasks are easily handled by different neural network models, but it is perhaps the identification of the different sub-problems that is the real progression forward.

Throughout the project, it became clear that there is tremendous amount of domain expertise available for the interpretation of invertebrate communities. As demonstrated, MLP models don't handle explicit prior knowledge well (especially if the prior knowledge is not of really good quality), so obvious

sources of information (i.e. the domain experts) are not being utilised. Other tools, such as probabilistic networks, are more suited to handling the domain knowledge adequately, as well as learning from data (Section 8.2.3). Again, the probabilistic networks lend themselves to hierarchical modelling, so information concerning different regions, from different experts, can be handled in a consistent and mathematically rigorous manner.

The selection of key indicator taxa was considered, and it was shown that is was mathematically possible (using ideas from information theory) to rank taxa in terms of their utility for classification. The idea of selecting key indicator has an intuitive appeal when applied to freshwater biomonitoring. An interesting conclusion is that it is not necessary, or even desirable, to use all of the invertebrate community to form the classification. Beyond a certain threshold any additional taxa appear only to add noise (i.e. they contribute no useful information) and are unreliable indicators (as compared to the other taxa).

The work on the Great Lakes sediment toxicity problem (Chapter 7) demonstrated the utility of the MLP models as classification and prediction tools. The MLPs were the most capable models tested, but their power must be tempered by the use of adequate data.

This dissertation has presented a thorough and principled investigation into the interpretation and classification of freshwater benthic invertebrate communities using artificial neural networks. As demonstrated, neural networks can be successfully applied to solve some of the many difficult problems in freshwater biomonitoring. Yet, the true worth of applying neural networks to freshwater biomonitoring has still to be determined, but the results of this dissertation clearly indicate that the methodology has considerable potential and warrants a more extensive investigation.

# References


[1] Y.S. Abu-Mostafa. Machines that learn from hints. *Scientific American*, pages 68–73, April 1995.

[2] J.A. Anderson, A. Pellionisz, and E. Rosenfeld, editors. *Neurocomputing 2: Directions for research.* MIT Press, Cambridge, MA, 1990.

[3] J.A. Anderson and E. Rosenfeld, editors. *Neurocomputing: Foundations of research.* MIT Press, Cambridge, MA, 1988.

[4] P.D. Armitage, D. Moss, J.F. Wright, and M.T. Furse. The performance of a new biological water quality score based on macroinvertebrates over a wide range of unpolluted running water sites. *Water Research*, 17(3):333–347, 1983.

[5] D. Balloch, C.E. Davies, and F.H. Jones. Biological assessment of water quality in three British rivers: the North Esk (Scotland), the Ivel (England) and the Taf (Wales). *Water Pollution Control*, 75:92–114, 1976.

[6] R. Battiti. Using mutual information for selecting features in supervised neural network learning. *IEEE Transactions on Neural Networks*, 5:537–550, 1994.

[7] R. Beale and T. Jackson. *Neural Computing: An Introduction.* Adam Hilger, Bristol, 1990.

[8] L. Belbin. *PATN Reference Manual.* CSIRO Division of Wildlife and Ecology, Canberra, 1988.

[9] P.E. Bertram and T.B. Reynoldson. Developing ecosystem objectives for the Great Lakes: Policy, progress and public participation. *Journal of Aquatic Ecosystem Health*, 1:89–95, 1992.

[10] C.M. Bishop and C.D. James. Analysis of multiphase flows using dual energy gamma densitometry and neural networks. Technical Report AEA-InTec-1032, United Kingdom Atomic Energy Authority, 1992.

[11] H. Bourlard and C.J. Wellekens. Speech pattern discrimination and multilayer perceptrons. *Computer Speech and Language*, 3:1–19, 1989.

[12] M. Boyd. *The application of methods of uncertain reasoning to the biological monitoring of river water quality (In preparation).* PhD thesis, Department of Civil Engineering, University of Aston, 1996.

[13] M. Boyd, W.J. Walley, and H.A. Hawkes. Dempster-Shafer reasoning for the biological surveillance of river water quality. In L.C. Wrobel and C.A. Brebbia, editors, *Proceedings of the Second International Conference on Water Pollution (Modelling, Measuring and Prediction).* Computational Mechanics Publications, 1993.

[14] J.S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman Soulié and J. Hérault, editors, *Neurocomputing: Algorithms, Architectures and Applications.* Springer-Verlag, 1990.

[15] D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive systems. *Complex Systems*, 2:321–355, 1988.

[16] A.V. Brown and P.P. Brussock. Comparisons of benthic invertebrates between riffles and pools. *Hydrobiologia*, 220:99–108, 1991.

[17] W.L. Buntine and A.S. Weigend. Bayesian back-propagation. *Complex Systems*, 5:603–643, 1992.

[18] W.L. Buntine and A.S. Weigend. Computing second derivatives in feed-forward networks: A review. *IEEE Transactions on Neural Networks*, 5(3):480–488, 1994.

[19] J.R. Chandler. A biological approach to water quality management. *Water Pollution Control*, 69:413–422, 1970.

[20] R.K. Chesters. Biological Monitoring Working Party. the 1978 national testing exercise. *DoE Water Data Unit Technical Memorandum*, 19:1–37, 1980.

[21] L.L. Conquest. Statistical approaches to environmental monitoring: Did we teach the wrong things? *Environmental Monitoring and Assessment*, 26:107–124, 1993.

[22] S.E.K. Cook. Quest for an index of community structure sensitive to water pollution. *Environmental Pollution*, 11:269–288, 1976.

[23] N. De Pauw, P.F. Ghetti, P. Manzini, and R. Spaggiari. Biological assessment methods for running waters. In P. Newman, A. Piavaux, and R. Sweeting, editors, *River Water Quality - Ecological Assessments and Control*, pages 217–248. Commission of European Communities, Brussels, 1993.

[24] N. De Pauw and H.A. Hawkes. Biological monitoring of river water quality. In W.J. Walley and S. Judd, editors, *River Water Quality Monitoring and Control*, pages 87–112. Aston University, 1993.

[25] N. De Pauw and D. Roels. Relationship between the biological and chemical indicators of surface water quality. *Verh. Internat. Verein. Limnol.*, 23:1553–1558, 1988.

[26] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, 1973.

[27] S. Džeroski, L. De Haspe, B.M. Ruck, and W.J. Walley. Classification of river water quality using machine learning. In P. Zanetti, editor, *Computer Techniques in Environmental Studies V (Proceedings of Fifth International Conference on the Development and Application of Computer Techniques to Environmental Studies-ENVIROSOFT '94) Vol. I: Pollution Modelling*, pages 129–137. Computational Mechanics Publications, Southampton, 1994.

[28] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, London, 1993.

[29] J.C. Ellis and P.J. Newman. Compliance with standards - the problems. In W.J. Walley and S. Judd, editors, *River Water Quality Monitoring and Control*, pages 115–134. Aston University, 1993.

[30] Environment Canada. *A Primer of Fresh Water.* Environment Canada, Reading, 1993.

[31] B.S. Everitt and G. Dunn. *Applied Multivariate Data Analysis.* Edward Arnold, London, 1991.

[32] C.A. Extence, A.J. Bates, W.J. Forbes, and P.J. Barnham. Biologically based water quality management. *Environmental Pollution*, 45:221–236, 1987.

[33] E.W. Fager. Determination and analysis of recurrent groups. *Ecology*, 38:586–595, 1957.

[34] S.E. Fahlman. An empirical study of learning speed in back-propagation. Technical Report CMU-CS-88-162, Carnegie Mellon University, 1988.

[35] D.P. Faith, P.R. Minchin, and L. Belbin. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69:57–68, 1987.

[36] D.P. Faith and R.H. Norris. Correlation of environmental variables with patterns of distribution and abundance of common and rare freshwater macroinvertebrates. *Biological Conservation*, 50:77–98, 1989.

[37] J.G. Field, K.R. Clarke, and R.M. Warwick. A practical strategy for analysing multispecies distribution patterns. *Mar. Ecol. Proq. Ser.*, 8:37–52, 1982.

[38] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, New York, 2nd edition, 1987.

[39] G. Fryer. Quantitative and qualitative: Numbers and reality in the study living organisms. *Freshwater Biology*, 17:177–189, 1987.

[40] K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183, 1989.

[41] M.T. Furse, D. Moss, J.F. Wright, and P.D. Armitage. The influence of seasonal and taxonomic factors on the ordination and classification of running-water sites in Great Britain and on the prediction of their macroinvertebrate communities. *Freshwater Biology*, 14:257–280, 1984.

[42] R.G. Gallager. *Information Theory and Reliable Communications.* Wiley, New York, 1968.

[43] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[44] P.E. Gill, W. Murray, and M.H. Wright. *Practical Methods of Optimization.* Academic Press, London, 1981.

[45] C. Girton. *Ecological studies on benthic invertebrate communities in relation to their use in river water quality surveillance.* PhD thesis, University of Aston, 1980.

[46] A. Giwer. The case for and against (Pan-European standards and systems). In W.J. Walley and S. Judd, editors, *River Water Quality Monitoring and Control*, pages 221–223. Aston University, 1993.

[47] S. Hashem. *Optimal linear combinations of neural networks.* PhD thesis, School of Industrial Engineering, Purdue University, 1993.

[48] B. Hassibi and D.G. Stork. Second order derivatives for network pruning: Optimal Brain Surgeon. In S.J. Hanson, J.D. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 164–171. Morgan Kaufman, San Mateo, CA, 1993.

[49] H.A. Hawkes. River zonation and classification. In B. Whitton, editor, *River Ecology*. Blackwell, Oxford, 1975.

[50] H.A. Hawkes. Manual on biological surveillance of rivers using benthic macroinvertebrates. Unpublished Report, University of Aston: Applied Hydrobiology Section, 1977.

[51] H.A. Hawkes. Invertebrates as indicators of river water quality. In A. James and L. Evison, editors, *Biological indicators of water quality*, pages 2.1–2.24. Wiley, Chichester, 1979.

[52] H.A. Hawkes. Water quality issues: An ecological reaction. *Chemistry and Industry*, March:201–204, 1979.

[53] H.A. Hawkes. Biological surveillance of rivers. *Water Pollution Control*, 81(3):329–342, 1982.

[54] H.A. Hawkes and L.J. Davies. Some effects of organic enrichment on benthic invertebrate communities in stream riffles. In E. Duffey and A.S. Watt, editors, *The Scientific Management of Animal and Plant Communities for Conservation*. Blackwell, London, 1971.

[55] J.M. Hellawell. *Biological Surveillance of Rivers*. Water Research Centre, Medmenham, 1978.

[56] J.M. Hellawell. *Biological Indicators of Freshwater Pollution and Environmental Management*. Elsevier, London, 1986.

[57] R.J. Henery. Methods for comparison. In D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors, *Machine Learning, Neural and Statistical Classification*, pages 107–124. Ellis Horwood, 1994.

[58] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA, 1991.

[59] M.O. Hill. Twinspan—A fortran program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. *ecology and systematics*, 1979.

[60] G.E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40:185–234, 1989.

[61] G.E. Hinton, C.K.I. Williams, and M.D. Revow. Adaptive elastic models for hand-printed character recognition. In J.E. Moody, S.J. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann, 1992.

[62] Y. Hirose, K. Yamashita, and S. Hijika. Back propagation algorithm which varies the number of hidden units. *Neural Networks*, 4:61–66, 1991.

[63] T. Hruby. Using similarity measures in benthic impact assessments. *Environmental Monitoring and Assessment*, 8:163–180, 1987.

[64] D.R. Hush and B.G. Horne. Progress in supervised neural networks: What's new since Lippmann? *IEEE Signal Processing Magazine*, pages 8–39, Jan 1993.

[65] H.B.N. Hynes. *The Biology of Polluted Waters.* Liverpool University Press, Liverpool, 1960.

[66] International Joint Commission. Guidance of characterization of toxic substances problems in areas of concern in the Great Lakes basin. Report from the Surveillance Work Group, pp179, 1987.

[67] International Joint Commission. Procedures for the assessment of contaminated sediment problems in the Great Lakes. Sediment Subcommittee, pp140, 1988.

[68] P. Jaccard. Nouvells recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270, 1908.

[69] D.A. Jackson. Multivariate analysis of benthic invertebrate communities the implication of choosing particular data standardizations, measures of association, and ordination methods. *Hydrobiologia*, 268:9–26, 1993.

[70] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[71] M. Jefferies. Water quality and wildlife. Report for Nature Conservancy Council, Contract HF 3 03 370, 1988.

[72] B. Jepson, A. Collins, and A. Evans. Post-neural network procedure to determine expected prediction values and their confidence limits. *Neural Computing and Applications*, 1(3):224–228, 1993.

[73] W.H. Joerding and J.L. Meader. Encoding a priori information in feed-forward networks. *Neural Networks*, 4:847–856, 1991.

[74] R.K. Johnson and T. Wiederholm. Classification and ordination of profundal macroinvertebrate communities in nutrient poor, olgio-mesohumic lakes in relation to environmental data. *Freshwater Biology*, 21:375–386, 1989.

[75] R.K. Johnson, T. Wiederholm, and D.M. Rosenberg. Freshwater biomonitoring using individual organisms, populations, and species assemblages of benthic macroinvertebrates. In D.M. Rosenberg and V.H. Resh, editors, *Freshwater Biomonitoring and Benthic Macroinvertebrates*, pages 40–125. Chapman & Hall, London, 1993.

[76] D.S. Jones. *Elementary Information Theory*. Clarendon Press, Oxford, 1979.

[77] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM alogorithm. *Neural Computation*, 6:181–214, 1994.

[78] I.G. Jowett. A method for objectively identifying pool, run, and riffle habitats from physical measurements. *New Zealand Journal of Marine and Freshwater Research*, 27:241–248, 1993.

[79] L. Kanal. On patterns, categories and alternate realities. *Pattern Recognition Letters*, 14:241–255, 1993.

[80] F. Kanaya and K. Nakogawa. On the practical implementation of mutual information for statistical decision making. *IEEE Transactions on Information Theory*, 37:1151–1156, 1991.

[81] G.K. Kanji. *100 Statistical Tests*. Sage, London, 1993.

[82] E.D. Karin. A simple procedure for pruning back propagation trained neural networks. *IEEE Transactions on Neural Networks*, 1:239–242, 1990.

[83] T. Kohonen. The self-organizing map. *Transactions of the IEEE*, 78(9):1464–1480, 1990.

[84] T. Kohonen, J. Kangas, and J. Laaksonen. The Self-Organizing Map Program Package. Laboratory of Computer and Information Science, Helsinki University of Technology. Program available by anonymous ftp from `cochlea.hut.fi`, 1992.

[85] R. Kolkwitz and M. Marsson. Oekolgie der pflanzlichen saprobien. *Ber. dtsch. bot. Ges.*, 26A:505–519, 1908.

[86] R. Kolkwitz and M. Marsson. Oekologie der kerischen saprobien. *Rev. Ges. Hydrobiol. Hydrogr.*, 2:126–152, 1909.

[87] D.C.L. Lam, I. Wong, D.A. Swayne, J. Storey, and J.P. Kerby. Application of the RAISON expert system for water pollution problems from acid rain to mine effluent. In P. Zannetti, editor, *Computer Techniques*

*in Environmental Studies III (Proceedings of Fifth International Conference on the Development and Application of Computer Techniques to Environmental Studies-ENVIROSOFT '90)*, pages 273–284. Computational Mechanics Publications, Southampton, 1990.

[88] M. Le Blanc and R. Tibshirani. Combining estimates in regression and classification. Technical Report. Department of Statistics, University of Toronto, 1993.

[89] Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Hubbard, and L. D. Kackel. Handwritten digit recognition with a back propagation network. *Neural Computation*, 1:541–551, 1989.

[90] Y. Le Cun, J.S. Denker, and S.A. Solla. Optimal brain damage. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 598–605. Morgan Kaufman, San Mateo, CA, 1990.

[91] R.P. Lippmann. Introduction to computing with neural nets. *IEEE ASSP Mag*, Apr:4–22, 1987.

[92] R.P. Lippmann. Pattern classification using neural networks. *IEEE Communications Magazine*, Nov.:47–50,59–64, 1989.

[93] P. Logan and M.P. Brooker. The macroinvertebrate fauna of riffles and pools. *Water Research*, 17(3):263–270, 1983.

[94] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[95] D.J.C. MacKay. A practical Bayesian framework for back propagation networks. *Neural Computation*, 4:448–472, 1992.

[96] P.S. Maitland. *A Coded Checklist of Animals Occurring in Fresh Waters in the British Isles.* Institute of Terrestrial Ecology, Edinburgh, 1977.

[97] R. Margalef. *Perspectives in Ecological Theory.* University of Chicago Press, London, 1968.

[98] C. Mason. *Biology of Freshwater Pollution.* Longman, London, 1981.

[99] W.T. Mason, P.A. Lewis, and C.I. Weber. An evaluation of benthic macroinvertebrate biomass methodology. *Environmental Monitoring and Assessment*, 5:399–422, 1985.

[100] W.S. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity forms. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[101] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley, New York, 1992.

[102] E.F. Menhinick. A comparison of some species—individuals diversity indices applied to samples of field insects. *Ecology*, 45:859–861, 1964.

[103] L.J. Metcalfe. Biological water quality assessment of running waters based on macroinvertebrates communities: History and present status in Europe. *Environmental Pollution*, 60:101–139, 1989.

[104] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors. *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, 1994.

[105] M.L. Minsky and S.A. Papert. *Perceptrons. An introduction to Computational Geometry. Expanded edition.* MIT Press, 1988.

[106] J.M.O. Mitchell. Classical statistical methods. In D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors, *Machine Learning, Neural and Statistical Classification*, pages 17–28. Ellis Horwood, 1994.

[107] J.E. Moody. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In J.E. Moody, S.J. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 847–854. Morgan Kaufman, San Mateo, CA, 1992.

[108] D.P. Morgan and C.L. Scofield. *Neural Networks and Speech Processing.* Kluwer Academic Publishers, 1991.

[109] D. Moss, M.T. Furse, J.F. Wright, and P.D. Armitage. The prediction of the macroinvertebrate fauna of unpolluted running water sites in Great Britain using environmental data. *Freshwater Biology*, 17:41–52, 1987.

[110] N. Murata, S. Yoshizawa, and S. Amari. A criterion for determining the number of parameters in an artificial neural network model. In T. Kohonen, K. Mäkisarra, O. Simila, and J. Kangas, editors, *Artificial Neural Networks*, pages 9–14. North Holland, Amsterdam, 1991.

[111] P.M. Murphy. The temporal variability in biotic indices. *Environmental Pollution*, 17:227–236, 1978.

[112] National Rivers Authority. The quality of rivers, canals and estuaries in England and Wales. Report of the 1990 Survey. National Rivers Authority, Water Quality Series No. 4, 1991.

[113] National Rivers Authority. Proposals for Statutory Water Quality Objectives. National Rivers Authority. Water Quality Series No. 5, 1991.

[114] National Water Council. River water quality—the next stage. Review of discharge consent conditions, 1978.

[115] Natural Environment Research Council. *Biological Monitoring.* NERC., London, 1977.

[116] R.M. Neal. Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical Report Technical Report CRG-TR-92-1, Department of Computer Science, University of Toronto, 1992.

[117] R.M. Neal. *Bayesian Learning for Neural Networks.* PhD thesis, Department of Computer Science, University of Toronto, 1994.

[118] E. Neapolitan. *Probabilistic reasoning in expert systems.* John Wiley, London, 1990.

[119] R.H. Norris and A. Georges. Analysis and interpretation of benthic macroinvertebrate surveys. In D.M. Rosenberg and V.H. Resh, editors, *Freshwater Biomonitoring and Benthic Macroinvertebrates*, pages 234–286. Chapman & Hall, London, 1993.

[120] S.J. Nowlan. *Soft Competitive Adaptation: Neural Network Learning Algorithms based on Fitting Statistical Mixtures.* PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1991.

[121] S.J. Nowlan and G.E. Hinton. Evaluation of adaptive mixtures of competing experts. In R.P. Lippmann, J.E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing Systems 3.* Morgan Kaufman, San Mateo, CA, 1991.

[122] W. Olthuis. Chemical sensing in freshwater - problems and opportunities. In W.J. Walley and S. Judd, editors, *River Water Quality Monitoring and Control*, pages 209–218. Aston University, 1993.

[123] R. Pantle and H. Buck. Die biologische Uberwachung der Gewässer und die Darstellung der Ergebnisse. *Gas und Wasserfach*, 96:604, 1955.

[124] Y-H. Pao. *Adaptive Pattern Recognition and Neural Networks.* Addison Wesley, New York, 1989.

[125] D.B. Parker. Learning-logic. Technical Report TR-47, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, 1985.

[126] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, California, 1988.

[127] G. Personne and N. De Pauw. Systems of biological indicators for water quality assessment. In O. Ravera, editor, *Biological Aspects of Freshwater Pollution.* Pergamon Press, 1979.

[128] L.C.V. Pinder and I.S. Farr. Biological surveillance of water quality (2). Temporal and spatial variation in the macroinvertebrate fauna in the river forme - A Dorset chalk stream. *Archiv Fur Hydrobiologie*, 109(3):321–331, 1987.

[129] L.C.V. Pinder, M. Ladel, T. Gledhill, J.A.B. Bass, and A.M. Matthews. Biological surveillance of water quality (1). A comparison of macroinvertebrate surveillance methods in relation to assessment of water quality in a chalk stream. *Archiv Fur Hydrobiologie*, 109(2):207–226, 1987.

[130] D.C. Plaut, S.J. Nowlan, and G.E. Hinton. Experiments on learning by back-propagation. Technical Report CMU-CS-86-126, Carnegie-Mellon University, Pittsburgh, PA, 1986.

[131] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C.* Cambridge University Press, Cambridge, 2 edition, 1992.

[132] C.F. Rabeni, S.P. Davies, and K.E. Gibbs. Benthic invertebrate response to pollution abatement: Structural changes and functional implications. *Water Quality Bulletin*, 21:489–497, 1985.

[133] K.H. Reckhow. Bayesian inference in non-replicated ecological studies. *Ecology*, 71(6):2053–2059, 1990.

[134] S.J. Renals. Radial basis function network for speech pattern classification. *Electronics Letters*, 25(7):437–439, 1989.

[135] S.J. Renals. *Speech and Neural Networks Dynamics.* PhD thesis, University of Edinburgh, 1990.

[136] T.B. Reynoldson, R.C. Bailey, K.E. Day, and R.H. Norris. Biological guidelines for freshwater sediment based on BEnthic Assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. *Australian Journal of Ecology*, 20:198–219, 1995.

[137] T.B. Reynoldson and K.E. Day. A study plan for the development of biological sediment guidelines. National Water Research Institute unpublished report, Burlington, Canada, 1991.

[138] T.B. Reynoldson and J.L. Metcalfe-Smith. An overview of the assessment of aquatic ecosystem health using benthic invertebrates. *Journal of Aquatic Ecosystem Health*, 1:295–308, 1992.

[139] T.B. Reynoldson and M.A. Zarull. The biological assessment of contaminated sediments - the Detroit River example. *Hydrobiologia*, 188/189:463–476, 1989.

[140] T.B. Reynoldson and M.A. Zarull. An approach to the development of biological sediment criteria. In S.J. Woodley, G. Francis, and J. Kay, editors, *Ecological Integrity and the Management of Ecosystems*, pages 177–200. St Lucie Press, Fl., 1993.

[141] B.D. Ripley. Statistical aspects of neural networks. In O.E. Barndorff-Nielsen, D.R. Cox, J.L. Jensen, and W.S. Kendall, editors, *Networks and Chaos - Statistical and Probabilistic Aspects*. Chapman & Hall, London, 1993.

[142] B.D. Ripley. Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society, Series B*, 56(3):409–456, 1994.

[143] S.J. Roberts and L. Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6:270–284, 1993.

[144] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7:777–781, 1994.

[145] R.J. Rohwer, M. Wynne-Jones, and F. Wysotzki. Neural networks. In D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors, *Machine Learning, Neural and Statistical Classification*, pages 84–106. Ellis Horwood, 1994.

[146] C. Rose. *The Dirty Man of Europe: the Great British Pollution Scandal*. Simon & Schuster, London, 1990.

[147] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–405, 1958.

[148] B.M. Ruck, W.J. Walley, and H.A. Hawkes. Biological classification of river water quality using neural networks. In G. Rzevski, Pastor J., and R.A. Adey, editors, *Applications of Artificial Intelligence in Engineering VIII, Vol. 2 Applications and Techniques*. Elsevier, 1993.

[149] B.M. Ruck, W.J. Walley, T.B. Reynoldson, and K.E. Day. A neural network predictor of benthic community structure in the Canadian waters of the Laurentian Great Lakes. In L.C. Wrobel and C.A. Brebbia, editors, *Proceedings of the Second International Conference on Water Pollution (Modelling, Measuring and Prediction)*. Computational Mechanics Publications, 1993.

[150] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations for error propagation. In D.E. Rumelhart, J.L. McClelland, and G.E. Hinton, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, 1986.

[151] D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, editors. *Parallel Distributed Processing: Explorations in the Microstrcuture of Cognition. Volume 1: Foundations.* MIT Press, Cambridge, MA, 1986.

[152] J.W. Sammon Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C18(5):401–409, 1969.

[153] SAS Institute Inc. *User's Guide, Version 6, Fourth Edition.* SAS Institute Inc., Cary, NC, 1990.

[154] W. Schaafsma. Selecting variables in discriminant analysis for improving classical procedures. In P.R. Krishnaiah and L. Kanal, editors, *Handbook of Statistics*, volume 2, pages 857–881. North-Holland, Amsterdam, 1982.

[155] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, 1949.

[156] E.H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.

[157] V. Sladecek. Systems of water quality from the biological point of view. *Arch. Hydrobiol. Beih.*, 7:1–218, 1973.

[158] T. Sorenson. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Bio. Skr. (K. danske. vidensk. Selsk. N.S.)*, 5:1–34, 1948.

[159] D.J. Spiegelhalter and R.G. Cowell. Learning in probabilistic expert systems. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 4*, pages 447–466. Clarendon Press, Oxford, 1992.

[160] D.J. Spiegelhalter, A.P. Dawid, and S.L. Lauritzen. Bayesian analysis in expert systems. *Statistical Science*, 8:219–247, 1993.

[161] L.E. Sucar, D.F. Gillies, and D.A. Gillies. Objective probabilities in expert systems. *Artificial Intelligence*, 61:187–208, 1993.

[162] H.H. Thodberg. Ace of Bayes: Application of neural networks with pruning. Technical Report Technical Report No. 1132E, Danish Meat Research Institute, 1993.

[163] S.B. Thrun and 23 co authors. The MONK's problems: A performance comparison of different learning algorithms. Technical Report CMU-CS-91-197, Department of Computer Science, Carnegie-Mellon University, 1991.

[164] H.H. Tolkamps. Biological assessment of water quality in running water using macroinvertebrates a case study for Limburg, The Netherlands. *Wat. Sci. Tech.*, 17:867–878, 1985.

[165] D. van Camp, T. Plate, and G.E. Hinton. The Xerion Neural Network Simulator. Department of Computer Science, University of Toronto. Program available by anonymous ftp from `ftp.cs.toronto.edu`, 1993.

[166] V.N. Vapnik and A. Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Application*, 2(10):264–280, 1971.

[167] W.J. Walley. Artificial intelligence in river water monitoring and control. In W.J. Walley and S. Judd, editors, *River Water Quality Monitoring and Control*, pages 179–194. Aston University, 1993.

[168] W.J. Walley. New approaches to the interpretation of water quality data based on techniques from the field of artificial intelligence. In *Proceedings of the Workshop on Monitoring Tailor-made*. Beekbergen, The Netherlands, 1994.

[169] W.J. Walley, M. Boyd, and H.A. Hawkes. An expert system for the biological monitoring of river pollution. In *Proceedings of the Fourth International Conference on Computer Techniques in environmental Studies*. Elsevier, Portsmouth, England, 1992.

[170] W.J. Walley, H.A. Hawkes, and M. Boyd. Application of Bayesian inference to river water quality surveillance. In D.E. Grierson, G. Rzevski, and R.A. Adey, editors, *Applications of Artificial Intelligence in Engineering VII*. Elsevier, 1992.

[171] H.G. Washington. Diversity, biotic and similarity indices a review with special relevance to aquatic ecosystems. *Water Research*, 18(6):653–694, 1984.

[172] P.D. Wasserman. *Neural Computing: Theory and Practice.* Chapman & Hall, Routledge, 1990.

[173] A.S. Weigend and N.A. Gershenfeld, editors. *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley, Reading, 1994.

[174] P. Werbos. *Beyond regression: New tools for prediction and analysis in the behavioural sciences.* PhD thesis, Harvard University, 1975.

[175] J.L. Whilm and T.C. Dorris. Biological parameters for water quality criteria. *Bioscience*, 18:477–481, 1968.

[176] I.T. Whitehurst. The *Gammarus:Asellus* ratio as an index of organic pollution. *Water Research*, 25:333–339, 1991.

[177] S.J. Wishart, J.P. Lumbers, and I.M. Griffiths. Expert systems for the interpretation of river water quality data. *Journal of the Institute of Water and Environmental Management*, 4:. 194–202, April 1990.

[178] D.H. Wolpert. Stacked generalisation. *Neural Networks*, 5:241–259, 1992.

[179] F.S. Woodiwiss. The biological system of stream classification used by the Trent River Board. *Chemistry and Industry*, 11:443–447, 1964.

[180] J.F. Wright, P.D. Armitage, and M.T. Furse. Prediction of invertebrate communities using stream measurements. *Regulated Rivers Research and Management*, 4:147–155, 1989.

[181] J.F. Wright, D. Moss, P.D. Armitage, and M.T. Furse. A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. *Freshwater Biology*, 14:221–256, 1984.

# Appendix A1:
# Taxa Recorded in the Severn-Trent Database†

**Planariidae**
- *Planaria torva*
- *Polycelis*
- *Polycelis felina*
- *Polycelis nigra*
- *Polycelis tenuis*
- *Dugesia*
- *Dugesia lugubris*
- *Dugesia polychroa*
- *Dugesia tigrina*

**Dendrocoelidae**
- *Dendrocoelum lacteum*
- *Theodoxus fluviatilis*
- *Viviparus*
- *Valvata*

**Hydrobiidae**
- *Potamopyrgus jenkinsi*
- *Bithynia*

**Physidae**
- *Physa fontinalis*

**Lymnaeidae**
- *Lymnaea auricularia*
- *Lymnaea glabra*
- *Lymnaea palustris*
- *Lymnaea peregra*
- *Lymnaea stagnalis*
- *Lymnaea truncatula*

**Planorbidae**
- *Armiger crista*
- *Planorbarius corneus*

**Ancylidae**
- *Ancylus fluviatilis*
- *Acroloxus lacustris*

**Unionidae**
- *Unio*
- *Anodonta*
- *Anodonta cygnea*

**Sphaeriidae**
- *Sphaerium*
- *Pisidium*

OLIGOCHAETA
- **Lumbriculidae**
- **Tubificidae**
- **Lumbricidae**
  - *Piscicola geometra*
- **Glossiphoniidae**
  - *Theromyzon tessulatum*
  - *Hemiclepsis marginata*
  - *Glossiphonia complanata*
  - *Glossiphonia heteroclita*
  - *Helobdella stagnalis*
- **Hirudinidae**
  - *Haemopis sanguisuga*
- **Erpobdellidae**
  - *Erpobdella octoculata*
  - *Erpobdella testacea*
  - *Trocheta*

HYDRACARINA

CLADOCERA

OSTRACODA

COPEPODA
- *Austropotamobius*
- **Asellidae**
  - *Asellus aquaticus*
  - *Asellus meridianus*
  - *Corophium curvispinum*
  - *Crangonyx pseudogracilis*
  - *Gammarus*
  - *Gammarus pulex*
  - *Gammarus tigrinus*
- **Baetidae**
  - *Baetis rhodani*
  - *Centroptilum luteolum*
  - *Cloeon*
  - *Cloeon dipterum*
- **Heptageniidae**
  - *Rhithrogena semicolorata*
  - *Heptagenia*

*Ecdyonurus*
*Ecdyonurus venosus*
**Leptophlebiidae**
*Paraleptophlebia*
*Paraleptophlebia submarginata*
*Habrophlebia fusca*
**Ephemeridae**
*Ephemera danica*
*Ephemera vulgata*
**Ephemerellidae**
*Ephemerella ignita*
**Caenidae**
*Caenis*
*Caenis luctuosa*
*Caenis rivulorum*
**Taeniopterygidae**
*Taeniopteryx nebulosa*
*Brachyptera risi*
**Nemouridae**
*Protonemura*
*Protonemura meyeri*
*Amphinemura*
*Nemurella picteti*
*Nemoura*
**Leuctridae**
*Leuctra geniculata*
**Capniidae**
**Perlodidae**
*Isoperla grammatica*
**Perlidae**
*Dinocras cephalotes*
**Chloroperlidae**
*Chloroperla torrentium*
*Platycnemis pennipes*
**Coenagriidae**
**Calopterygidae**
*Calopteryx splendens*
*Calopteryx virgo*
**Libellulidae**
**Mesovelidae**
**Hydrometridae**
**Veliidae**
**Gerridae**
*Nepa cinerea*
**Notonectidae**
**Corixidae**
**Haliplidae**
*Brychius elevatus*
*Haliplus*
*Hygrobia hermanni*
**Dytiscidae**
**Gyrinidae**
**Hydrophilidae**
*Hydraena*
**Scirtidae**
**Elmidae**
*Elmis aenea*
*Esolus parallelepipedus*
*Limnius volckmari*
*Oulimnius*
**Curculionidae**
**Sialidae**
*Sialis fuliginosa*
*Sialis lutaria*
**Rhyacophilidae**
*Rhyacophila dorsalis*
*Glossosoma*
*Agapetus*
**Hydroptilidae**
*Hydroptila*
**Philopotamidae**
**Psychomyiidae**
*Tinodes*
*Tinodes waeneri*
**Polycentropidae**
*Polycentropus flavomaculatus*
**Hydropsychidae**
*Hydropsyche angustipennis*
*Hydropsyche contubernalis*
*Hydropsyche instabilis*
*Hydropsyche pellucidula*
*Hydropsyche siltalai*
**Phryganeidae**
*Phryganea grandis*
*Brachycentrus subnubilus*
**Lepidostomatidae**
*Lepidostoma hirtum*
**Limnephilidae**
*Drusus annulatus*
*Ecclisopteryx guttulata*

*Micropterna*
*Potamophylax*
*Glyphotaelius pellucidus*
*Limnephilus*
*Limnephilus extricatus*
*Limnephilus fuscicornis*

**Goeridae**

**Beraeidae**

*Beraeodes minutus*

**Sericostomatidae**

*Sericostoma personatum*
*Odontocerum albicorne*

**Molannidae**

**Leptoceridae**

*Athripsodes*
*Athripsodes commutatus*
*Leptocerus*
*Mystacides*

**Tipulidae**

*Pedicia rivosa*
*Dicranota*

**Psychodidae**

*Pericoma*

**Ptychopteridae**

**Dixidae**

**Chaoboridae**

**Culicidae**

**Ceratopogonidae**

**Simuliidae**

**Chironomidae**

*Chironomus riparius*

**Stratiomyidae**

*Atherix ibis*

**Tabanidae**

**Empididae**

**Syrphidae**

*Limnophora riparia*

†This list details all the taxa that were identified in the Severn-Trent database (Section 4.3). Note that it is not taxonomically rigorous.

# Appendix A2:
# Taxa Recorded in the National NRA Database

Acroloxidae
Bithyniidae
Ceratopogonidae
Crangonyctidae
Dugesiidae
Chaoboridae
Ecnomidae
Empididae
Enchytraeidae
Glossosomatidae
Hebridae
HYDRACARINA
Hydraenidae
OLIGOCHAETA
Chironomidae
Valvatidae
Hydrobiidae
Lymnaeidae
Physidae
Planorbidae
Sphaeriidae
Glossiphoniidae
Hirudinidae
Erpobdellidae
Asellidae
Baetidae
Sialidae
Piscicolidae
Mesoveliidae
Hydrometridae
Gerridae
Nepidae
Naucoridae
Notonectidae
Pleidae
Corixidae
Haliplidae
Hygrobiidae
Dytiscidae
Gyrinidae
Hydrophilidae
Clambidae
Scirtidae
Dryopidae
Elmidae
Chrysomelidae
Curculionidae
Hydropsychidae
Tipulidae
Simuliidae
Planariidae
Dendrocoelidae
Neritidae
Viviparidae
Ancylidae
Hydroptilidae
Unionidae
Corophiidae
Gammaridae
Platycnemididae
Coenagriidae
Caenidae
Nemouridae
Rhyacophilidae
Polycentropidae
Limnephilidae
Astacidae
Lestidae
Calopterygidae
Gomphidae
Cordulegasteridae
Aeshnidae
Corduliidae
Libellulidae
Psychomyiidae
Philopotamidae
Lumbriculidae
Muscidae
Naididae
Noteridae

OSTRACODA
Psychodidae
Ptychopteridae
Rhagionidae
Syrphidae
Stratiomyidae
Tubificidae
Tabanidae
Veliidae
Siphlonuridae
Heptageniidae
Leptophlebiidae
Ephemerellidae
Potamanthidae
Ephemeridae
Taeniopterygidae
Leuctridae
Capniidae
Perlodidae
Perlidae
Chloroperlidae
Aphelocheiridae
Phryganeidae
Molannidae
Beraeidae
Odontoceridae
Leptoceridae
Goeridae
Lepidostomatidae
Brachycentridae
Sericostomatidae
Lumbricidae
Daphniidae
LEPIDOPTERA
Thaumaleidae
Culicidae
Spongillidae
Dixidae
Sisyridae
Osmylidae

# Appendix A3:
# Great Lakes Sediment Guidelines Project Species List

GASTROPODA
- **Bithyniidae**
  - *Bithynia tentaculata**
- **Hydrobiidae**
  - *Amnicola limosa*
  - *A. walkeri*
  - *Marstonia decepta*
  - *Probythinella lacustris*
  - *Hydrobiidae immatures*
- **Lymnaeidae**
  - *Fossaria obrussia*
- **Physidae**
  - *Physella integra*
  - *P. spp.**
- **Planorbidae**
  - *Armiger crista*
  - *Gyraulus circumstriatus*
  - *G. deflectus*
  - *Helisoma anceps*
  - *Promenetus exacuous*
- **Valvatidae**
  - *Valvata lewisi*
  - *V. piscinalis**
  - *V. sincera*
  - *V. tricarinata**
- **Viviparidae**
  - *Campeloma decisum*
- **Unknown spp.***

PELYCEPODA
- **Sphaeridae**
  - *Pisidium casertanum**
  - *P. compressum**
  - *P. ferrugineum*
  - *P. henslowanum**
  - *P. nitidum**
  - *P. ventricosum*
  - *P. unknown**
  - *Sphaerium nitidum*
  - *S. simile*
  - *S. striatum*
  - *S. unknown*
  - *Musculim partinium*
  - *M. securis**
  - *M.transversum*
- **Unionidae**
  - *Elliptio camplanata*
  - *Lampris radiata*
- **Dreissenidae**
  - *Dreissena polymorpha**

DIPTERA
- **Chironomidae**
  - *Chironomus**
  - *Cladopelma**
  - *Clanotanytarsus*
  - *Cryptochironomus**
  - *Cryptotendipes**
  - *Dicrotendipes**
  - *Demicryptochironomus**
  - *Endochironomus**
  - *Glypotendipes**
  - *Harnischia*
  - *Micropsectra**
  - *Microtendipes**
  - *Nilothauma*
  - *Pagastiella*
  - *Parachironomus*
  - *Paracladoplema*
  - *Paralauterborniella*
  - *Paratendipes*
  - *Polypedium**
  - *Pseudochironomus*
  - *Stictochironomus**
  - *Tanytarsus**
  - *Tribelos*
  - *Stempellina*
  - *Zavreliella*
  - *Unknown Chironominae*
  - *Corynoneura*

*Cricotopus*
*Epoicricotopus*
*Heterotrissocladius**
*Nanocladius*
*Parakiefferiella*
*Psectrocladius**
*Unknown orthocladinae*
*Ablabesmyia*
*Clinotanypus*
*Coelotanypus**
*Larsia*
*Macropelopia*
*Procladius**
*Tanypus*
*Monodiamesia*

**Ceratopogonidae**
*Bezzia spp.*
*Ceratopogon spp.*
*Culicoides spp.*
*Mallochohelea spp.*
*Probezzia spp.*
*Serronmyia spp.*

**Chaoboridae**
*Chaoborus spp.**

**Empididae**

EPHEMEROPTERA

**Ephemeridae**
*Hexagenia limbata*

**Caenidae**
*Caenis spp.**

COLEMBOLA

TRICHOPTERA

**Polycentropidae**
*Polycentropus spp.*
*Cernatina spp.*
*Phylocentropus spp.*

**Helicopsychidae**
*Helicopsyche spp.*

**Leptoceridae**
*Leptocerus americanus*
*Mystacides spp.*
*Nectopsyche spp.*
*Oecetis spp.*
*Setodes spp.*

**Molannidae**
*Molanna spp.*

**Hydroptilidae**
*Agraylea spp.*

POLYCHAETA

**Sabellidae**
*Manayunkia speciosa**

OLIGOCHAETA

**Lumbriclidae**
*Eclipidrilus lacustris*
*Lumbriculus variegatus*
*Stylodrilus herringlanus**

**Enchytreidae***

**Naididae**
*Arcteonais lomondi**
*Chaetogaster diaphanus*
*Nais barbata*
*N. elinguis*
*N. pseudobtusa*
*N. simplex*
*N. variabilia**
*Piguetiella michiganensis*
*Pristina leidyi*
*Pristinella acuminata*
*Ophidonais serpentina*
*Specaria josinae**
*Stylaria lacustris**
*Uncinais uncinata*
*Vejdovskyella intermedia**

**Tubificidae**
*Immatures with hair chaetae**
*Immatures without hair chaetae**
*Aulodrilus americana*
*A. limnobius*
*A. pigueti**
*A. pluriseta**
*Branchiura sowerbyi*
*Ilyodrilus templetoni*
*Limnodrilus claparedianus*
*L. cervix*
*L. hoffmeisteri**
*L. profundicola*
*Potamothrix bedoti*
*P. moldaviensis**
*P. vejdovskyi**
*Quadradrilus multisetosus**

*Spirosperma ferox**
*Tasserkidrilus superiorensis*
*Tubifex tubifex**
HIRUDINEA
**Glossiphoniidae**
*Alboglassiphonia heteroclita*
*Gloiobdella elongata*
*Helobdella stagnalis*
**Piscicolidae**
*Myzobdella lugubris*
PLATYHELMINTHES*
ISOPODA
**Asellidae**
*Caecidotea communis*
*C. intermedius**
*C. spp.**
AMPHIPODA
**Gammaridae**
*Gammarus lacustris*
**Haustoriidae**
*Diporeia hoyi**
**Taliridae**
*Hyalella azteca*
COELENTERATA
**Hydridae**
*Hydra americana**
PORIFERA*
TARDIGRADA
**Milnesiidae**
*Milnesium tardigradum*

Taxon denoted by * were used in the ordination of the community structure (see Section 7.3.3).