Applications of Artificial Intelligence for the Biological Surveillance of River Quality

Technical Report
E52

Applications of Artificial Intelligence for the Biological Surveillance of River Quality

R&D Technical Report E52

W J Walley, V N Fortama and R W Martin

Research Contractor:
School of Computing, Staffordshire University

Further copies of this report are available from:
Environment Agency R&D Dissemination Centre, c/o
WRc, Frankland Road, Swindon, Wilts SN5 8YF

tel: 01793-865000 fax: 01793-514562 e-mail: publications@wrcplc.co.uk

Publishing Organisation:
Environment Agency
Rio House
Waterside Drive
Aztec West
Almondsbury
Bristol BS32 4UD

Tel: 01454 624400 Fax: 01454 624409

TH-01/98-B-BASA

Dissemination status
Internal: Released to Regions
External: Released to Public Domain

Statement of use
This report represents the findings of an investigation into applications of artificial intelligence for the biological surveillance of river quality. It is intended for use by the Agency's staff and others interested in the computer-based interpretation of biological data into river quality terms.

Research contractor
This document was produced under R&D Project E1/i621 by:

School of Computing
Staffordshire University
The Octagon
Beaconside
Staffordshire
ST18 0AD

Tel: 01785 353510

Environment Agency Project Manager
The Environment Agency's Project Manager for R&D Project E/i621 was:
Dr John Murray-Bligh, Thames Region

CONTENTS

TABLES

FIGURES

APPENDICES

GLOSSARY

The following list gives brief definitions of the technical terms and acronyms that are used throughout the report. Some of them are defined in greater detail within the report. In these cases the appropriate Section number is given.

ALK	Alkalinity (mg/l of $CaCO_3$).
ALT	Altitude (m).
ASPT	Average Score Per Taxon.
Back-propagation	A commonly used training algorithm for supervised-learning neural networks.
BBN	Bayesian Belief Networks (Section 1.5).
BLDS	Percentage of boulders in the substrate.
BMWP	Biological Monitoring Working Party score system.
BOD	Biochemical Oxygen Demand.
Conformity index	A measure of how consistent a given taxon's state is relative to the rest of the sample (Section 5.5).
DEPTH	Average depth of river (cm).
DISCH	Discharge category.
DO	Dissolved Oxygen (percentage saturation).
EQI	Environmental Quality Index.
F1 and F2	Data files derived from IFE614 database, each having 307 records.
F1/F2	A cross-validated training scheme: trained on F1, tested on F2.
Feature map	A topographic map showing the variation of a given variable over the output array of a SOM (Sections 4.4.2 and 4.4.3).
GQA	General Quality Assessment – a river quality assessment scheme used by the Environment Agency.
GTM	Generative Topographic Mapping – a type of unsupervised neural network.
GTM10	A site specific GTM with a 10x10 output array (Section 4.2).
IFE	Institute of Freshwater Ecology.
IFE614	A database of 614 'unpolluted' sites, from which RIVPACS III was developed by IFE.
Impact analysis	A procedure for determining the relative importance of individual input nodes in a supervised neural network. (Section 3.3.4).
LDIST	Log_{10} of distance from source (km).
LSLOPE	Log_{10} of slope (m/km).
$M(C,X)$	Mutual information between class (C) and attribute (X) - (2.1.2).
$M'(C,X)$	Indifferent mutual information between C and X (Section 2.1.2).
N13DASPT	A dependent neural network predictor of ASPT having 13 inputs.

N13DNFAM	A dependent neural network predictor of NFAM having 13 inputs.
N5DASPT	A dependent neural network predictor of ASPT having 5 inputs.
N5DRASPT	A dependent neural network predictor of ASPT (based on revised scores) having 5 inputs.
N5XASPT	An independent neural network predictor of ASPT having 5 inputs.
N5XRASPT	An independent neural network predictor of ASPT (based on revised scores) having 5 inputs.
N7DNFAM	A dependent neural network predictor of NFAM having 7 inputs.
N7XNFAM	An independent neural network predictor of NFAM having 13 inputs.
NFAM	Number of BMWP families.
NNBMWP	A neural network trained on GQA class 'a' data (Section 3.2).
NNIFE614	A neural network trained on the IFE614 database (Section 3.2).
NNRSCR	A neural network trained on GQA (revised score) class 'a' data (3.2).
Organic River Quality	A river quality classification based upon the EQI(ASPT) component only of the GQA classification (Section 2.4).
PBLS	Percentage of pebbles in the substrate.
Pool	A site having ≥70% sand and silt in its substrate.
Revised scores	The revised BMWP family scores derived by Walley and Hawkes (1996, 1997).
Riffle	A site having ≥70% boulders and pebbles in its substrate.
RIVPACS	A computer package, originally developed for the prediction of the macroinvertebrate fauna of unpolluted running water sites in Great Britain, but since extended to provide river quality classifications via the use of EQIs.
RIVPACS III	The third, and current, version of RIVPACS.
SAND	Percentage of sand in the substrate.
SILT	Percentage of silt in the substrate.
SOM	An unsupervised neural network called a Self-Organising Map.
SOM10	A site specific SOM with a 10x10 output array (Section 4.2).
SOM20	A general SOM classifier with a 20x20 output array (Section 4.4).
State of existence	The absence (0) or abundance category (1-4) of a taxon.
Supervised learning	A training procedure for neural networks based on known target values.
Unsupervised learning	A training procedure for neural networks in the absence of known target values.
WIDTH	Average width of river (m).
X	Global easting of the National Grid Reference.
Y	Global northing of the National Grid Reference.

EXECUTIVE SUMMARY

This Technical Report presents the findings of a study into potential applications of Artificial Intelligence (AI) in the biological monitoring of river quality. It was carried out as the second stage of National R&D Project E1/i621 "Applications of Artificial Intelligence in River Quality Surveys".

Artificial Intelligence is defined as a discipline concerned with the building of computer programs that perform tasks requiring intelligence when done by humans. In this particular case, the task is the interpretation of biological data into river quality terms. When this is done by human experts it involves two complimentary mental processes: scientific reasoning and pattern recognition. Two AI techniques which are capable of modelling these processes, Bayesian reasoning and neural networks, were investigated. However, emphasis was placed on the development and testing of the neural networks, because the ready availability of the necessary data provided a unique opportunity to gain rapid results. The most advanced of the Bayesian methods requires elicitation of knowledge from experts, a process that can prove very time consuming. Thus the method investigated, known as naive Bayesian inference, was not the most advanced but had the benefit of being able to draw its 'knowledge' from the data. This permitted a rapid assessment of the potential of Bayesian methods to be made.

Two types of neural network were tested: supervised-learning networks, which require both input data and desired or target outputs; and unsupervised-learning networks, which require input data only. A commonly used supervised-learning network, known as back-propagation, was shown to slightly outperform RIVPACS in the task of predicting 'unpolluted' ASPT and NFAM, and an unsupervised-learning network, known as a Self-Organising Map (SOM), was shown to have considerable potential for the diagnosis of different types of pollution. Full details of the theory, development and testing of these networks are given in the report, together with details of several other networks that perform different tasks.

Bayesian classifiers of river quality were developed separately for spring and autumn samples. They provide classifications in probabilistic terms, and include a conformity index which is used to identify anomalies in the community composition. They have potential for use as diagnostic and quality assurance tools.

Information theory was used to define the 'indicator values' of the BMWP taxa in terms of the amount of information they provide about river quality class. Values were derived using presence/absence and abundance data, thus permitting: a) the taxa to be listed in rank order of their overall worth as indicators of river quality; and b) the added value gained from recording abundance to be quantified.

The report concludes that the AI methods tested have considerable potential for use in river quality classification and river management. The results will be of interest to those working in river quality management and environmental monitoring.

Keywords: artificial intelligence, AI, neural networks, Bayesian, information theory, river, pollution, RIVPACS, ASPT, biological monitoring, benthic, macroinvertebrates.

1. INTRODUCTION

1.1 Background to Study

The work described in this report stems from research carried out in the early 1990s by a small team of researchers at Aston University. The team was first established in 1989 when W. J. Walley, a specialist in artificial intelligence (AI), and H. A. Hawkes, an expert river ecologist, agreed to collaborate on the development of a biomonitoring system based upon AI techniques. They subsequently engaged two PhD students: M. Boyd to work on a knowledge-based (or expert systems) approach; and B. M. Ruck to take a neural networks approach. Thus the project developed along two parallel strands of research, one based upon methods of reasoning under conditions of uncertainty (Walley *et al.*, 1992a, 1992b; Boyd *et al.*, 1993; Boyd, 1995), and the other based on the pattern recognition capabilities of neural networks (Ruck *et al.*, 1993a; Ruck, 1995). Walley (1993, 1994) described the principles underlying both approaches and the progress made on each. In addition, a small project on the application of neural networks to predict benthic community structures in the Great Lakes was carried out in collaboration with the National Water Research Institute of Canada (Ruck *et al.*, 1993b; Ruck *et al.*, 1996). In 1993, a group of Slovenian AI researchers specialising in machine learning methods of rule induction approached the Aston team with a view to testing their techniques in the biomonitoring field. This resulted in a joint paper on the use of machine learning to classify river water quality (Džeroski *et al.*, 1994) and a further one that compared Bayesian, neural and machine learning methods of classification (Walley and Džeroski, 1995). Collaboration with the Slovenian group has continued, resulting in further publications on the application of machine learning techniques, mainly in relation to the induction of rules from bioindicator data (Džeroski *et al.*, 1996, 1997a, 1998).

This Environment Agency project (National R&D Project 621) has produced three AI-based publications (Walley and Fontama, 1997, 1998 and in press). To the authors' knowledge the papers cited above are the only publications to date on the application of AI-techniques to the biological monitoring of river[1] quality.

The need for the improvement of existing biomonitoring methods was demonstrated by a comprehensive reappraisal of the Biological Monitoring Working Party system based upon an analysis of data from the 1990 River Quality Survey of England and Wales (Walley and Hawkes, 1996, 1997), and later by a preliminary reappraisal of the saprobic system based on a similar analysis of Slovenian data (Džeroski *et al.*, 1997b).

1.2 Prior Knowledge

This project, which commenced on 1st September 1995, had the benefit of knowledge gained from earlier studies. This included knowledge gained:

- over many decades, as a result of the development of various biomonitoring systems, principally in Europe;
- from the development of the RIVPACS system; and
- from the development of the AI systems outlined above.

[1] Throughout the remainder of this report the term 'river quality' rather than 'water quality' has been used in recognition of the fact that environmental stresses other than the quality of water affect the biological community. For example, contamination of the substrate and engineering works.

De Pauw and Hawkes (1993) comprehensively reviewed several European systems, clearly outlining their main features, strengths and weaknesses, and Hawkes (1998) provided a detailed account of the history and development of the Biological Monitoring Working Party score system. The knowledge gained from these earlier systems stems more from the basic principles of biological monitoring than from the systems themselves, since the systems are all essentially simple *ad hoc* numeric or tabular algorithms. Early in the development of the AI approach to biomonitoring it was concluded that none of these *ad hoc* systems provided an adequate model of the expertise of river ecologists (Walley, 1993, 1994). The basic knowledge gained from these systems was that:

- the most suitable biota for use in biomonitoring are the benthic macroinvertebrates, for reasons that were concisely stated by De Pauw and Hawkes (1993);
- different taxa have different sensitivities to pollution and are therefore indicative of different river qualities;
- some taxa are more useful indicators of river quality than others;
- some taxa are naturally more abundant that others; thus the use of a single scale of abundance categories is not ideal;
- several factors other than river quality are important determinants of community composition, the principal ones being site type (i.e. eroding or depositing), geographic location and time of year; and
- the sampling procedure used significantly affects the recorded community composition.

The essential knowledge gained from the RIVPACS study was that:

- the community composition of unpolluted running-water sites can be predicted with a reasonable degree of accuracy from the geographical location and environmental characteristics of the site.

The additional knowledge gained from the AI-based studies mentioned earlier was that:

- experts use two complementary mental processes when interpreting biological data, plausible (or probabilistic) reasoning based upon their scientific knowledge of the ecological system and pattern recognition based upon their experience of past cases;
- the relationships between river quality and the occurrence of individual taxa (hence community composition) are inherently uncertain;
- the absence of a commonly occurring taxon gives useful evidence about river quality;
- different abundance levels are generally indicative of different qualities, thus abundance-based data are more discriminating than present-only data;
- high abundance levels are generally more discriminating than low ones; and
- the use of an inappropriate abundance scale for a given taxon will results in loss of information.

Walley and Fontama (in press) gave a comprehensive discussion and justification of these statements.

1.3 Appropriate Techniques

The prior knowledge outlined above has important implications for the selection of appropriate data interpretation techniques. Firstly, and most importantly, uncertainty in the

relationships between river quality and the states of existence of the taxa (i.e. different abundance levels, including absence) means that the evidence provided by a sample is not exact in its meaning but vague. That is, there is uncertainty in the meaning of the data, in addition to any uncertainty in its value. Thus, the principal requirement of an appropriate technique is that it should be capable of handling this uncertainty with minimal loss of information. This is not achieved by existing systems based upon scores, indices or look-up tables, because they make no attempt to account for variability in the meaning of the data. They effectively eliminate such variability by the use of averages or single-valued representative scores or indices, such as BMWP family scores, average score per taxon and saprobic indices. This results in the loss of valuable information and undermines the system's ability to reach a reliable conclusion. Secondly, the techniques chosen should be capable of modelling the plausible reasoning and/or pattern recognition processes used by experts. Thirdly, they should be capable of using the evidence given by different states of existence of the taxa in a way that permits discrimination between the different river qualities they represent. Finally, they must be capable of separating out the effects on community composition of factors other than river quality.

Artificial Intelligence covers a wide range of research topics, including natural language processing, neural networks, speech recognition, computer vision, expert systems and robotics. Two of these, neural networks and expert systems, are particularly relevant to the problem at hand. The pattern recognition capabilities of neural networks, and the probabilistic reasoning capabilities of Bayesian expert systems, offer the prospect of emulating the two mental processes used by experts. Both techniques effectively handle uncertainty in the meaning of the data, the first does so implicitly via its inherent abilities to generalise and to interpret input patterns as a whole in a non-linear way, and the second does so explicitly via probability theory. In addition, both techniques are able to give different meanings to evidence provided by different states of existence. As regards their ability to separate out the effects of factors other than river quality, this is clearly within their capabilities, but at the outset of the project it was not known how this would be best achieved or how successful it would prove to be.

There are other important factors supporting the use of neural networks in conjunction with probabilistic expert systems. Neural networks use historical data, whereas expert systems use scientific knowledge. Our industry is rich in both, so the application of these two techniques also permits maximal use to be made of available resources. Furthermore, a comparative study of Bayesian, neural and machine learning methods of classifying river water quality from biological data (Walley and Džeroski, 1995) concluded that the best performance was achieved by the Bayesian models, followed closely by the neural networks.

1.4 Neural Networks

Neural networks were originally devised as simple models of the structure and function of the brain, but some have now abandoned the original brain-based concept. What they all have in common is the ability to 'learn' from data. They can be classified into two types: networks that learn in a supervised way, and those that learn in an unsupervised way.

In supervised learning, the network is presented with many different examples of input data, together with their desired outputs. Thus, during the training phase the network is able to compare its outputs with the desired outputs to determine the magnitude of its errors. The

training algorithm enables the network to modify its internal parameters, so that next time the same examples appear its predictions are less erroneous. This process is repeated many times until the network's performance is maximised. The trained network can then be applied in practice to provide predictions and classifications for new cases.

In unsupervised learning the desired outputs are not given and the network 'learns' to classify the examples by recognising different patterns in the data, in a similar way to a child learning to recognise faces without the aid of a teacher. The ability to recognise different faces, however, has limited utility until one is able to attach names to the faces. Thus the benefit of not having to supply unsupervised networks with desired outputs is offset by the need to identify and label their output categories once training is complete.

This report explores potential applications of both types of network, and fuller descriptions of how they function are given in the relevant sections. However, those seeking a more detailed introduction to the theory and application of neural networks are referred to the introductory text by Beale and Jackson (1990) and the more advanced texts by Haykin (1994), Bishop (1995) and Kohonen (1995).

1.5 Probabilistic Reasoning

Traditional expert systems are based on classical logic and operate by the chaining of "If.....Then....." rules. These systems are perfectly adequate for problems involving exact relationships, but have serious weaknesses with respect to problems involving uncertain relationships. Under conditions of uncertainty it is necessary to employ one of the methods of 'inexact' or plausible reasoning, such as Dempster-Shafer theory of evidence, fuzzy logic or Bayesian inference (Giarratano and Riley, 1989). The latter has a long history but was originally considered to be computationally too demanding for use in complex knowledge-based systems. However, this problem has been overcome through the development of updating algorithms based on local computations within graphical representations of dependencies (Lauritzen and Spiegelhalter, 1988). Bayesian methods now provide the most powerful and consistent means of reasoning under uncertainty.

A Bayesian Belief Network (BBN) is an expert system in which the knowledge-base consists of two distinct parts: a causal network that defines the 'cause-effect' relationships between variables; and a set of conditional probability matrices that relate the state of each 'effect' (or child) variable to the states of its 'cause' (or parent) variables. Figure 1.1 shows the causal network of a simple BBN of river ecology. This is presented merely to illustrate the structure of the network, not as a valid working model, so the relationships may appear over-simplistic. Each node in the network represents a variable, and the arrows between variables represent causal relationships and the direction of causality (i.e. cause to effect). Each variable has a number of possible states (e.g. *Altitude* may be high, medium or low) and the likelihood of each, given the current state of evidence, is defined by a probability (or belief). The conditional probability matrices mentioned earlier define the probability of the variable being in each of its possible states, given the states of its parents (i.e. those variables at the tail ends of its incoming causal links). For example, if *Geology* had just two states, igneous and sedimentary, and *Distance from Source* had three states, short, medium and long, then the conditional probability matrix for *Altitude* would define the probability of each of its three states for each of the six possible combinations of *Geology* and *Distance from Source*. In the case of variables like *Geology*, that have no parents, their probability matrices reduce to a vector of prior probabilities. The conditional and prior probabilities may be derived

subjectively by elicitation from experts, or more objectively by data analysis if a suitable database is available.

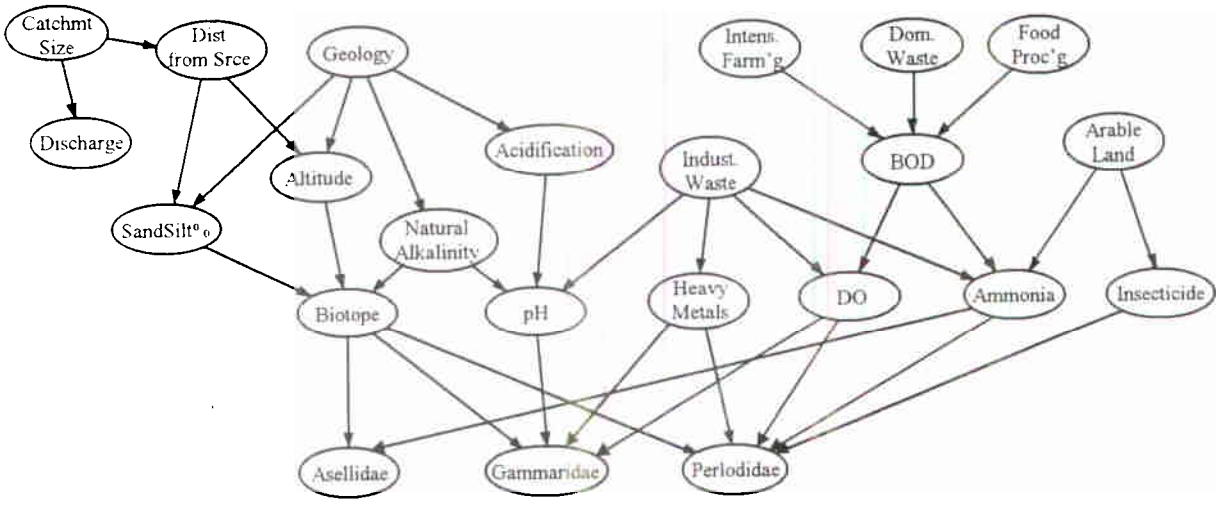


Figure 1.1 The causal network of a simple BBN model of river ecology.

Initially, when no evidence at all has been presented to the network, the beliefs in the states of all variables are equal to their priors (NB. the priors of child nodes are automatically derived from their conditional probabilities and the prior probabilities of their parents). When evidence is presented to the network, in terms of the observed state of one or more of its variables, the beliefs in the states of all the other variables are updated using algorithms that are soundly based in probability theory. This is probabilistic reasoning in its most advanced form.

Chong and Walley (1996) demonstrated the superiority of BBNs over rule-based systems. However, the development of BBNs that are capable of performing well on real-world problems is a time-consuming process requiring considerable knowledge elicitation and knowledge structuring. Like all knowledge-based systems they suffer from the so-called knowledge elicitation bottleneck. Readers seeking detailed accounts of the theory and application of BBNs should refer to Pearl (1988), Neapolitan (1990) and Jenson (1996).

If a problem can be expressed as a multi-valued hypothesis (e.g. several river quality classes) that can be determined from several conditionally independent items of evidence (e.g. the presence or absence of given biota), then it can be solved using 'naive' Bayesian inference. This assumes that the causal network consists of a single 'cause' linked to many 'effects', as illustrated in Figure 1.2. In this case the river quality is assumed to be the sole cause of the states (i.e. present or absent) of each of the taxa included in the model.

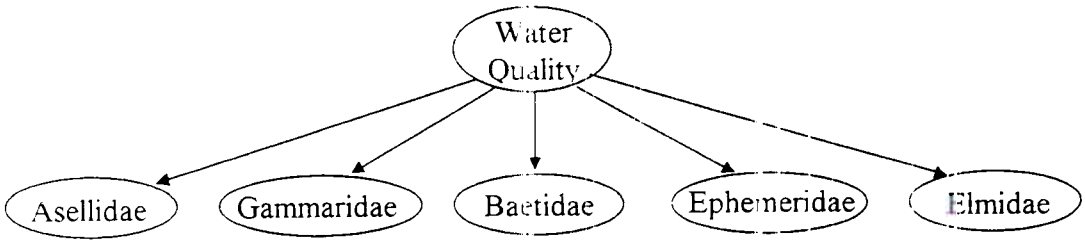


Figure 1.2 The causal network of a simple naive Bayesian classifier

This method can be used to classify river quality without the need to model the complex interdependencies of the ecological system, provided that:

a) other causal factors, like site type and season, can be effectively separated out, and;
b) the assumption of conditional independence of the evidence can reasonably be assumed to hold (i.e. independent within quality classes, not between them, e.g. it is assumed that Asellidae and Gammaridae are independent within any given quality class, but not independent across the classes).

These requirements may appear demanding, but they are the same assumptions that are implicit in the BMWP and saprobic systems. However, it is important to state them explicitly, so as to remain conscious of the limitations of the method.

The principal advantages of Bayesian methods are that they offer a mathematically sound and consistent way of reasoning under conditions of uncertainty and provide probabilistic outputs, thus indicating the degree of uncertainty in their conclusions. The models investigated in this study were based solely on naive Bayesian inference.

1.6 Outline of the Study

1.6.1 Objectives

The overall objective of the project was:

"To develop computer-based systems for the interpretation of biological data in river quality terms using AI techniques, and to assess their potential use as expert assistants and river quality classifiers on a national scale, in order to help the Agency determine whether this approach could fulfil its need for systems that analyse biological river quality data more fully."

This was subdivided into two specific objectives:

- to analyse data from the National Biological Database to help develop the AI systems and to provide distributions of taxa in geographical and river quality terms; and
- to develop and evaluate classification and interpretation systems based on artificial intelligence techniques so that the Agency can assess the utility of artificial intelligence for analysing biological river quality survey data.

Full details of the distributions of taxa in geographical and river quality terms were published in R&D Technical Report E12 (Walley and Martin, 1997). The results of the other data analyses carried out during the study are presented in this report.

Although the stated objectives of the study did not define the AI techniques to be used, the Technical Plan did refer to them. This required that the main emphasis of the study be placed on the use of neural networks, since these made best use of the vast amount of data available from the River Quality Surveys. That is, the study was intended to be data-based not knowledge-based.

It was also intended that the AI systems should not build on existing methods such as BMWP and RIVPACS, but should start afresh from basic AI principles. However, the work by Walley and Hawkes (1996, 1997) on the reappraisal of the BMWP scores resulted in a slight deviation from this intention. The Agency felt that there was a need for an investigation into the impact of the 'revised' scores on GQA classifications. This occurred at a time when

networks were being developed within the project to classify site type based upon the ASPT 'potential' of the site. It was realised that an opportunity existed to develop a neural network predictor of 'revised' ASPT, and hence a revised GQA classification, without impacting too much on the development of the new AI-based approach. This led to the development of neural network predictors of ASPT and number of families (NFAM), based on data supplied by the Institute of Freshwater Ecology (IFE). and subsequently a predictor of 'revised' ASPT.

1.6.2 Selected taxa and abundance scale

Table 1.1 lists the 76 BMWP families used in this study. Eight of the original BMWP families have since been split into separate families, but for the purpose of this report they were recombined into their original families. Consequently, each of these 'families' consisted of a group of families as detailed below:

Planariidae	-	includes Dugesiidae
Hydrobiidae	-	includes Bithyniidae
Ancylidae	-	includes Acroloxidae
Gammaridae	-	includes Crangonyctidae and Niphargidae
Dytiscidae	-	includes Noteridae
Hydrophilidae	-	includes Hydraenidae
Rhyacophilidae	-	includes Glossosomatidae
Psychomyiidae	-	includes Ecnomidae

Table 1.1 The 76 BMWP families used in the study

Planariidae	Gammaridae	Calopterygidae	Rhyacophilidae
Dendrocoelidae	Astacidae	Aeshnidae	Philopotamidae
Neritidae	Siphlonuridae	Corduliidae	Polycentropidae
Viviparidae	Baetidae	Libellulidae	Psychomyiidae
Valvatidae	Heptageniidae	Hydrometridae	Hydropsychidae
Hydrobiidae	Leptophlebiidae	Gerridae	Hydroptilidae
Lymnaeidae	Ephemerellidae	Nepidae	Phryganeidae
Physidae	Potamanthidae	Naucoridae	Limnephilidae
Planorbidae	Ephemeridae	Aphelocheiridae	Molannidae
Ancylidae	Caenidae	Notonectidae	Beraeidae
Unionidae	Taeniopterygidae	Corixidae	Odontoceridae
Sphaeriidae	Nemouridae	Haliplidae	Leptoceridae
Oligochaeta	Leuctridae	Dytiscidae	Goeridae
Piscicolidae	Capniidae	Gyrinidae	Lepidostomatidae
Glossiphoniidae	Perlodidae	Hydrophilidae	Brachycentridae
Hirudididae	Perlidae	Scirtidae	Sericostomatidae
Erpobdellidae	Chloroperlidae	Dryopidae	Tipulidae
Asellidae	Platycnemidae	Elmidae	Chironomidae
Corophiidae	Coenagriidae	Sialidae	Simuliidae

Three BMWP families (Lestidae, Pleidae and Hygrobiidae) were included in the 1995 survey but were not found in any samples. Three other families were removed from the list because they are not used by the Agency and have been excluded from RIVPACS III. These are Clambidae, Chrysomelidae and Curculionidae.

The abundance scale used throughout the project was based upon a logarithmic banding of the number of individuals found, as set out below:

0 - not found
1 - 1 - 9 individuals
2 - 10 - 99 individuals
3 - 100 - 999 individuals
4 - ≥ 1000 individuals.

1.6.3 Data validation and analyses

It was initially intended that the project would be based upon data from the 1990 River Quality Survey of England and Wales, but a delay in the commencement of the project made it possible to use the 1995 data. Thus all results presented in this report are based upon data from the 1995 River Quality Survey of England and Wales. For the purpose of this project, England and Wales was divided into ten regions[2] based upon the original ten administrative regions of the former National Rivers Authority (NRA), as it is in the Agency's national Biological Database. The relationships between these regions and the present eight administrative regions of the Environment Agency are given in Table 1.2.

Table 1.2 Relationship between the former ten NRA administrative regions used as the basis of this study and the present eight administrative regions of the Environment Agency.

No.	NRA Region	Present Environment Agency Administrative Region/Area
1	Anglian	Anglian Region
2	Northumbrian	Northumbria Area of North East Region
3	North West	North West Region
4	Severn Trent	Midlands Region
5	Southern	Southern Region
6	South West	Devon and Cornwall Areas of South West Region
7	Thames	Thames Region
8	Welsh	Welsh Region
9	Wessex	North Wessex and South Wessex Areas of South West Region
10	Yorkshire	Dales and Ridings Areas of North East Region

Although the 1995 database was known to be of higher quality than the 1990 database it still had to be validated to ensure that the data used were complete, representative and as error free as possible. The original database, prior to validation, contained 13,296 samples from

[2] The fact that the project was based on ten regions (i.e. the original NRA Regions) as opposed to the present eight Environment Agency Regions led to confusion over the use of 'Region' and 'region', because some of the projects regions were no longer administrative Regions. The convention adopted was to use 'Region' only when referring to the proper name of one of the present eight Regions of the Environment Agency (e.g. Midlands Region) or when referring to all eight collectively.

approximately 6700 sites. During the validation process 1,220 samples were removed, leaving 12,076 validated samples from 6038 sites (i.e. two samples, spring and autumn, from each site). A brief description of the validation process and reasons for removing sites is given in R&D Technical Report E12 (Walley and Martin, 1997). A comprehensive account of the data validation exercise is given in the Project Record (Walley and Martin, 1998).

The data analyses carried out on the validated data in relation to the distributions of taxa reported in R&D Technical Report E12 (Walley and Martin, 1997) were:

- geographical distributions of sampling sites;
- geographical distributions of taxa by occurrence and abundance category; and
- frequency distributions of taxa by river quality and abundance category.

Data analyses and manipulations carried out on the validated data in relation to the development of the AI systems described in this report were:

- derivation of indicator values of taxa based on information theory;
- construction of various input vector files based on different groupings of taxa;
- construction of a database of exemplars based on an analysis of site types;
- construction of a database of matched chemical and biological sites;
- analysis of the distribution of sites by EQI(ASPT) and EQI(NFAM); and
- derivation of the conditional probabilities need for the Bayesian classifier.

In addition, many different analyses were carried out on the results produced by the various models. These are detailed in the sections where the models are described.

1.6.4 Subjectivity and the need for exemplars

One problem that arises in the development of any system to classify river quality in biological terms is the lack of a well-defined standard set of quality classes. River quality is so complex and multi-dimensional that any attempt to express it in meaningful and readily understandable general terms must involve a degree of subjectivity. Thus, the real problem facing us is how to minimise the effects of subjective errors on the performance of the systems. That is, how can we best handle the subjective aspect of the problem to produce the most meaningful and consistent classification system? We can attempt to minimise its extent and maximise its quality, and we can examine the stage at which it enters the process, so as to avoid building 'objective methods' on subjective foundations.

The Biological Monitoring Working Party (BMWP) system is based upon family scores that were allocated subjectively by a committee of experts. Thus, this system is *founded* on subjectivity, albeit the best available expert opinion at the time (1976-78). Hawkes (1998) gave a comprehensive account of the development of the system, the problems faced and the decisions made. A computer-based reappraisal of the BMWP scores (Walley and Hawkes, 1996, 1997) demonstrated how the accuracy of these subjectively derived values could be enhanced through an analysis of field data.

An alternative approach to that taken by BMWP could be to use a committee of experts to classify a large set of biological field samples into their various river quality classes. To achieve this the committee would first have to develop an agreed overview of the ecological nature of each river quality class. They would then have to classify a large number of samples

covering all quality classes and a wide range of site types. However, once the experts had agreed this set of examples (i.e. a database of exemplars), it could then be used as the standard by which biological river quality is defined. An advantage of this approach is that the subjective component defines a set of end-points, which are then used as targets for the classification systems, not the foundation on which they must build. This is precisely what is needed for the training of supervised-learning neural networks (i.e. a training set consisting of the desired outputs corresponding to a set of example inputs). On the other hand, unsupervised-learning neural networks do not require desired outputs during their training phase. They simply learn to sort the input examples into groups based upon the patterns in their data, leaving the experts to label the groups into river quality terms at the end of the training process. Once again, the subjective judgements provided by the experts are used to define end points. However, some subjectivity still remains at the beginning of the modelling process, the subjectivity relating to model selection (e.g. the type of network, its topology and training algorithm). The effect of this can be minimised by optimising the model's performance on a set of exemplars, or alternatively, by maximising its mutual information value.

In this study, a database of exemplars was needed for the training of the river quality classifier based on supervised-learning neural networks. It was also required for the derivation of the conditional probabilities for the Bayesian models, because the data-based approach to the project required that these be derived by data analysis, not through knowledge elicitation.

The original intention was that a database of exemplars would be constructed by asking Agency biologists to classify their sites subjectively, without reference to their GQA classification, using their own knowledge and understanding of the ecology of the sites. However, this proved to be impracticable within the time-scale of the project. Thus, it was decided to construct the database using classifications derived by two existing methods, but after removing all dubious classifications. Full details of how this was done are given in Section 2.4 (Database of Exemplars). The justification for using existing classifications, some of which may have been incorrect, was that both the neural and Bayesian systems were capable of handling noisy data, and would not be unduly disturbed by a few remaining erroneous classifications. Neither of the two methods depends upon every desired output being absolutely correct, provided the database is fairly large and the proportion of errors is relatively small and well distributed across the classes.

1.6.5 Indicator values of taxa

It has long been known that some taxa are better indicators of river quality than others. Indeed, the saprobic system uses indicator values as weighting coefficients in the mathematical formulation of the saprobic index. These weights are determined from the shape of the subjectively derived distribution of saprobic valency, using rules proposed by Sládeček (1964). A broadly tolerant species with a flat distribution is given an indicator value (or weight) of unity, whereas a species which is highly specific, appearing in only one of the five saprobic classes, is given an indicator value of five.

The BMWP system makes no use of indicator values, but Walley and Hawkes (1997) suggested some modifications to the Biological Monitoring Working Party score system that incorporated abundance rating, biotope type and indicator value. They defined indicator value in terms of the inverse of the standard deviation of the taxon's distribution with respect to the

Average Score Per Taxon (ASPT) of the sites at which the taxon occurred. This, in effect, produces similar results to Sládeček's rules.

At the outset of this project it was decided that there was a need for a definition of indicator value, which truly reflected the information value of the evidence provided by the taxon. It was felt that this would prove valuable when attempting to optimise the input vectors to the neural networks. Indicator values based upon the mutual information between the biological GQA class and the state of existence of the taxa were derived using present-only and abundance data. These were subsequently used as the basis of an investigation into how the size of the input vector to the neural networks might be minimised without jeopardising their performance. The derivation and analysis of these indicator values are fully described in Section 2.1, and a detailed account of the tests on the input vector is given in Section 2.2.

1.6.6 Overview of models developed

The models developed in this project fall into three categories:

- supervised-learning neural networks;
- unsupervised-learning neural networks; and
- naive Bayesian classifiers.

Extensive exploratory tests were carried out to determine which of the various supervised-learning and unsupervised-learning neural networks were best suited to the tasks at hand. In addition tests were carried out to determine the best configuration of each type. These investigations are described in the introductions to Sections 3 (Supervised Neural Networks) and 4 (Unsupervised Neural Networks).

The models that were developed were:

Supervised-learning networks

- Classifiers of site type
- Predictors of ASPT and NFAM
- Predictors of BOD, DO and ammonia
- Classifiers of biological river quality

Unsupervised-learning networks

- Classifiers of biological river quality

Naive Bayesian Classifiers

- Classifier of biological 'organic' river quality

Four of these were investigated in greater detail than the others, namely: the predictors of ASPT and NFAM; the unsupervised classifiers of biological river quality; and, to a lesser degree, the classifiers of site type and the naive Bayesian classifiers.

1.7 Operational Value of the Study

The efficient and reliable interpretation of field data is of prime importance to the Agency. Each year vast amounts of data are collected and processed at considerable expense. Thus, it

is desirable to use the best available techniques for its analysis and interpretation to ensure that as much useful information as possible is extracted from it. Prior to this project the Agency had not investigated the use of AI methods. The results show that at least two AI techniques, neural networks and Bayesian reasoning, offer considerable potential for use in an operational setting. In addition, the project has produced some valuable spin-off benefits, and several of the outputs will prove useful to operational staff. These include:

- the identification / correction of errors in the GQA databases, especially the 1995 database;
- maps showing the geographical distribution of the BMWP taxa based (for the first time) on average abundances (R& D Technical Report E12);
- histograms showing the frequency of occurrence of the BMWP taxa by river quality and abundance category (R& D Technical Report E12);
- rankings of BMWP taxa in terms of their information value (or indicator value) with respect to river quality (2.1 and Appendix A);
- evidence showing which additional taxa could usefully be added to the BMWP list (2.1.3);
- the demonstration of how abundance data (currently collected but not used) can be used to improve the information values of the BMWP taxa (2.1.3) and hence the performance of classification / diagnostic systems;
- new information on the relationship between taxa and environmental variables (3.3.4 - 3.3.8);
- results which indicate that the BMWP system could benefit from the use of the revised scores derived by Walley and Hawkes (1996, 1997) (3.3.9);
- results which give an indication of the overall reliability of GQA classifications (3.3.10);
- a computer package called SOMVIEW (available on disk or via the Web at http:// www.soc.staffs.ac.uk/research/groups/cies/somview/somview.htm) that enables users to view any two SOM feature maps, and hence to visually explore the relationships between them in data space; and
- a computer package called RBMS (available free to Agency staff, although not a contract deliverable) which provides user-friendly access to the national biological database (validated sites only), plus the GQA biological classifications, various alternative classifications produced by the project and some useful analytical tools (NB. This is the first time that data from all Regions have been made readily available to users).

These outputs provide operational staff with more reliable data, some useful analytical tools and the means of gaining further insight into the relationships between the taxa, river quality and environmental variables. If operational biologists make full use of these outputs, their abilities to interpret their field data will be enhanced.

It is anticipated that further developments of this work, which are now being pursued via a new R&D contract, will produce two river quality diagnostic systems: one based on an improved version of the SOM neural network, and the other based on Bayesian Belief Networks. These will be available for field testing late in 1999.

2. DATA ANALYSES AND MANIPULATION

2.1 Indicator Values

2.1.1 Background

This study commenced before valid abundance-based data had been acquired from the North West, Midlands and Welsh Regions and before the development of the site-type classifiers (Section 3.2). Thus the original analysis was based on a database containing data from only seven of the ten former NRA regions listed in Table 1.2, and a method of classifying sites into *Riffles* and *Pools*[3] first used by Walley and Hawkes (1996, 1997), as defined below.

Site Type	Nature of Substrate	Number of Samples
Riffles	≥ 70% boulders and pebbles	7416
Pool	≥ 70% sand and silt	2174
Riffle/Pool	neither *Riffle* nor *Pool*	2488

Indicator values were only derived for sites classified as either *Riffles* or *Pools*.

The analysis was repeated later using validated data from all ten regions. The same definitions of site types into *Riffles* and *Pools* were used, because a change to the site-types derived in this study (Section 3.2) would have required major changes to the software and the data files. The results of these analyses are given in Appendix A.

The term 'indicator value' is generally used rather loosely. This can lead to confusion since there are two very different interpretations. The term is often used to represent the value of a taxon as an indicator of river quality, when it is found in a sample. This, in fact, is a conditional indicator value, since it only applies when it is known that the taxon is present. Thus, by this definition, a very rare taxon can have a high indicator value. Such indicator values provide an appropriate means of weighting the evidence given by a taxon when found in a sample. However, if one is trying to rationalise a taxonomic list for survey purposes, or developing a data interpretation system based upon several possible states of existence of the taxa, including absence, then the most appropriate definition is one based on the information value of each state of existence weighted in proportion to its probability of occurrence. This provides an overall (or unconditional) measure of the utility of each taxon as a sensor of river quality. Thus a rare taxon that is indicative of a particular river quality will score lowly on this measure, because its most commonly occurring state (i.e. absent) tells us nothing about the quality of the water. A commonly occurring taxon that is indicative of a particular river quality scores highly because it frequently provides valuable information via its occurrence, and even its absence provides useful information.

2.1.2 Mathematical formulation

The mutual information $M(C,X)$ is a well-recognised measure of the common information shared by two variables C (class) and X (attribute). In our case, C represents six river quality classes (i.e. biological GQA classes a-f, numbered 1 to 6 here) and X represents five possible states of existence (i.e. abundance categories 0 - 4) of a given taxon.

[3] The terms *Riffles* and *Pools* are used throughout this report to mean river sites having substrate compositions as defined by Walley and Hawkes. They should not be taken to mean riffles and pools as generally understood be river ecologists , although the two meanings are clearly closely related.

Now, if:

n_{ij} = number of samples of the i th class in which the given taxon is in the j th state;

N_i = total number of samples of the i th class;

M_j = total number of samples in which the given taxon is in the j th state; and

T = total number of samples,

then the mutual information between C and X is given by:

$$M(C,X) = \sum_{ij} p_{ij} \log\left(\frac{p_{ij}}{q_i r_j}\right) \tag{1}$$

where:

$p_{ij} = \frac{n_{ij}}{T}$ = probability of observing the given family in the i th class and the j th state;

$q_i = \frac{N_i}{T}$ = probability of the i th class; and

$r_j = \frac{M_j}{T}$ = probability of the given family being in the j th state.

However, this formulation is not ideal for a new definition of the indicator values of bioindicators, since the derived values of the probabilities are dominated by data from the most commonly occurring classes. Since indicator values are intended to represent how well each taxon discriminates between the different classes, it is important that the data are not weighted towards any particular class. This can be achieved by defining the probabilities in a way that gives equal weight to all classes, thus:

$$p'_{ij} = \frac{n_{ij}}{6N_i} \qquad q'_i = \frac{1}{6} \qquad r'_j = \sum_{i=1}^{6} p'_{ij}$$

This is equivalent to applying the Principle of Indifference and produces a set of probabilities that can be used to define indicator values based upon what we have called the *Indifferent Mutual Information* value, $M'(C,X)$, defined as:

$$M'(C,X) = \sum_{ij} p'_{ij} \log\left(\frac{6p'_{ij}}{r'_j}\right) \tag{2}$$

Riffle and *Pool* data were used to derive $M'(C,X)$ values for all 76 taxa, based on abundance data and present-only data. In the latter case, equation (2) had to be modified to use just two states of existence (i.e. present and absent) instead of five. The abundance-based values, $M'_a(C,X)$, and the present-only values, $M'_p(C,X)$, were used to define an improvement ratio (R_j), thus:

$$R_j = \frac{M'_a(C,X)}{M'_p(C,X)} \tag{3}$$

This ratio represents the benefit gained by recording the abundance of the taxon as opposed to just its presence. A value of 1.0 represents no benefit at all, whereas a value of 1.5 represents a 50% increase in information value.

Riffle and *Pool* $M'(C,X)$ values were also derived for 17 non-BMWP taxa. However, not all Regions recorded these taxa so the results are based upon data covering only part of England and Wales. Two of these taxa produced $M'(C,X)$ values that would have placed them in the top 50 indicator taxa for both *Riffles* and *Pools*. They were Hydracarina and Ceratopogonidae, which would have been placed 34th and 30th respectively in the *Riffle* list and 29th and 45th in the *Pool* list. Thus, if new taxa are to be added to the BMWP list at a future date, then these two should be prime candidates, at least from the information theory viewpoint.

2.1.4 Results of regional analysis

In order to examine what variation might exist in the information values of individual taxa between regions, $M'(C,X)$ values based on abundance data were derived for each of the ten regions. The results are given in Tables A3 for *Riffle* sites and Table A4 for *Pool* sites. Both tables also give the average and standard deviation of the $M'(C,X)$ values taken across all ten regions, plus the national $M'(C,X)$ value (as given in Table A1 and A2) and the ratio between the regional (average) and national values. The results for North East (Northumbria) and South West (Devon and Cornwall) in Table A4 are given in italics, because there were so few pool sites in these regions that the derived statistics were considered unreliable. Thus they were excluded from the calculation of regional average $M'(C,X)$ values for *Riffles*. Tables A3 and A4 reveal that almost all taxa have a higher regional average $M'(C,X)$ value than national value. This is because the regional analysis eliminates:

a) more of the site-type effect; and
b) part of the national variation in the mix of species within families.

The list of ratios of regional to national values indicates that some taxa produce much higher ratios than others. For example, Oligochaeta (ratio = 3.6 in *Riffles* and 3.9 in *Pools*), Chironomidae (3.3 and 3.9), Asellidae (2.0 and 4.2), Glossiphoniidae (2.9 and 1.6), Erpobdellidae (3.9 and 2.9), Sphaeriidae (1.9 and 1.6) and Lymnaeidae (3.3 and 2.2). High values are also produced by many of the less common taxa, but these are considered unreliable, due to distortions caused by very small sample sizes in some regions. A high ratio implies that the various states of the taxon are indicative of different river qualities in different regions. This is most likely to occur where a taxon is represented by different species in different regions. Conversely, a ratio around unity indicates that the taxon is a fairly consistent indicator across the whole country (e.g. Elmidae with 0.967 for *Riffles* and 1.047 for *Pools*). There were, as expected, very few taxa with ratios noticeably less than unity (i.e. 7 taxa in *Riffles* and 3 taxa in *Pools* had ratios < 0.9), the lowest being 0.812 for Caenidae in *Pools* and 0.844 for Gyrinidae in *Riffles*. An ecological explanation for these low values has not been found, but it is possible that they were due to sampling error, since some of the regional samples were fairly small.

The rank order of the taxa based upon regional average $M'(C,X)$ differs from that based upon national $M'(C,X)$. The most significant changes from national to regional ranking are as follows.

Riffles (upgradings): Gammaridae (from 12th to 4th), Simuliidae (19th to 10th), Hydrobiidae (21st to 11th), Asellidae (25th to 12th), Ancylidae (23rd to 14th), Oligochaeta (36th to 16th) and Glossiphoniidae (33rd to 18th).

Riffles (downgradings): Leuctridae (6th to 13th), Lepidostomatidae (8th to 19th), Perlodidae (11th to 21st), Gyrinidae (10th to 23rd) and Ephemerellidae (15th to 25th).

The above formulation was designed for use on the national database, where *Riffle* and *Pool* sites are represented in all six river quality classes. In some of the regional databases, however, *Riffle* and *Pool* sites are not found in all six quality classes. Thus the equations were modified to account for the reduced number of classes, wherever this occurred. The benefit of splitting the analyses between *Riffles* and *Pools* was that the mutual information between C and X was enhanced, because the split partially removed the site type effect.

2.1.3 Results of national analysis

The mutual information, $M(C,X)$, and indifferent mutual information $M'(C,X)$ were derived for each taxon based on the validated national dataset, using both abundance data and presence/absence data split between *Riffles* and *Pools*. Although it was stated earlier that $M(C,X)$ is not the appropriate measure to use as an indicator value, its values were derived to provide a baseline against which to compare the proposed indicator values, $M'(C,X)$. In addition, *Riffle* and *Pool* $M'(C,X)$ values were derived for each of the ten regions of the former NRA (as defined in Table 1.2) using abundance data only. The results are given in Appendix A.

Tables A1 and A2 give the overall $M(C,X)$ and $M'(C,X)$ values of 76 BMWP taxa for *Riffles* and *Pools* as derived from nation-wide 'present-only' and 'abundance' data. The tables also give the improvement ratio, which provides a measure of the added benefit gained by using abundance data as opposed to present-only data. The taxa are listed in order of their abundance-based $M'(C,X)$ value. That is, they appear in order of their value as indicators of river quality if their abundance levels are recorded and used in the classification. Inspection of the *Riffles* and *Pools* lists reveals a very different ordering of the taxa in each. Only six of the top twenty taxa are common to both lists, namely Elmidae, Leptoceridae, Baetidae, Caenidae, Gammaridae and Ephemeridae, the first five of which are the overall top five indicator taxa listed in rank order. It is interesting to note that the average $M'(C,X)$ values of all 76 taxa listed in Tables A1 and A2 are 0.0746 and 0.0610 for *Riffles* and *Pools* respectively. The equivalent figures based on the top twenty taxa in each site type are 0.1948 and 0.1436 respectively. Thus, samples taken from *Riffles* are about 25% to 35% better (i.e. in information terms) at discriminating between river qualities than those taken from *Pools*. This confirms in quantitative terms what river ecologists have known for a long time.

Another interesting point to note is that a few taxa are markedly better indicators when their abundance, as opposed to just their presence, is recorded. Oligochaeta was exceptional in this respect, showing nine-fold and twenty-fold increases in *Riffle* and *Pool* $M'(C,X)$ values respectively when derived from abundance data. Other notable taxa were Chironomidae (69% and 109% increases respectively), Asellidae (62% and 67%), Erpobdellidae (52% and 26%), Physidae (88% [relatively small sample] and 10%), Sphaeriidae (13% and 21%), Lymnaeidae (14% and 16%), Gammaridae (14% and 15%) and Baetidae (20% and 8%). The lack of noticeable improvements by the other taxa was probably the result of the $\log_{10}$-based abundance scale being too coarse for their particular cases. The fact that most of the high improvement ratios (Tables A1 and A2) were achieved by taxa that naturally occur in high numbers implies that the $\log_{10}$-based scale is better suited to these taxa than the naturally less numerous taxa. Walley *et al.* (1992b) used five different abundance scales to maximise the information input to their Bayesian classifier, thus ensuring that each taxon was represented by an appropriate scale. If the Agency were to adopt a finer scale for the naturally less numerous taxa it would undoubtedly draw out more useful information from these taxa.

Pools (upgradings): Gammaridae (6th to 1st), Sphaeriidae (7th to 3rd), Asellidae (32nd to 7th), Oligochaeta (33rd to 9th), Lymnaeidae (20th to 10th), Chironomidae (36th to 13th) and Erpobdellidae (31st to 16th).

Pools (downgradings): Dytiscidae (13th to 22nd), Coenagriidae (9th to 23rd) and Calopterygidae (12th to 28th).

Table A5 shows the top 12 *Riffle* taxa in each of the ten regions listed in order of the *M'(C,X)* values. Eight of the ten regions are very similar in terms of the rank order of their top taxa. Elmidae ranks first in all eight, and Hydropsychidae ranks either second or third in seven of the eight, the odd one out being Southern Region were it ranks 11th. In the two regions where Elmidae does not rank first (i.e. Anglian and Thames), Gammaridae is the leading taxon. Gammaridae also ranks highly (3rd) in Southern Region. Thus visual inspection of these lists indicates that Anglian, Thames and Southern are somewhat different from the other regions in terms of their top indicator taxa for *Riffle* sites.

Table A6 lists the top *Pool* taxa in eight of the ten regions. The other two regions, North East (Northumbria) and South West (Devon and Cornwall) had too few *Pool* sites to produce reliable estimates of *M'(C,X)*. These lists show far less similarity between regions than was the case for *Riffles.*

The top taxon overall, in terms of average regional ranking position, was Gammaridae, followed by Leptoceridae, Baetidae, Hydrobiidae and Elmidae in rank order.

In order to compare the regions in terms of their overall information values of the taxa, the average *M'(C,X)* of each region's top 40 taxa was determined. The highest average was achieved by Thames Region in the case of *Riffle* sites (value = 0.201) and Midlands Region in the case of *Pool* sites (value = 0.186). Table 2.1 below gives the values for the other regions expressed as a percentage of these maximum values. The variations from region to region may be due to several factors. Firstly, the existing classification system may suit some regions better than others and/or one site type better than another. For example, Midlands Region performs well with respect to *Pool* sites but poorly with respect to *Riffle* sites, whereas in the case of Thames Region it is *vice versa.* The sites in some regions may be better suited to biomonitoring than in other regions, and there may be differences in the standard of sampling and analysis procedures. It is not possible at this stage to identify the most likely cause of these variations, but they are interesting to note and may be worthy of further investigation.

Table 2.1 Average indifferent mutual information values, *M'(C,X)*, of the top 40 taxa for *Riffle* and *Pool* sites, expressed as a percentage of the maximum for the given site type.

Site Type	Ang	N.East Nrthm	NWest	Mid	Sthn	SWest D&C	Thms	Welsh	SWest Wssx	NEast D&R
Riffle	97%	89%	70%	67%	68%	98%	100%	76%	97%	78%
Pool	88%	61%	N/A	100%	88%	N/A	80%	71%	82%	85%
Avg	92%	75%	N/A	84%	78%	N/A	90%	73%	90%	81%

2.2 Input Vectors

In view of the large number of potential inputs to the proposed models (i.e. 76 taxa), it was decided to investigate ways of combining or eliminating taxa without jeopardising the performance of the models. In fact, it was thought that combining rare taxa into groups based upon their pollution sensitivities might even improve performance. Thus, the main purpose of the investigation was to maximise performance without over-parameterisation of the models. The results of the study of indicator values provide a means of ranking the taxa in terms of their information value, and the revised BMWP scores derived by Walley and Hawkes (1996) provided a means of grouping taxa into sensitivity bands. Various input vectors were tested on supervised and unsupervised neural networks in an attempt to determine an optimum input configuration. These include:

- 76 inputs comprising all 76 BMWP families, using their abundance categories as inputs;
- 77 inputs comprising all 76 BMWP families (as above) plus NFAM, the number of families present;
- 50 inputs comprising the top 50 BMWP families, based on their national indifferent mutual information values;
- 51 inputs comprising the top 50 BMWP families plus NFAM;
- 21 inputs based upon taxonomic groups of families[4], using the total of their abundance categories as the input values;
- 21 inputs based upon taxonomic groups of families, using the number of families present in each group as the input value; and
- 16 inputs based upon groups of taxa, each having similar sensitivities to organic pollution as indicated by their overall revised BMWP scores (Walley and Hawkes, 1996), and using the total of their abundance categories as the input value.

It was hoped that one of the vectors based upon groups of families would perform as well, if not better, than the full vector of 76 families. It was expected that the vector comprising the 16 sensitivity groups would perform particularly well, but this was not found to be the case. For the unsupervised networks, the best input vector was found to be the one consisting of 76 BMWP families, whereas for the supervised networks the best was the 77 input vector consisting of 76 BMWP families plus NFAM. This does not however mean that these are the most cost-effective input vectors. It may well be possible to achieve 95% of optimum performance using 50% percent of the taxa, but optimisation based upon benefit / cost ratio was not the purpose of the exercise.

[4] The 21 groups consisted of Flatworms, Mollusca (3 groups), Worms, Leeches, Crustacea (2 groups), Mayflies (3 groups), Stoneflies, Damselflies, Dragonflies, Bugs, Beetles (2 groups), Caddisflies (2 groups), Chironomidae and one miscellaneous group (consisting of Tipulidae, Simuliidae and Sialidae). Where major groups were subdivided, this was achieved by clustering the taxa on the basis of their overall revised scores.

2.3 Distribution of Sites by River Quality Class

The distribution of the 1995 validated sites by biological GQA class is given in Table 2.2 below.

Table 2.2 Distribution of the number of sites by biological GQA class

	Biological GQA Class						Total No. of Sites
	a	b	c	d	e	f	
Number of Sites	1762	1747	1271	638	488	132	6038
Percentage of Total	29.2%	28.9%	21.0%	10.6%	8.1%	2.2%	100%

The rules used by the Environment Agency when allocating sites to these GQA classes were based on the threshold values for EQI(ASPT) and EQI(NFAM) given in Table 2.3 below:

Table 2.3 Threshold values of EQIs for the allocation of GQA classes

GQA Class	EQI(ASPT)	EQI(NFAM)
a	1.00	0.85
b	0.90	0.70
c	0.77	0.55
d	0.65	0.45
e	0.50	0.30
f	-	-

A site is allocated to the highest class in which its EQI(ASPT) and EQI(NFAM) equal or exceed the stated threshold values for the class.

Table 2.4 gives a breakdown of the 1995 biological GQA classification into its EQI(ASPT) and EQI(NFAM) component parts. The totals in the bottom row show the distribution between the classes if the classification was based on EQI(ASPT) alone, and the totals in the right-hand column show the distribution based on EQI(NFAM) alone. The matrix gives the overall distribution with respect to these two component parts. For example, 15 sites achieved a 'b' classification on the basis of their EQI(ASPT), but only a 'd' classification on the basis of their EQI(NFAM). Since 'd' is the lower of the two, their overall biological GQA class is 'd'. Thus, a low EQI(NFAM) caused their EQI(ASPT) class to be downgraded by two class intervals. This is not exceptional since several sites were downgraded by three or more class intervals due to either a low EQI(ASPT) or a low EQI(NFAM). For example, three were downgraded from 'a' to 'd' due to low EQI(NFAM)s and, conversely, 12 sites were downgraded from 'a' to 'd' due to low EQI(ASPT)s. All 15 sites were thus given the same overall biological GQA classification of 'd', despite being represented by two very different biological communities, one indicative of organic pollution, (i.e. low EQI(ASPT) and high EQI(NFAM)) and the other indicative of toxic pollution (i.e. high EQI(ASPT) and low EQI(NFAM)).

Although these may be considered to be of equal severity in terms of environmental stress, it is desirable to label them in a way that preserves their different identities. For example, the first three sites could be labelled 'ae' to indicate high ASPT but low NFAM, and the other 12

could be labelled 'ea' to indicate low ASPT and high NFAM. This would in effect produce a more detailed classification system in which both the organic and toxic dimensions of river pollution are represented. The problem that the existing classification system presents to this study, especially the naive Bayesian classifiers, is that it results in examples of the lower quality classes (d-f) being represented by markedly different biological communities, even within a given site type and river quality class.

Table 2.4 Relationship between river quality classifications based upon EQI(ASPT) only and EQI(NFAM) only

GQA Class based on EQI(NFAM) only	GQA Class based on EQI(ASPT) only						Total
	a	b	c	d	e	f	
a	1762	1269	267	12			3310
b	140	338	460	61	1		1000
c	39	129	376	258	19		821
d	3	15	84	205	84		391
e	3	10	25	143	203	5	389
f		2	1	18	71	35	127
Total	1947	1763	1213	697	378	40	6038

2.4 Database of Exemplars

The main purpose in constructing this database was to provide data for the training and testing of supervised-learning neural networks and the derivation of conditional probabilities for the naive Bayesian classifiers. These models required both inputs and target outputs (i.e. river quality classes). The problem was how to derive target outputs, bearing in mind:

a) the undesirability of using the GQA classifications, since these were dependent upon existing (imperfect) systems that could not be considered to be absolute standards; and
b) complications concerning the 'organic' and 'toxic' components of the GQA classification highlighted above.

It was subsequently decided to use the EQI(ASPT) component of the GQA class as the basis of an 'organic' classification, which was derived using two different procedures in an attempt to reduce its dependence on existing methods. The reasons for defining the targets in this way are listed below.

- Errors in both the observed and predicted values of NFAM are so great as to make EQI(NFAM) unreliable. This is supported by:
  a) the 1995 audit of samples by IFE, which showed an average of 1.88 gains and 0.22 losses per sample taken (NB. Walley and Martin, 1997, contains a summary of gains and losses by region); and
  b) the results of tests carried out during the development of the neural network predictors of ASPT and NFAM as part of this project (Section 3.3.7), which showed that predictions of NFAM are far less reliable than those of ASPT.
- The combination of EQI(ASPT) and EQI(NFAM) to form one GQA classification results in unnecessary confusion in the relationships between community composition and river

quality class. This would have undermined the performance of the naive Bayesian classifiers and, to a lesser degree, the supervised-learning neural networks.

Thus it was thought better to separate the components into their 'organic' (based on ASPT) and 'toxic' (based on NFAM) parts, and to use only the former, being the reliable part, as the target river quality classifications in the database of exemplars.

It should be noted that the decision to split the GQA quality classification into its two component parts was only relevant to the supervised-learning networks and the derivation of conditional probabilities for the Bayesian classifiers. The unsupervised-learning networks did not require target classifications and were therefore unaffected by this decision.

The first stage in the construction of the database of exemplars was to combine the taxonomic strings of the spring and autumn samples. This was achieved by taking the highest of each family's two abundance levels ('absence' being treated as zero). The next stage was to derive reliable site-type and 'organic' river quality classifications for each site.

Site type was classified into five ASPT bands (labelled 1-5) based upon three different methods of predicting unpolluted ASPT, two based on neural networks (NNRSCR & NNIFE614 - see Section 3.2 for details) and the other on RIVPACS III The final site-type classification was allocated by 'majority vote' of the three methods. In the rare cases where none of the three methods agreed, a fourth method (based upon neural network NNBMWP) was introduced to break the tie. Section 3.2 gives a detailed explanation of the site types and the reasons for classifying them. The geographic distributions of the 6038 sites by site type are shown in Figures B1 to B5 in Appendix B.

In the case of 'organic' river quality, the final classification required complete agreement between two methods based on EQI(ASPT) bandings. One used predicted ASPT derived by RIVPACS III and the other used values derived by a neural network predictor of ASPT (called NX5ASPT - see Section 3.3.5). If the classifications given by the two methods disagreed then the site was rejected from the database of exemplars. Both methods were based upon original BMWP scores, not revised scores, since appropriate data on the latter were not available at the time.

The final number of sites remaining in the database of exemplars was 4960, of which site types 1, 2, 3, 4 and 5 contributed 1048, 1001, 955, 996 and 960 respectively.

2.5 Database of Matched Biological and Chemical Sites

In order to develop the predictors of BOD, DO and ammonia it was necessary to link data on these three chemical variables to the validated biological data. This required the matching of chemical sampling sites to biological sampling sites. Since many 'matched' chemical and biological sites had different Ordnance Survey grid references, software was developed to calculate their distance apart. Analysis of these distances uncovered many grid reference errors, and these were either corrected or the sites removed from the database of matched sites. To ensure that the database only included valid matches, all pairs of sites that were greater than 400 metres apart were excluded, with the exception of sites in North East (Northumbria) where a threshold of 1 km was used. The reason for this exception was that the region had

very few matched sites that were less than 400 metres apart. It was thought desirable to increase the threshold in this case to avoid under-representation of the region in the database.

Unfortunately, many biological sites did not have a matching chemical site, and the final database consisted of just 3556 matched sites, which included 1897 exact matches (53.3%) and a further 1270 matches (35.7%) that were up to 150 metres apart. The distribution of the sites by biological GQA class is given in Table 2.5.

Table 2.5 Distribution of matched biological and chemical sites by biological GQA class

	Biological GQA Class						Total No. of Sites
	a	b	c	d	e	f	
Number of Sites	1119	1102	746	320	204	65	3556
Percentage of Total	31.5%	31.0%	21.0%	9.0%	5.7%	1.8%	100%

2.6 Conditional Probabilities

Bayesian methods are based upon probabilities that can either be elicited from experts or estimated from data, if available. When probabilities are subjectively derived from experts they are normally referred to as beliefs. In this project all probabilities were derived from data. Probabilities may be conditional probabilities or prior (i.e. unconditional) probabilities. A conditional probability is the probability of an event occurring given that some condition is known to exit. For example, Table 2.7 shows that the probability of finding Asellidae in abundance category 2 (i.e. state 2) is 0.423 given that: a) it is spring; b) the site is a lowland pool; and c) the 'organic' river quality is class 'b'. This is just one element of the conditional probability matrix for Asellidae. If Table 2.7 had included the probability tables for both seasons (spring and autumn) and all five site types, it would have represented the complete conditional probability matrix for Asellidae conditioned on season, site type and 'organic' river quality. In the absence of any information on these three variables, the relevant probability would be the prior probability of Asellidae being in state 2, which is about 0.20.

The probabilities required for the Bayesian models were the conditional probabilities, $P(e_{jk}|Q,T)$, of taxon j occurring in state k given the river quality class (Q) and the site type (T). In our case, there were 76 BMWP taxa, five states of existence (i.e. absence = 0 and abundance categories = 1 to 4), six river quality classes (a - f) and five site types (1 to 5). Values of $P(e_{jk}|Q,T)$ were derived using the combined spring and autumn biological data from the 5008 exemplar sites[5]. The distribution of the sites with respect to site type and 'organic' river quality class is given in Table 2.6. The performance of the resulting Bayesian classifier indicated that better results might be achieved by developing separate classifiers for spring and autumn. Thus, conditional probabilities were also derived for spring and autumn data taken separately (i.e. $P(e_{jk}|Q,T,S)$, where S is the season).

[5] The database of exemplars used in this case was constructed in a slightly different way from that described in Section 2.4. Instead of splitting ties in site types (such as 2, 3 and 4) by referring to a fourth classification, they were split by taking the middle value (i.e. 3 in this case) provided that all three were direct neighbours. This increased the size of the database slightly, from 4960 to 5008 sites.

quality class. This would have undermined the performance of the naive Bayesian classifiers and, to a lesser degree, the supervised-learning neural networks.

Thus it was thought better to separate the components into their 'organic' (based on ASPT) and 'toxic' (based on NFAM) parts, and to use only the former, being the reliable part, as the target river quality classifications in the database of exemplars.

It should be noted that the decision to split the GQA quality classification into its two component parts was only relevant to the supervised-learning networks and the derivation of conditional probabilities for the Bayesian classifiers. The unsupervised-learning networks did not require target classifications and were therefore unaffected by this decision.

The first stage in the construction of the database of exemplars was to combine the taxonomic strings of the spring and autumn samples. This was achieved by taking the highest of each family's two abundance levels ('absence' being treated as zero). The next stage was to derive reliable site-type and 'organic' river quality classifications for each site.

Site type was classified into five ASPT bands (labelled 1-5) based upon three different methods of predicting unpolluted ASPT, two based on neural networks (NNRSCR & NNIFE614 - see Section 3.2 for details) and the other on RIVPACS III The final site-type classification was allocated by 'majority vote' of the three methods. In the rare cases where none of the three methods agreed, a fourth method (based upon neural network NNBMWP) was introduced to break the tie. Section 3.2 gives a detailed explanation of the site types and the reasons for classifying them. The geographic distributions of the 6038 sites by site type are shown in Figures B1 to B5 in Appendix B.

In the case of 'organic' river quality, the final classification required complete agreement between two methods based on EQI(ASPT) bandings. One used predicted ASPT derived by RIVPACS III and the other used values derived by a neural network predictor of ASPT (called NX5ASPT - see Section 3.3.5). If the classifications given by the two methods disagreed then the site was rejected from the database of exemplars. Both methods were based upon original BMWP scores, not revised scores, since appropriate data on the latter were not available at the time.

The final number of sites remaining in the database of exemplars was 4960, of which site types 1, 2, 3, 4 and 5 contributed 1048, 1001, 955, 996 and 960 respectively.

2.5 Database of Matched Biological and Chemical Sites

In order to develop the predictors of BOD, DO and ammonia it was necessary to link data on these three chemical variables to the validated biological data. This required the matching of chemical sampling sites to biological sampling sites. Since many 'matched' chemical and biological sites had different Ordnance Survey grid references, software was developed to calculate their distance apart. Analysis of these distances uncovered many grid reference errors, and these were either corrected or the sites removed from the database of matched sites. To ensure that the database only included valid matches, all pairs of sites that were greater than 400 metres apart were excluded, with the exception of sites in North East (Northumbria) where a threshold of 1 km was used. The reason for this exception was that the region had

very few matched sites that were less than 400 metres apart. It was thought desirable to increase the threshold in this case to avoid under-representation of the region in the database.

Unfortunately, many biological sites did not have a matching chemical site, and the final database consisted of just 3556 matched sites, which included 1897 exact matches (53.3%) and a further 1270 matches (35.7%) that were up to 150 metres apart. The distribution of the sites by biological GQA class is given in Table 2.5.

Table 2.5 Distribution of matched biological and chemical sites by biological GQA class

	Biological GQA Class						Total No. of Sites
	a	b	c	d	e	f	
Number of Sites	1119	1102	746	320	204	65	3556
Percentage of Total	31.5%	31.0%	21.0%	9.0%	5.7%	1.8%	100%

2.6 Conditional Probabilities

Bayesian methods are based upon probabilities that can either be elicited from experts or estimated from data, if available. When probabilities are subjectively derived from experts they are normally referred to as beliefs. In this project all probabilities were derived from data. Probabilities may be conditional probabilities or prior (i.e. unconditional) probabilities. A conditional probability is the probability of an event occurring given that some condition is known to exit. For example, Table 2.7 shows that the probability of finding Asellidae in abundance category 2 (i.e. state 2) is 0.423 given that: a) it is spring; b) the site is a lowland pool; and c) the 'organic' river quality is class 'b'. This is just one element of the conditional probability matrix for Asellidae. If Table 2.7 had included the probability tables for both seasons (spring and autumn) and all five site types, it would have represented the complete conditional probability matrix for Asellidae conditioned on season, site type and 'organic' river quality. In the absence of any information on these three variables, the relevant probability would be the prior probability of Asellidae being in state 2, which is about 0.20.

The probabilities required for the Bayesian models were the conditional probabilities, $P(e_{jk}|Q,T)$, of taxon j occurring in state k given the river quality class (Q) and the site type (T). In our case, there were 76 BMWP taxa, five states of existence (i.e. absence = 0 and abundance categories = 1 to 4), six river quality classes (a - f) and five site types (1 to 5). Values of $P(e_{jk}|Q,T)$ were derived using the combined spring and autumn biological data from the 5008 exemplar sites[5]. The distribution of the sites with respect to site type and 'organic' river quality class is given in Table 2.6. The performance of the resulting Bayesian classifier indicated that better results might be achieved by developing separate classifiers for spring and autumn. Thus, conditional probabilities were also derived for spring and autumn data taken separately (i.e. $P(e_{jk}|Q,T,S)$, where S is the season).

[5] The database of exemplars used in this case was constructed in a slightly different way from that described in Section 2.4. Instead of splitting ties in site types (such as 2, 3 and 4) by referring to a fourth classification, they were split by taking the middle value (i.e. 3 in this case) provided that all three were direct neighbours. This increased the size of the database slightly, from 4960 to 5008 sites.

Most taxa showed noticeable differences in their distributions of $P(e_{jk}|Q,T,S)$ from site type 1 to site type 5, but few, mainly stoneflies, showed marked differences between spring and autumn.

Table 2.7 illustrates a typical case of differences in conditional probability distributions between site types 1 (upland riffle) and site type 5 (lowland pool). The taxon in question, Asellidae, is essentially a pollution-tolerant pool species that invades organically polluted riffles. Consequently, it is very common in pools across all river qualities except the most severely polluted, but only common in the poorer quality riffles, as the table shows.

Table 2.8 shows an extreme difference between spring and autumn probability distributions. The taxon in question, Chloroperlidae, is far more abundant in spring than autumn.

Table 2.6 Distribution of data used to derive the conditional probabilities.

Site Type	'Organic' river quality class						Total
	a	b	c	d	e	f	
1	441	418	95	50	36	8	1048
2	359	287	176	108	107	11	1048
3	296	187	209	165	94	6	957
4	339	234	261	135	37	7	1013
5	250	305	241	105	35	6	942
Total	1685	1431	982	563	309	38	5008

Table 2.7 Conditional probability distributions, $P(e_{jk}|Q,T,S)$, for Asellidae in site types 1 and 5 during spring.

Spring - Site Type 1 (Upland riffle)

State	'Organic' River Quality Class					
	a	b	c	d	e	f
0	0.966	0.888	0.621	0.420	0.333	0.625
1	0.032	0.096	0.295	0.420	0.250	0
2	0.002	0.014	0.084	0.160	0.389	0.375
3	0	0.002	0	0	0.028	0
4	0	0	0	0	0	0

Spring - Site Type 5 (Lowland pool)

'Organic' River Quality Class					
a	b	c	d	e	f
0.188	0.148	0.174	0.257	0.229	0.667
0.436	0.403	0.469	0.371	0.457	0.333
0.356	0.423	0.320	0.324	0.229	0
0.020	0.026	0.037	0.048	0.057	0
0	0	0	0	0.029	0

Table 2.8 Conditional probability distributions, $P(e_{jk}|Q,T,S)$, for Chloroperlidae in site type 1 during spring and autumn.

Spring - Site Type 1 (Upland riffle)

State	'Organic' River Quality Class					
	a	b	c	d	e	f
0	0.141	0.287	0.705	0.920	1	1
1	0.508	0.462	0.263	0.080	0	0
2	0.347	0.249	0.032	0	0	0
3	0.005	0.002	0	0	0	0
4	0	0	0	0	0	0

Autumn - Site Type 1 (Upland riffle)

'Organic' River Quality Class					
a	b	c	d	e	f
0.726	0.837	0.979	1	1	1
0.245	0.151	0.021	0	0	0
0.029	0.012	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

3. SUPERVISED-LEARNING NEURAL NETWORKS

3.1 Introduction

A brief introduction to neural networks, including supervised and unsupervised learning, was given in Section 1.4. The most commonly used supervised-learning neural network is the standard back-propagation network, an example of which is shown in Figure 3.1. Typically, it has a layer of input nodes, a layer of output nodes (just one in this case) and normally one or two hidden layers of nodes. The latter are hidden in the sense that they do not interface with the user, unlike the input and output layers. The nodes in the hidden and output layers are different to those in the input layer in that they process data. For this reason they are referred to as processing elements. Networks are normally fully connected, in that each node in a given layer is connected to every node in its adjacent layers. The input vector $(x_1,..x_i,..x_n)$ is presented to the network via the n input nodes, shown here as boxes to distinguish them from processing elements, which are shown as circles.

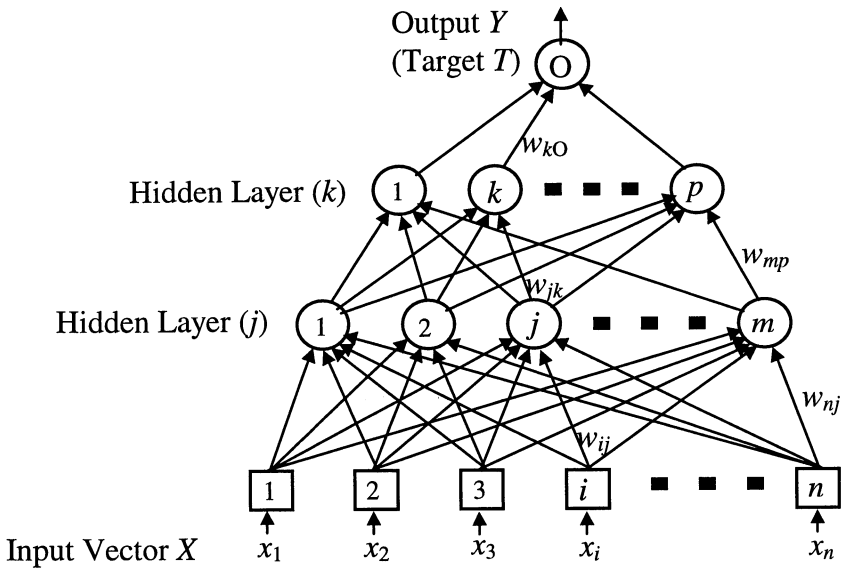


Figure 3.1 A typical standard back-propagation neural network

In the brain analogy, a processing element is analogous to the cell body of a neuron. It receives signals from many input links, which are analogous to the neuron's dendrites, and transmits them in modified form to other processing elements via output links, analogous to the neuron's axon terminals. These output signals are magnified or suppressed by weights (w) before arriving at the next processing element. The weights are analogous to the conductivities across the synaptic gaps, where nerve impulses transfer from the axon terminals of one neuron to the dendrites of another. It is these conductivities that hold the key to learning. When we learn, the conductivities at some of the many billions of synapses in our brain are modified, some suppressing the transfer of nerve impulses, others enhancing them. The back-propagation training algorithm attempts to model this learning process through the iterative adjustment of the weights on the links between layers. The modification of the nerve impulse within the cell body of the neuron is modelled by a mathematical function, called a transfer or activation function, that is associated with each processing element.

Prior to commencing the training process, the raw data are processed to form the input vectors ($x_1,..x_i,..x_n$). This generally involves transforming each individual variable to a scale ranging from -1 to +1, but more sophisticated transforms are sometimes used in an attempt to enhance the value of the input vectors. In addition, all weights are randomly initialised to small values about a mean of zero. The training process proceeds as follows.

1. A typical case is presented to the network as an input vector ($x_1,..x_i,..x_n$)
2. Inputs x_i are multiplied by their weights w_{ij} to produce the inputs to the processing elements in hidden layer (j).
3. The processing elements in layer j modify the weighted sum of their inputs to produce outputs (y_j), using the transfer function:

$$y_j = f(X_j) \tag{4}$$

where:

$$X_j = \sum_{i=1}^{n} w_{ij} x_i \, .$$

A commonly used transfer function is the hyperbolic tangent,

$$y_j = \tanh(X_j) \, ,$$

but other S-shaped functions are sometimes used.

4. The outputs y_j are multiplied by weights w_{jk} to produce the inputs to the processing elements in hidden layer (k).
5. The processing elements in layer k modify the weighted sum of their inputs to produce outputs (y_k), using the transfer function:

$$y_k = f(X_k) \tag{5}$$

where:

$$X_k = \sum_{j=1}^{m} w_{jk} y_j \, .$$

6. The outputs y_k are multiplied by weights w_{kO} to produce the inputs to the processing elements in the output layer (i.e. in this case the single element, "O").
7. The output Y is then derived by:

$$Y = f(X_O) \tag{6}$$

where:

$$X_O = \sum_{k=1}^{p} w_{kO} y_k \, .$$

8. Output Y is compared with the desired (or target) output T to determine the error E in the network's prediction.
9. The back-propagation algorithm then modifies all of the weights in the network, such that if the same input data are presented again the network's error will be less. Note that only a fraction of the error, as defined by the 'learning rate', is eliminated at each stage. A full description of the back-propagation process is beyond the scope of this report, but the

mathematics of this and other training algorithms can be found in several texts (Beale and Jackson, 1990; Haykin, 1994; Bishop, 1995).

10. Steps 1 to 9 are repeated many times using different examples, and if necessary, when all examples have been used, the same set of examples is used all over again. As time progresses the average error decreases, rapidly at first but then more slowly, and the algorithm progressively reduces the learning rate.

If the training process is allowed to continue unchecked it reaches a point where the network begins to model the noise in the data as well as its underlying features, and thereby begins to 'memorise' individual cases. This undermines the network's ability to generalise its 'knowledge' to new (i.e. previously unseen) cases. To maximise the network's performance on such cases its training has to be stopped at a stage before it starts to memorise (i.e. over-fit) the data. This is best achieved by testing the network's performance on independent data at intervals during training. Training is stopped at the point when the errors on the independent test data are minimised. Once trained, the network is ready to be put to work, making predictions for totally new cases.

Most other supervised-learning networks are similar to the standard back-propagation network described above. Many only differ in the formulation of their training algorithm (i.e. in the way that they modify the weights to reduce prediction errors), but others differ more fundamentally, in terms of their structure and function.

3.2 Site Classifiers

The environmental characteristics of a sampling site have a significant influence on its community composition, and must be accounted for in any system that attempts to classify river quality from biological data to ensure an accurate classification. It was thought desirable to classify sites into a relatively small number of types and to relate these to the principal site factors governing community composition. Consideration was given to:

a) grouping the 35 RIVPACS III site types into their parent groups; and
b) classifying the sites from their environmental characteristics using a Self-Organising Map (SOM).

Unfortunately, both of these methods resulted in site groupings that did not discriminate well between the predicted ASPTs of the sites. Such discrimination was considered to be important because ASPT represents a measure of community composition that is closely related to river quality.

Finally, it was decided to base the site-type classifications on predicted ASPTs divided into bands. The basic principle used was to divide the predicted two-season ASPTs into five ASPT bands, labelled 1 to 5, band 1 being the highest scoring and 5 the lowest. The thresholds between bands were chosen such that each band contained an approximately equal number of sites. This was done to facilitate uniformity of precision in the classification models. To ensure that the final site-type classifications were as reliable as possible four different methods were used to predict the two-season ASPTs, and hence to classify the site types (as detailed in Section 2.4).

1) A neural network predictor (NNBMWP) trained and tested on the observed ASPTs of the Biological GQA Class 'a' sites in the 1995 database of validated sites;

2) A neural network predictor (NNRSCR) trained and tested on the observed *revised* ASPTs (i.e. calculated using the site-abundance related scores derived by Walley & Hawkes, 1997) of Biological GQA Class 'a' sites in the 1995 database of validated sites;

3) RIVPACS III two-season predictions (NB. RIVPACS III was derived using the IFE 614 dataset); and

4) A neural net predictor (NNIFE614) trained and tested on the observed ASPTs of sites in the IFE 614 dataset;

Thus two of the methods were developed using data from Biological GQA Class 'a' (presumably 'clean') sites that were sampled in the 1995 Survey, and the other two were based on the IFE614 dataset. Three of the methods were based on neural networks and one on multivariate statistical methods (i.e. RIVPACS). The reason for using data based on Biological GQA Class 'a' sites was to overcome the site-type bias that existed in the IFE614 database, as illustrated in Table 3.1. The reason for using a method based upon the revised ASPT scores was to guard against the system being too sensitive to the revision of BMWP scores.

Table 3.1. Number of IFE614 sites falling within a specific alkalinity/substrate class expressed as a percentage of the corresponding number of 1995 National Survey sites

	Nature of Substratum (Percentage Sand + Silt)					
Alkalinity	0 to <10	10 to <30	30 to <50	50 to <70	70 to <90	90 to 100
0 to <15	25.6%	23.6%	0.0%	0.0%	0.0%	N/A
15 to <30	22.2%	5.2%	12.6%	0.0%	0.0%	N/A
30 to <50	22.6%	8.2%	6.8%	0.0%	0.0%	0.0%
50 to <75	20.2%	9.0%	4.4%	3.6%	0.0%	0.0%
75 to <100	23.8%	5.6%	6.2%	6.6%	6.0%	2.8%
100 to <150	18.4%	11.2%	5.0%	4.8%	7.0%	6.6%
150 to <200	11.8%	8.2%	2.2%	6.6%	1.8%	3.8%
200 to <250	15.0%	8.0%	8.2%	7.6%	7.6%	6.6%
250 to <300	0.0%	6.4%	4.2%	3.4%	6.8%	2.8%
300 +	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Preliminary tests carried out on neural network NNBMWP showed that three environmental factors dominated the prediction of ASPT, namely alkalinity, altitude and the percentage of silt in the substrate. However, these tests also showed that virtually the same level of performance (i.e. correlation coefficient = 0.8094 compared to 0.8150) was achieved using the combined percentage of sand and silt instead of just silt alone. It was felt that this negligible loss of performance was offset by certain advantages of combining the sand and silt percentages. Firstly, it meant that one of the three environmental inputs was identical to that used by Walley and Hawkes (1996, 1997) as the basis of their site-type classifications. Secondly, it reduced the likely error in the measurement of this input, since the recorded percentage of sand+silt in the substratum is subject to slightly less error than silt alone. It should be noted that the predicted ASPTs are only used as a means of ranking the sites in terms of their 'ASPT potential', and that their absolute values are irrelevant to the process of site classification.

Each of the four methods was used to classify all 6038 validated sites, irrespective of their river quality class. The number of sites allocated to each site type by the four methods, acting separately and in combination (i.e. based upon unanimity of classification), are given in Table 3.2, together with their combined (or joint) classification rates expressed as a percentage.

Table 3.2 Results of site classifications based on various combinations of four different models. The results from combinations of two or more methods are listed in order of their joint classification rate. Also given is the average mutual information between site class and the biological data.

Method				No. sites allocated to class : *Percentage of max. possible allocation*									
1	2	3	4	Site Type 1		Site Type 2		Site Type 3		Site Type 4		Site Type 5	
•				1208		1209		1210		1209		1210	
	•			1209		1209		1209		1209		1210	
		•		1231		1199		1217		1207		1192	
			•	1208		1211		1210		1208		1209	
•	•			1165	*95.4*	1113	*92.1*	1030	*85.2*	1026	*84.9*	1155	*95.4*
•			•	1127	*93.3*	1042	*86.2*	932	*77.0*	884	*73.1*	1071	*88.6*
	•		•	1109	*91.8*	995	*82.3*	901	*74.5*	891	*73.7*	1084	*89.7*
•	•		•	1084	*90.1*	986	*81.6*	829	*68.6*	793	*65.6*	1047	*86.6*
•		•		1054	*87.3*	807	*67.3*	617	*51.0*	610	*50.5*	818	*68.6*
		•	•	1052	*87.1*	816	*68.1*	598	*49.4*	603	*50.0*	808	*67.8*
	•	•		1037	*85.8*	795	*66.3*	578	*47.8*	579	*48.0*	796	*66.8*
•	•	•		1022	*84.6*	755	*63.0*	534	*44.2*	513	*42.5*	781	*65.5*
•		•	•	1012	*83.8*	724	*60.4*	498	*41.2*	479	*39.7*	756	*63.4*
	•	•	•	987	*81.7*	702	*58.5*	471	*39.0*	468	*38.8*	748	*62.8*
•	•	•	•	981	*81.2*	696	*58.0*	452	*37.4*	429	*35.5*	736	*61.7*

Note: Method 1 = NNBMWP, Method 2 = NNRSCR, Method 3 = RIVPACS III
Method 4 = NNIFE614.

The purpose in combining the classifications was to determine what proportion of sites were consistently classified into the same site type by different methods of classification. It is clear from the results given in Table 3.2 that greater consistency existed between the three neural network methods than between RIVPACS and any one neural network.

The highest level of consistency between any two methods was achieved by the NNBMWP and NNRSCR models (average joint classification rate across all five site types = 90.6%), which was encouraging since it demonstrated that the classification system was fairly insensitive to the revision of BMWP family scores. The closest matching model to that based on RIVPACS predictions was NNBMWP, but their average joint classification rate across all five site types was only 64.9%.

The variation in joint classification rate (%) across the site types is worthy of comment. All combinations tended to agree on the classification of site type 1, the lowest joint classification rate being 81.2% when all four methods were combined. Site types 2 and 5 were also classified with a fair degree of consistency, giving joint classification rates across all four methods of 58.0% and 61.7% respectively. Site type 4 had the lowest rate (35.5%), followed

closely by site type 3 (37.4%). However, these low values were almost entirely due to a mismatch between the RIVPACS method and the three neural networks. The corresponding values based on the three networks alone were 68.6% and 65.6% respectively.

It was originally intended to construct the database of exemplars using only sites that were classified to the same site type by all four methods, or at least three of them, but this would have differentially reduced sample sizes, leaving site types 3 and 4 with very few samples. Thus it was decided to base the final site types on the majority vote of just three methods, NNRSCR, RIVPACS III and NNIFE614, leaving NNBMWP to provide the casting vote in the event of a tie. Figures B1 to B5 in Appendix B give the spatial distributions of site types 1 to 5 over England and Wales. These show that type 1 sites (i.e. having low percentage sand+silt, low alkalinity and a tendency to the higher altitudes), occur predominantly in the upland regions of the Lake District, Pennines, Wales, Devon and Cornwall, whereas type 5 sites (i.e. having high percentage sand+silt, high alkalinity and low altitude), occur predominantly in the flat, lowland, soft rock areas of England, such as East Anglia, Lincolnshire, Somerset, Cheshire, Oxfordshire and parts of Yorkshire. Table 3.3 gives the regional distributions of sites by their site types.

Table 3.3 Regional distributions of sites by site type.

	Percentage of Sites in Site Type:				
Region	1	2	3	4	5
Anglian	0.0%	0.5%	7.2%	25.4%	66.9%
North East (Northumbria)	27.6%	25.7%	28.7%	9.9%	8.1%
North West	29.7%	31.9%	15.9%	9.1%	13.3%
Midlands	6.6%	17.2%	34.9%	26.8%	14.4%
Southern	2.6%	24.7%	17.6%	26.3%	28.7%
South West (Devon & Cornwall)	58.9%	30.8%	7.9%	1.3%	1.0%
Thames	0.2%	3.5%	23.6%	45.0%	27.7%
Welsh	46.0%	29.3%	9.7%	9.2%	5.8%
South West (North & South Wessex)	2.0%	5.8%	28.9%	40.8%	22.5%
N.East (Dales & Riding)	23.7%	28.2%	17.7%	12.2%	18.2%
National Average	20.3%	20.1%	19.2%	20.3%	20.2%

Clearly, site types 1 and 5 are best visualised as upland riffles and lowland pools respectively. The other three types represent a progression between these two extremes. The precise definition of the five site types involves a continuous relationship between three variables (i.e. alkalinity, altitude and the percentage of sand+silt in the substrate), thus they cannot be defined by simple rules. However, they can be represented approximately in tabular form. Figures B6 to B9 show the distribution of site types 1 - 5 with respect to alkalinity, percentage sand+silt and altitudinal range, and provide a convenient means of visualising the different types.

3.3 Predictors of ASPT and NFAM

These networks were not developed as part of the mainstream AI approach to biomonitoring, but as an extension to the work on site classifiers. They were designed to provide a like-for-like comparison between RIVPACS III (Wright *et al.,* 1995) and neural network predictors of 'unpolluted' average score per taxon (ASPT) and number of families (NFAM).

3.3.1 The data

The networks were developed using the same data set (IFE614) that was used to develop RIVPACS III. That is, 13 environmental variables plus observed ASPTs and NFAMs from each of 614 'unpolluted' river sites covering the whole of Great Britain. The environmental variables were as follows:

a) location (National Grid Reference or NGR);
b) altitude (m - above Ordnance datum);
c) distance of site (km) downstream from its source;
d) discharge category (on a 1 to 10 log-type scale);
e) slope of the river bed (m/km) based on the distance between 50-metre contour lines;
f) average width (m) and average depth (m) of the river at the time of sampling;
g) nature of the river bed (or substrate) expressed as four average percentages of the plan area covered by boulders, pebbles, sand and silt; and
h) average alkalinity (mg/l of $CaCO_3$).

All of these contributed to the initial input vector to the networks, although some were represented in a different form. The National Grid Reference, for example, was converted to global co-ordinates X and Y, based upon an origin at grid reference SV 000 000. In addition, the logarithmic values of slope and distance from source were used in place of their straight values, because initial tests showed that these transforms improved performance. Thus the 13 environmental variables used in this exercise were those listed in Table 3.4 below.

Table 3.4 List of the 13 variables in the full environmental input vector

Variable	Description	Variabl	Description
X	Global easting of NGR	DISCH	Discharge category
Y	Global northing of NGR	BLDS	Boulders (% of substrate)
ALT	Altitude (m)	PBLS	Pebbles (% of substrate)
LDIST	Log_{10} distance from source (km)	SAND	Sand (% of substrate)
LSLOPE	Log_{10} of slope (m/km)	SILT	Silt (% of substrate)
WIDTH	Average width of river (m)	ALK	Alkalinity (mg/l of $CaCO_3$)
DEPTH	Average depth of river (cm)		

The data used as target outputs were the observed ASPT and NFAM based upon two-season (i.e. spring and autumn) and three-season (i.e. spring, autumn and winter) combined samples. The final models, however, were based on two-season data only, because the Agency's biological monitoring programme was based upon two-season sampling.

3.3.2 Training and testing

Relatively small data sets, like IFE614, present a dilemma. In the interests of precision it is tempting to use all of the data to train the network, but this leaves no means of testing its performance on independent data. As mentioned earlier, the danger in doing this is that the network may over-fit the data and thereby give a false impression of high performance. This problem was overcome in this study by randomly partitioning the data into two equal sub-sets of 307 sites, labelled F1 and F2. These were used interchangeably as training and testing sets, thus producing two networks - one trained on F1 and tested on F2 and the other trained on F2 and tested on F1 - the final prediction being based on the average of the two. This process, known as two-fold cross validation, enabled the final model (i.e. the average of two independent predictions) to be based upon all 614 records, while at the same time permitting independent testing on all 614 records. Over-fitting was avoided by terminating training at the point where performance on the independent test sets began to deteriorate.

3.3.3 Choice of network type

Before embarking on the development of the final network, preliminary tests were carried out to determine which type of network was best suited to the problem at hand. Various types of network were trained and tested using all 13 environmental variables in the input vector. These included the standard back-propagation, modular, directed random search, genetic reinforcement learning, radial basis function, general regression and Bayesian neural networks. The results of performance tests on these networks are given in Table 3.5.

Table 3.5 Results of performance tests on various neural networks.

	Correlation coefficients between target and predicted ASPTs		
Type of Network	Trn F1/Tst F2	Trn F2/Tst F1	Mean
Back-propagation Network	0.851	0.809	0.830
Modular Network with 3 experts	0.854	0.810	0.832
Directed Random Search	0.840	0.763	0.802
Genetic Reinforcement Learning	0.640	0.602	0.621
Radial Basis Function	0.816	0.815	0.816
General Regression Neural Network	0.780	0.760	0.770
Bayesian Neural Network	0.806	0.785	0.795

The best performing networks were the standard back-propagation network and the modular network. The number of experts (i.e. modules) used in the modular networks were found to have little effect on their performance. The one that used three modules produced the best overall correlation (i.e. 0.832) between target and predicted ASPTs, but this was only marginally better than that achieved by the standard back-propagation network (i.e. 0.830). It was decided to use the standard back-propagation network for the development of the ASPT and NFAM predictors, because this network has far fewer parameters and is thus faster to train and less prone to over-fitting.

3.3.4 Identification of key variables

The relevance of each environmental variable to the prediction of ASPT and NFAM was determined using leave-one-out impact analysis. Thus networks were trained using all 13 input variables and then tested on the independent test set, using the correlation coefficient between predicted and desired outputs as the measure of performance. The effect of disabling each input node in turn was then determined, in terms of the resulting percentage reduction in the correlation coefficient. This provided a measure of the importance of each input variable to the network's predictions. The results of the impact analyses on the initial 13-input networks are given in Table 3.6.

These results showed that three variables, ALT, ALK and SILT, dominated the prediction of ASPT, with impacts of 31.31%, 22.65% and 11.5% respectively. In the case of NFAM, two variables dominated the predictions, Y and LDIST with impacts of 33.28% and 29.71% respectively, but several other inputs had impacts greater than 5% (i.e. LSLOPE, ALK, ALT, DISCH and X).

Table 3.6 Results of impact analyses on the 13-input neural network predictors of ASPT and NFAM. The variables are listed in order of their percentage impact on performance of the networks. F1/F2 means trained using data file F1 and tested using data file F2.

Input	ASPT Predictor (% impacts)			Input	NFAM Predictor (% impacts)		
Variable	F1/F2	F2/F1	Mean	Variable	F1/F2	F2/F1	Mean
ALT	20.71	41.90	31.31	Y	36.79	29.78	33.28
ALK	20.61	24.39	22.65	LDIST	44.14	15.28	29.71
SILT	13.18	9.81	11.50	LSLOPE	8.30	6.40	7.35
DISCH	3.82	3.91	3.87	ALK	3.42	10.03	6.73
LSLOPE	1.73	4.43	3.08	ALT	3.96	9.45	6.70
WIDTH	-0.31	4.70	2.20	DISCH	10.43	1.20	5.81
LDIST	0.87	2.31	1.59	X	7.77	3.16	5.47
DEPTH	1.20	1.39	1.30	WIDTH	7.42	-1.21	3.10
X	0.63	1.17	0.90	SAND	2.81	2.43	2.62
Y	0.75	0.46	0.61	BLDS	3.24	1.74	2.49
PBLS	0.40	0.56	0.48	DEPTH	-0.64	5.44	2.40
SAND	0.46	0.49	0.48	PBLS	-0.33	1.57	0.62
BLDS	-0.16	-0.47	-0.32	SILT	-1.74	2.97	0.62

3.3.5 Development of a two-season predictor of ASPT

Although the results of the impact tests on the 13-input predictor of ASPT enabled the input variables to be ranked in order of importance, it would be wrong to assume that, for example, the top eight variables from this list would produce the best eight-input predictor. This is because the removal of just one variable might change the order of importance of the remaining 12 variables. The best possible eight-input predictor, for example, is produced by progressive removal of the weakest variables until just eight variables remain. At each stage, a new network is trained and subjected to impact analysis to give a new ranking of its input variables. In this study, impact analysis was used to progressively reduce the number of input

variables from 13 to one. Each network was trained and tested using cross validation between F1 and F2. The order in which the variables were deleted was based on their average percentage impact on the two networks F1/F2 and F2/F1. The change in performance level resulting from the reduction in input variables was monitored by means of the correlation coefficient between the predicted and target ASPTs. The results of these analyses are given in Table 3.7.

These show that the average correlation coefficient remained virtually unchanged as the number of environmental inputs was reduced from thirteen to five. Further reduction to three inputs produced a slight decrease, but after this the decline was rapid. Since the aim of the exercise was to reduce the number of input variables to the minimum possible without impacting significantly on overall performance, it was decided to adopt the five-input model, called N5XASPT (NB: N for neural, X for cross validated), as the final two-season predictor of ASPT. Both of the networks making up this model (i.e. F1/F2 and F2/F1) used ALK, ALT, SILT, LSLOPE and DISCH as their input variables. The average impacts of these variables on the final models were 30.64%, 26.92%, 19.82%, 3.83% and 1.62% respectively.

Table 3.7 Development of two-season predictor of ASPT - Results of progressive removal of the weakest input variables.

Number of Input Variables	Correlation Coefficients between Predicted and Target ASPT		
	Network F1/F2	Network F2/F1	Average
13	0.8503	0.8007	0.8255
12	0.8522	0.8070	0.8296
11	0.8510	0.7986	0.8248
10	0.8549	0.8084	0.8317
9	0.8591	0.8085	0.8338
8	0.8542	0.7979	0.8261
7	0.8543	0.8018	0.8281
6	0.8576	0.8093	0.8335
5	0.8512	0.8072	0.8292
4	0.8357	0.7957	0.8157
3	0.8368	0.7932	0.8150
2	0.8110	0.7553	0.7832
1	0.4335	0.4334	0.4335

The reason for the generally lower correlations produced by network F1/F2 (i.e. trained on F1 and tested on F2) was explored, and it was concluded that F1 contained more outliers than F2. The data were subsequently re-partitioned, resulting in greater similarity between the two sets.

3.3.6 Development of a two-season predictor of NFAM

The development of the NFAM predictor proceeded in the same way as that for ASPT. However, in this case the correlation coefficient started to decline at a much earlier stage in the progressive elimination of input variables. It appeared that the best model would require about seven or eight environmental inputs. To clarify the situation the same tests were carried out using the three-season data. The results are summarised in Table 3.8.

Table 3.8 Development of two-season predictor of NFAM - Results of progressive removal of the weakest input variables.

Number of Input Variables	Average Correlation Coefficient of F1/F2 and F2/F1 Networks		
	2-season model	3-season model	Overall avg. 2/3-season
13	0.5987	0.6391	0.6189
11	0.6007	0.6417	0.6212
9	0.6050	0.6372	0.6211
8	0.6000	0.6486	0.6243
7	0.5973	0.6609	0.6291
6	0.5706	0.6457	0.6081
5	0.5699	0.6355	0.6027
3	0.5702	0.6094	0.5898
1	0.4571	0.4530	0.4551

The results show that seven inputs produced the best overall correlation coefficient. In addition, the seven variables were found to be the same in both the two- and three-season models. Thus the seven-input model, N7XNFAM, was adopted as the final two-season predictor of NFAM. The seven input variables were Y, ALT, SAND, LDIST, LSLOPE, DEPTH and X in order of their average impact percentages, which were 53.8%, 16.2%, 15.7%, 15.5%, 9.1%, 7.1% and 4.9% respectively.

3.3.7 Comparison with RIVPACS III

The neural network predictors, N5XASPT and N7XNFAM, were developed using two-fold cross validation and just five and seven environmental variables respectively, whereas RIVPACS III was developed using all 614 sites and all 13 environmental variables. To provide a 'like-for-like' comparison between the two methods, it was necessary to train and test additional neural networks using all 614 sites. This was done using five and seven inputs as before, and also using all 13 inputs. These networks were named N5DASPT, N7DNFAM, N13DASPT and N13DNFAM, where the 'D' refers to the fact that, like RIVPACS III, they were tested on dependent data (i.e. of necessity). It will be recalled that the 'X' used in the names of the earlier networks referred to the fact that they were cross-validated. The performance of the neural network predictors, expressed in terms of their linear regression statistics, are given in Table 3.9 together with equivalent figures for RIVPACS III.

Table 3.9 Performance of various predictors of ASPT and NFAM expressed in terms of the correlation coefficient (r), slope coefficient (a) and intercept (c) of the linear regression lines relating predicted values to observed values.

	ASPT Predictors				NFAM Predictors		
Model	r	a	c	Model	r	a	c
N5XASPT	0.8261	0.9861	0.0822	N7XNFAM	0.5860	0.8669	3.6168
N5DASPT	0.8358	1.0007	-0.0040	N7DNFAM	0.6704	1.0070	-0.1752
N13DASPT	0.8468	1.0016	-0.0086	N13DNFAM	0.6665	1.0070	-0.2149
RIVPACS III	0.8444	1.0274	-0.1624	RIVPACS III	0.6703	1.0808	-2.3578

The 'like-for-like' comparison between the ASPT predictor N13DASPT and RIVPACS III showed that the neural network slightly out-performed RIVPACS in terms of its correlation

coefficient, but more noticeably in terms of its slope coefficient (ideally 1.0000) and intercept (ideally 0.0000). The 'like-for-like' comparison between the NFAM predictor N13DNFAM and RIVPACS III showed that RIVPACS slightly out-performed the neural network in terms of its correlation coefficient, but that the network noticeably out-performed RIVPACS III in terms of slope coefficient and intercept. Indeed, a slope of 1.08 and intercept of -2.36 implies that RIVPACS III is biased in its predictions of high and low values of NFAM. The overall performance of N13DNFAM was therefore judged to be slightly better than RIVPACS III.

It is also worth noting that N5DASPT matched RIVPACS III in terms of overall performance and that N7DNFAM was the best overall predictor of NFAM. This clearly demonstrates that spatial variations in 'unpolluted' ASPT and NFAM can be predicted using a relatively small number of environmental variables. However, only ALT and LSLOPE were common to the two models, leaving just three of the 13 environmental variables as totally redundant (i.e. PBLS, BLDS and WIDTH). Clearly, at least one of the four substrate variables was bound to be redundant.

The overall performance of the cross validated predictor N5XASPT appears to be marginally worse than that of RIVPACS III, but one has to remember that in this case the test results were based on independent data, not dependent data as was the case for RIVPACS III. Thus, it seems likely that this network would prove to be a better predictor of ASPT for new data than would RIVPACS. Indeed the only true measure of the worth of a model is its performance on independent data.

The performance of the cross-validated predictor N7XNFAM appeared to be much worse than that of all the dependent models, including RIVPACS III. In order to gain some measure of the difference to be expected between performances on dependent and independent data, a series of tests were carried out in which identical networks were trained and tested dependently and independently (i.e. using combinations F1/F1, F2/F2, F1/F2 and F2/F1). The results showed that the average difference between the correlation coefficients produced by the ASPT predictor when tested on dependent and independent data was only 0.0287, indicating that its performance on previously unseen data was only marginally below that achieved on its training set. In the case of the NFAM predictor the average difference was 0.1671, representing a 23% reduction in correlation coefficient between tests on dependent data and previously unseen (independent) data. The implications of these findings are:

a) that N5XASPT is the best overall predictor of ASPT;
b) that N7XNFAM is probably the best overall predictor of NFAM;
c) that NFAM predictions are far less reliable than ASPT predictions; and
d) that the EQI(NFAM) component of biological GQA classifications are far less reliable than previously thought.

The predictions made by the neural networks were achieved directly from the input variables in one non-linear mathematical mapping, whereas RIVPACS (Moss *et al.,* 1987) used three consecutive mappings (i.e. environmental data to site type, site type to community structure, community structure to predicted ASPT and NFAM). It has been suggested that RIVPACS adds ecological knowledge to the prediction process, because its three mappings are based on ecological concepts, but this is not so. The three steps simply provide convenient staging posts for three linear (or relatively simple non-linear) mathematical mappings. RIVPACS, like the neural networks, is purely a mathematical model calibrated from data. Even the use of

temperature and temperature range within RIVPACS is a mathematical function of the input variables (e.g. X, Y and ALT). Thus, the neural networks would have automatically accommodated, via their complex non-linear mappings, any temperature effects that related to X, Y and ALT. If necessary, the neural networks could have been trained to perform the same three mappings as RIVPACS, thus providing predictions of biological site-type and community structure. However, the aim of the exercise was solely to predict ASPT and NFAM for the purpose of river quality classification, so there was no point in introducing unnecessary steps. The single non-linear mapping makes the classification process simpler without any loss of validity.

3.3.8 Sources of error and bias

In order to check the predicted values of ASPT and NFAM for spatial bias, the EQIs of the English and Welsh sites in the IFE614 database were plotted on maps. Since the database contains only 'unpolluted' sites their EQIs should equal unity, but in reality they are scattered around unity. Figures C1 and C2 in Appendix C show the distributions over England and Wales of the deviations from unity of EQI(ASPT) and EQI(NFAM) as derived from the neural network predictions of ASPT and NFAM. Figures C3 and C4 show the corresponding distributions based on RIVPACS predictions. None of these maps reveals any evidence of overall spatial bias in the deviations of the EQIs. However, they do indicate spatial bias on a river basin scale, because EQIs greater or less than unity tend to persist along some rivers (e.g. the Usk and the Stour in Kent on Figure C3). This implies that some relevant environmental factors are missing from the present profile of environmental characteristics. The existing variables predominantly represent properties of the site, only alkalinity can be considered to incorporate characteristics of the catchment. Thus, there appears to be a deficiency in catchment characteristics, such as geology, soil type and mean altitude of the catchment. The importance of geology in determining the community composition was demonstrated by Ruse (1996). The amount of woody litter at the site could also be a useful addition to the site variables. On the other hand, the suitability of alkalinity as an environmental input variable has to be questioned, bearing in mind its relationship to some forms of pollution. Since it was found to be the most important variable governing the prediction of ASPT, it would clearly be advantageous to devise a means of predicting natural alkalinity from other environmental variable that are independent of pollution. Perhaps the inclusion of catchment geology would help in this regard.

Despite the fact that RIVPACS and the neural networks gave almost identical levels of performance, they produced some very different predictions for individual sites, as can be seen from Figures 3.2a and 3.2b. The correlation coefficients between the two systems were 0.9562 and 0.8525 for ASPT and NFAM predictions respectively, indicating better agreement between ASPT predictions than NFAM predictions, as one might expect given that both RIVPACS and the networks performed better with respect to ASPT than NFAM.

Various analyses were carried out to determine the causes of the differences between:

a) the values predicted by the two systems; and

b) the observed and predicted values (i.e. the so-called prediction 'errors').

The results of these analyses are outlined below, but discussed in greater detail by Walley and Fontama (1998)

The likelihood of error and bias were found to be greatest at sites where one or more environmental variables were exceptionally high or low. This was especially true for sites having: high or very low altitude; and/or high or very low alkalinity; and/or a high percentage of sand and silt. Note that all of these variables were found to be key environmental factors, so if they were not truly represented in the data at the extremes of their ranges it could result in noticeable distortions in the models' predictions. It was surprising to find that high altitude was associated with poor predictions, since high altitude sites were over-represented in the data. However, detailed investigation revealed that out of five very high altitude sites (i.e. over 450 metres), three had unusually high alkalinities. It is possible that these three sites distorted the models and hence the predictions for these and other high altitude sites.

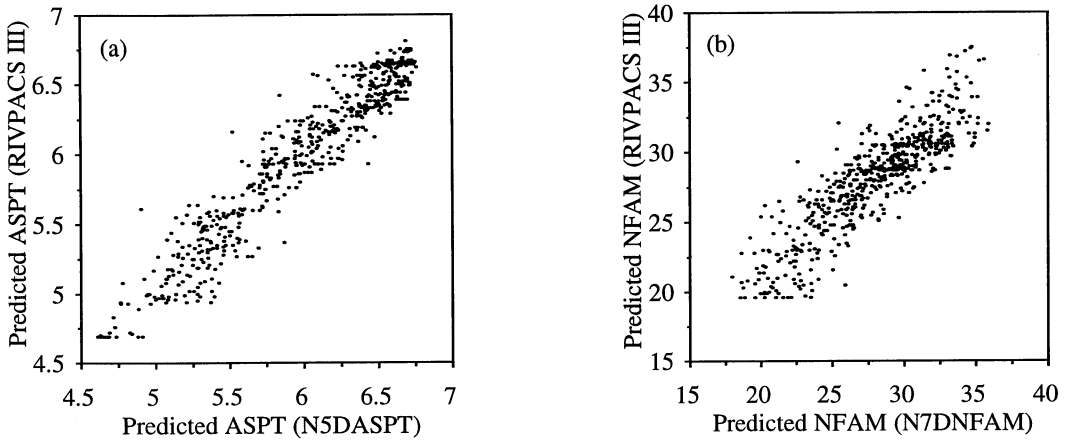


Figure 3.2 Graphs showing: (a) RIVPACS predicted ASPT against N5DASPT predicted ASPT; and (b) RIVPACS predicted NFAM against N7DNFAM predicted NFAM.

There were several cases where two sites, having almost identical sets of environmental variables, had observed values of ASPT or NFAM that differed very substantially, but very similar predicted values. For example, two of the sites had observed NFAMs of 18 and 44, whereas the values predicted by RIVPACS were 29.1 and 30.2 respectively and those predicted by the neural networks were 30.8 and 31.5 respectively. This case clearly supports the view that there are relevant environmental factors missing from the current set of environmental input variables. Although observed values of NFAM are known to be particularly sensitive to sampling effort, this could not account for the differences in this case. However, it is reasonable to assume that variations in sampling effort during the collection of the IFE614 data will have contributed to the NFAM prediction 'errors', and hence the poor performance of the NFAM models.

When discussing prediction 'errors' it is also important to realise that 'unpolluted' ASPT and NFAM are in fact non-stationary time series, since they are governed by natural variations in the biological community that are determined by factors such as seasonal variation, population dynamics, hydrological events etc. Thus the 'unpolluted' ASPT and NFAM for a given site cannot rightly be considered as constant values, since they are really stochastic variables defined by some probability distribution. Thus, a significant component of the apparent prediction errors may not be due to errors at all, but simply natural variation.

There is also the matter of what is meant by 'unpolluted'. One might expect that pollution includes things like pesticides, but does it include disease and the stresses caused by river engineering works (e.g. channel regrading, weed removal, removal of bankside vegetation, flow regulation, water abstraction etc.)? In the absence of a clear definition of the term 'unpolluted', or preferably 'unstressed', there can be no proper basis for the selection of reference sites.

3.3.9 Two-season predictor of ASPT based on revised BMWP scores.

After completing the development of the ASPT and NFAM predictors described in Sections 3.3.5 and 3.3.6, the Institute of Freshwater Ecology supplied values of observed ASPT for the IFE614 sites, based upon the revised BMWP family scores derived by Walley and Hawkes (1996, 1997). The opportunity was taken to train and test two-season predictors of ASPT using three different sets of target ASPTs, based upon the overall, site-related and site-abundance-related revised scores. The correlation coefficients achieved by these networks are given in Table 3.10.

Table 3.10 Correlation coefficients achieved by ASPT predictors based on revised BMWP scores

	N5XRASPT			N5DRASPT		
Type of revised score	F1/F2	F2/F1	Avg.	F1/F1	F2/F2	Avg.
Overall score	0.8790	0.8904	0.8847	0.9019	0.8957	0.8988
Site-related score	0.8830	0.8861	0.8846	0.9029	0.9025	0.9027
Site-abundance-related score	0.8645	0.8735	0.8690	0.8874	0.8853	0.8864

These results show that all of these networks performed better than their corresponding networks based on original BMWP scores (Table 3.9). Table 3.10 shows that the networks based on overall scores and site-related scores performed equally well, giving the best overall results. These results provide a clear indication that the revised BMWP scores derived by Walley and Hawkes (1996, 1997) are an improvement on the original BMWP scores.

3.3.10 GQA classifications based on neural network predictions of ASPT and NFAM

The neural network predictors NX5ASPT and N7XNFAM were used to derive EQI(ASPT) and EQI(NFAM) values, and hence to provide biological GQA classifications for all validated sites in the 1995 database. The EQI threshold values used to separate the different classes were not those used for the original classification based on RIVPACS, because these resulted in significant changes to the overall number of sites allocated to each class. Instead, the original threshold values were marginally adjusted to give the same overall distribution of sites between classes.

This same exercise was repeated using the neural network predictor of ASPT based on overall revised scores (i.e. using N5XRASPT in place of N5XASPT). It is worth noting that this model produced a correlation coefficient of 0.8847 between the predicted and observed ASPT, which is noticeably better than anything previously achieved. Once again the threshold EQIs used to separate the quality classes were marginally adjusted so as to achieve the same overall distribution of sites between quality classes as that produced by RIVPACS. Table 3.11 shows

the original EQI threshold values that were used to produce the biological GQA classifications from RIVPACS predictions and the adjusted values used to produce the neural network classifications based on revised scores.

Table 3.11 Original and adjusted EQI classification thresholds.

River Quality Class	Original Thresholds RIVPACS Model		Adjusted Thresholds (Revised scores) N5XRASPT / N7XNFAM Model	
	EQI(ASPT)	EQI(NFAM)	EQI(ASPT)	EQI(NFAM)
a	1.00	0.85	0.989	0.850
b	0.90	0.70	0.889	0.711
c	0.77	0.55	0.771	0.553
d	0.65	0.45	0.664	0.453
e	0.50	0.30	0.535	0.300
f	<0.50	<0.30	<0.535	<0.300

3.3.11 Comparisons between GQA classifications based on RIVPACS and neural network

Table 3.12 compares the biological GQA classifications derived using RIVPACS with those derived using the neural networks based on the original BMWP scores (i.e. the N5XASPT / N7XNFAM models).

Table 3.12 Distribution of validated sites by biological GQA class based on EQIs derived by RIVPACS and the neural networks N5XASPT and N7XNFAM.

Equivalent GQA Classification based on EQIs derived by N5XASPT & N7XNFAM	GQA classification based on EQIs derived by RIVPACS						
	a	b	c	d	e	f	Total
a	1505	252	4	0	0	0	1761
b	255	1320	172	0	0	0	1747
c	2	175	999	95	1	0	1272
d	0	0	96	479	66	0	641
e	0	0	0	64	403	20	487
f	0	0	0	0	18	112	130
Total	1762	1747	1271	638	488	132	6038

The number of sites given the same biological GQA classification by RIVPACS and the neural networks based on BMWP scores (i.e. using N5XASPT / N7XNFAM) was 4818, representing 79.8% of the 6038 validated sites. Hence, a change in the method of predicting ASPTs and NFAMs from a statistical approach (i.e. RIVPACS) to a neural networks approach, resulted in 20.2% of sites being classified to a different GQA class (i.e. 20.1% by ± 1 class and 0.1% by ± 2 classes).

Table 3.13 compares the biological GQA classifications derived using RIVPACS with those derived using the neural networks based on the overall revised scores (i.e. the N5XRASPT / N7XNFAM models). In this case, the networks classified 4379 sites to the same class as

RIVPACS, representing 72.5% of all sites. This means that a further 7.3% of sites were classified to a different class when the basis of the neural network predictors was changed from BMWP scores to the overall revised scores derived by Walley and Hawkes (1996, 1997), thus putting 27.5% of sites in a different GQA class to that allocated by RIVPACS (i.e. 27.2% by ± 1 class and 0.3% by ± 2 classes). This does not necessarily mean that over 20% of sites are presently classified incorrectly, but it does mean that classifications based upon EQIs are sensitive to the method used to derive the them (even though the methods have the same level of accuracy) and, to a lesser degree, the accuracy of the scores allocated to each family.

Table 3.13 Distribution of validated sites by biological GQA class based on EQIs derived by RIVPACS and the neural networks N5XRASPT and N7XNFAM.

Equivalent GQA Classification based on EQIs derived by N5XRASPT & N7XNFAM	GQA classification based on EQIs derived by RIVPACS						
	a	b	c	d	e	f	Total
a	1403	348	10	0	0	0	1761
b	352	1165	230	0	0	0	1747
c	7	233	900	129	2	0	1271
d	0	1	131	422	85	0	639
e	0	0	0	87	378	21	486
f	0	0	0	0	23	111	134
Total	1762	1747	1271	638	488	132	6038

3.4 Predictors of BOD, DO and Ammonia

Supervised-learning networks were developed to predict BOD, DO and ammonia using the standard back-propagation network with an input vector consisting of the states of existence (i.e. 0 to 4) of the 76 BMWP. The final networks had two hidden layers, one with 60 nodes and the other with 30, and a single output node representing either BOD, DO or ammonia.

The database of matched biological and chemical sites consisted of 3556 validated sites, but for the purpose of this study, all sites that were greater than 150 m apart were eliminated from the database. The figure of 150 m was chosen on the basis that an error of ± 100 m between both the NGR eastings and northings of the chemical and biological sites would result in a distance apart of 141 m. This was considered to be an acceptable error in the grid references of sites that were truly matched. The remaining 3167 sites were divided into the five site types and then randomly partitioned to provide test sets of exactly 100 sites and training sets of approximately 500 sites. Training was continued until performance on the test set started to deteriorate, subject to a maximum of 200,000 cycles. The results of performance tests on the final models, based on the independent test sets of 100 sites, are given in Table 3.14.

Although the correlation coefficients given in Table 3.14 are not as high as one might like, it is worth noting that the dissolved oxygen predictor was as good as the independent predictor of NFAM (see Table 3.9) and the predictor of ammonia was noticeably better. If the correlations had been better, a more detailed study would have been carried out involving two-

fold cross validation, with a view to producing a biological equivalent of the chemical GQA classification.

Table 3.14 Correlation coefficients between observed and predicted values of BOD, DO and ammonia for the 100 independent test sites.

Chemical Parameter	Site Type					Overall Mean
	1	2	3	4	5	
BOD	0.3277	0.6330	0.5105	0.5626	0.4513	0.4970
Dissolved Oxygen	0.6379	0.6666	0.5923	0.5587	0.4362	0.5783
Ammonia	0.6106	0.6063	0.7106	0.6206	0.7509	0.6598

3.5 Classifiers of River Quality

Supervised-learning classifiers of 'organic' river quality were trained and tested for each of the five site types. The database of exemplars (Section 2.4) was partitioned to produce a representative test set of exactly 100 sites for each site type, leaving the remaining sites (i.e. approx. 900 sites per site type) to make up the training set. All networks had two hidden layers, with between 10 and 60 nodes in each. The output layer had six nodes, one for each quality class.

The initial tests were based upon three of the seven input vectors defined in Section 2.2, namely the 16 sensitivity groups; 21 taxonomic groups (inputs = total abundance); and 50 top families plus NFAM. Networks based on these three input vectors were tested against each other using data for site types 1, 3 and 5. The target outputs for these tests were the 'organic' river quality classes, as defined in Section 2.4. The overall 'success' rates achieved by the network are given in Table 3.15. It should be noted that success in this context means that the network classified the sample to the same class as the 'organic' component of its GQA classification. This assumes that the target 'organic' classes had all been correctly assigned, which was unlikely, despite all our efforts to eliminate sites with dubious classifications. Nevertheless, the tests did provide a valid measure of the networks' abilities to model the data.

Table 3.15 Results of initial tests on supervised classifiers.

	Overall 'Success' Rates (%)		
Model (input vector)	Site Type 1	Site Type 3	Site Type 5
50 families	56.9	65.6	59.8
50 families + NFAM	67.7	68.6	68.6
76 families + NFAM	88.2	74.5	69.6

In the light of these results, networks were trained to predict the 'organic' river quality for all five site types using an input vector comprising the 76 BMWP families plus NFAM. These models were developed using the database of exemplars, thus the target classifications were the 'organic' classes as defined in that database (Section 2.4). During the training phase the networks were tested at intervals of 1,000 cycles until their performance on the test set started to deteriorate, subject to a maximum of 200,000 training cycles. The results of performance tests on these networks are given in Table 3.16. The success rates are based on the

assumption that the 'organic' classifications in the database of exemplars are correct. Since some will be incorrect, despite all our efforts to eliminate misclassifications from the database of exemplars, the results are best interpreted as lower bounds of the true performance.

Table 3.16 Performance of 77-input predictors of 'organic' river quality class.

Site Type	Success rates (%) by 'organic' class						Average per class	Overall rate (%)
	a	b	c	d	e	f		
1	74.1	75.0	64.7	80.0	57.1	0.0	58.5	69
2	77.4	84.2	66.7	71.4	84.0	33.3	69.5	77
3	79.5	55.6	70.0	81.0	71.4	0.0	59.6	74
4	98.2	46.2	55.6	86.7	75.0	0.0	60.3	83
5	86.1	68.4	47.8	86.7	60.0	0.0	58.2	71
Avg	83.1	65.9	61.0	81.2	69.5	6.7	61.2	75

Although the overall 'success' rate was 75%, it was not very uniform across the classes, falling to only 6.7% for class 'f' and giving an average per class of just 61.2%. This kind of problem often occurs when classifiers are developed using data in which one or more classes are poorly represented. In this case it was the consequence of there being very few class 'f' sites in the database. One way to overcome the problem would have been to repeat the class 'f' (and even class 'e') data in the training set as many times as necessary to provide adequate representation. Clearly, repeated data are poor substitutes for extra data, but this procedure would have served to improve performance on class 'f'. Plans to remedy this problem and to re-train the networks using two-fold cross validation analysis were abandoned in favour of work on Self-Organising Maps (SOM), which were showing considerable promise.

assumption that the 'organic' classifications in the database of exemplars are correct. Since some will be incorrect, despite all our efforts to eliminate misclassifications from the database of exemplars, the results are best interpreted as lower bounds of the true performance.

Table 3.16 Performance of 77-input predictors of 'organic' river quality class.

Site Type	Success rates (%) by 'organic' class						Average per class	Overall rate (%)
	a	b	c	d	e	f		
1	74.1	75.0	64.7	80.0	57.1	0.0	58.5	69
2	77.4	84.2	66.7	71.4	84.0	33.3	69.5	77
3	79.5	55.6	70.0	81.0	71.4	0.0	59.6	74
4	98.2	46.2	55.6	86.7	75.0	0.0	60.3	83
5	86.1	68.4	47.8	86.7	60.0	0.0	58.2	71
Avg	83.1	65.9	61.0	81.2	69.5	6.7	61.2	75

Although the overall 'success' rate was 75%, it was not very uniform across the classes, falling to only 6.7% for class 'f' and giving an average per class of just 61.2%. This kind of problem often occurs when classifiers are developed using data in which one or more classes are poorly represented. In this case it was the consequence of there being very few class 'f' sites in the database. One way to overcome the problem would have been to repeat the class 'f' (and even class 'e') data in the training set as many times as necessary to provide adequate representation. Clearly, repeated data are poor substitutes for extra data, but this procedure would have served to improve performance on class 'f'. Plans to remedy this problem and to re-train the networks using two-fold cross validation analysis were abandoned in favour of work on Self-Organising Maps (SOM), which were showing considerable promise.

4. UNSUPERVISED NEURAL NETWORKS

4.1 Introduction

A brief introduction to neural networks, including supervised and unsupervised learning, was given in Section 1.4. The main advantage of using unsupervised neural networks is that they do not require target values for their outputs. They simply recognise different patterns in the input data and allocate them to a number of discrete categories. Three different types of network were tested to determine which was best suited to the task at hand. These were:

a) Self-Organising Maps (SOM)
b) Adaptive Resonance Theory (ART2)
c) Generative Topographic Mapping (GTM)

It was soon established that the choice was really between SOM and GTM, both of which gave very similar levels of performance in the initial tests. It was therefore decided to proceed with the development of the river quality classifier using SOM and GTM networks.

Since unsupervised networks do not require output targets, it was not necessary to restrict the training data to the database of exemplars. Neither was it necessary to partition the data into training and testing sets. Consequently, the full set of 6038 validated sites was used as the training data. A series of tests were carried out using SOM and GTM based upon 36 output categories (i.e. a square 6x6 map). The input vectors tested included biological inputs only and various mixtures of biological and environmental inputs. The results indicated that the patterns in the combined biological and environmental input vectors were too complex to permit the networks to effectively discriminate between them. Thus it was decided that the effects of environmental factors on community composition would have to be accounted for by developing separate networks for each of the five site types defined earlier.

4.2 Development of Site Specific SOM and GTM Classifiers

Networks having a 10x10 output array (SOM10 and GTM10) were trained for each site type, using combined spring and autumn data (i.e. abundance levels of the 76 BMWP families) as the input vector. The SOMs were trained over 60,000 cycles and the GTMs over 20 cycles (NB. GTM training is a fundamentally different process from that of SOMs), which were roughly equivalent in terms of training effort. Once trained, the networks were used to categorise each of the 6038 sites to one of the 100 output 'bins' associated with its particular site type. On average, therefore, each bin contained approximately 12 sites with very similar community structures. Both the SOM and GTM networks allocated the sites to bins in such a way that the community structures of neighbouring bins were closely related. Thus, any attribute of the sites that is related to community structure (e.g. river quality, alkalinity, DO or the occurrence of a given family) should vary in a relatively smooth and continuous way across the 10x10 output array. A contour map of any given attribute, commonly called a feature map, can be plotted using the attribute's average values within each bin of the output array. To test the relative performance of the SOM and GTM models, the standard deviation of each attribute within each bin was derived. The average standard deviation of the attribute over the 100 output bins was then derived for each site type. This value represented the *noise* between the attribute's feature map and its values within the samples in the bins. Thus, the lower the *noise* the better the fit between the model and the data with respect to that particular attribute. Table 4.1 shows the *noise* levels (i.e. average standard deviations) on the GTM10

and SOM10 feature maps of 13 key attributes (i.e. ASPT, NFAM, three key environmental variables and the top eight taxa). The columns labelled 1 to 5 give the *noise* levels on the attribute's feature maps for site types 1 to 5. Also given is the average noise level, across all five site types, for GTM10 and SOM10. The column at the extreme right hand side of the table gives the ratio of the average *noise* level on the SOM to that on the GTM for each the attributes. The figures in this column highlight the fact that there was negligible difference between the two networks in terms of their overall ability to categorise the data.

Table 4.1 Noise levels (average standard deviations) on the SOM and GTM feature maps of 13 key attributes.

	GTM10 for Site Type:					GTM10	SOM10 for Site Type:					SOM10	Ratio
Attribute	1	2	3	4	5	Avg(G)	1	2	3	4	5	Avg(S)	S/G
ASPT	0.29	0.34	0.33	0.32	0.35	0.33	0.28	0.31	0.33	0.31	0.33	0.31	0.95
NFAM	2.88	3.12	2.98	3.15	3.02	3.03	2.83	3.07	3.06	3.15	3.02	3.02	1.00
ALK	18.96	33.49	38.56	40.28	52.02	36.66	18.58	33.60	39.06	42.47	50.98	36.94	1.01
ALT	60.63	38.99	31.15	25.63	20.01	35.28	58.48	40.37	31.89	26.71	19.20	35.33	1.00
SILT	4.05	7.45	10.91	16.23	25.81	12.89	3.78	6.88	10.92	16.89	25.55	12.80	0.99
Leptoceridae	0.49	0.54	0.47	0.51	0.45	0.49	0.52	0.50	0.47	0.49	0.46	0.49	0.99
Gammaridae	0.63	0.71	0.66	0.60	0.58	0.63	0.63	0.71	0.65	0.63	0.58	0.64	1.01
Elmidae	0.44	0.47	0.46	0.54	0.50	0.48	0.44	0.46	0.47	0.54	0.48	0.47	0.99
Baetidae	0.50	0.52	0.58	0.57	0.58	0.55	0.49	0.54	0.58	0.60	0.61	0.56	1.03
Caenidae	0.45	0.50	0.47	0.51	0.45	0.48	0.48	0.54	0.48	0.52	0.46	0.49	1.04
Hydrobiidae	0.62	0.74	0.67	0.66	0.65	0.67	0.61	0.74	0.70	0.67	0.60	0.66	1.00
Limnephilidae	0.55	0.53	0.56	0.56	0.55	0.55	0.54	0.53	0.57	0.57	0.54	0.55	1.01
Hydropsychidae	0.48	0.59	0.59	0.58	0.48	0.54	0.49	0.56	0.57	0.58	0.47	0.53	0.98
												Avg.	1.00

Both networks produced well-structured feature maps of the key attributes, indicating that they were performing much better than the earlier models. For example, the variations in average ASPT and NFAM across the 10x10 output array of the SOM (site type 2) were smooth and had a good range of values, as can be seen from Figure 4.1.

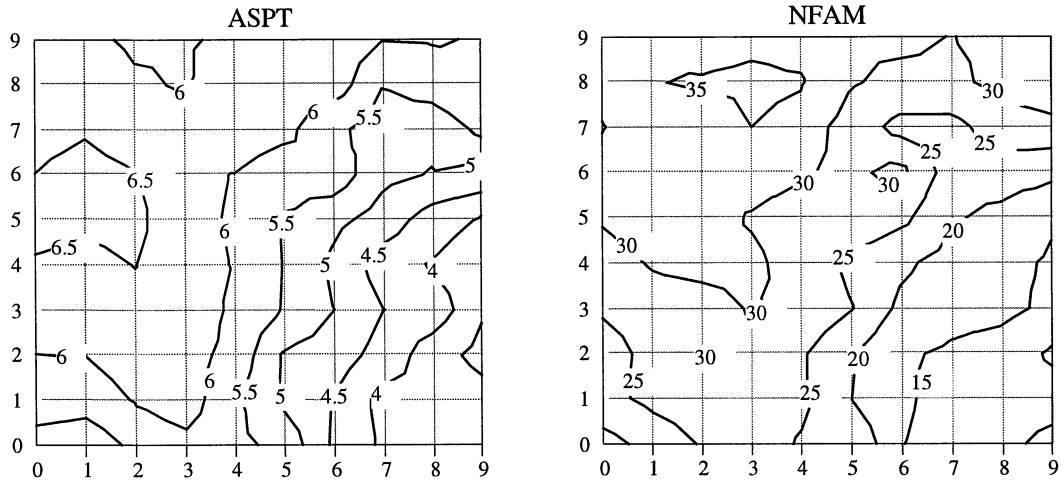


Figure 4.1 Feature maps of ASPT and NFAM produced by the SOM for site type 2.

It should be noted that the vertical and horizontal scales merely serve to provide the 'x' and 'y' co-ordinates of the 100 'classification' bins located at the grid intersections. For these bins to have any real meaning they must first be labelled in terms of their river quality. This is not

a trivial task. It requires the attention of experts to examine the characteristic features of each bin, perhaps with the aid of feature maps.

However, before proceeding further, a choice had to be made about which type of network should be chosen for use as the unsupervised classifier of river quality. It was finally decided to use the SOM network because:

a) it had a long and well-proven record, whereas GTM was a recent innovation;
b) reliable SOM software was readily available (i.e. SOM-PAK); and
c) it performed better than GTM with respect to ASPT (see Table 4.1), the most relevant single feature with respect to river quality.

In view of this decision a brief description is now given of the structure and function of the SOM network. Readers are referred to Kohonen (1995), the originator of SOM, for a detailed account of the theory and application of these networks.

4.3 Structure and Function of Self-Organising Maps (SOM)

A Self-Organising Map (SOM) is an unsupervised neural network in which the output takes the form of a two-dimensional array of output nodes, as shown in Figure 4.2.

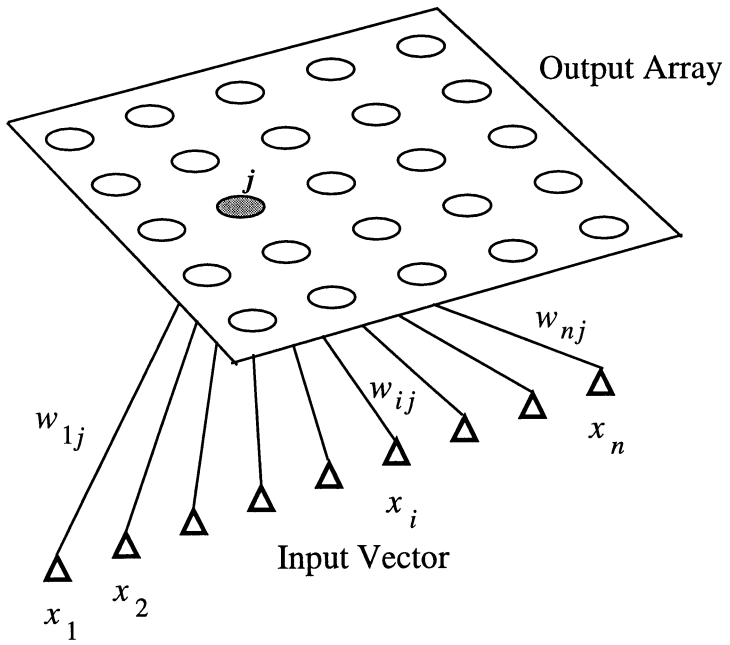


Figure 4.2 Topology of a Self-Organising Map with a 5x5 output array. Note that all output nodes are connected to the input vector, not just node *j* as shown.

Each output node (j) is fully connected to the input vector (x_1, x_2,... x_i,... x_n), and represents a particular pattern in the data, as defined by the set of weights on the links connecting the node to the input vector. That is, the weight vector (w_{1j}, w_{2j},... w_{ij},... w_{nj}) is the exemplar pattern represented by node j. These patterns are determined by the training algorithm during the learning phase. Initially, all weights are randomised, so each output node represents an arbitrary pattern. Representative input data are then presented to the network and compared to the exemplar pattern of each output node to determine which gives the best match. The similarity metric most commonly used to determine the best match is the Euclidean distance:

$$d_j = \left\{ \sum_{i=1}^{n} (x_i - w_{ij})^2 \right\}^{0.5} \quad (7)$$

The exemplar pattern of the winning node, and all nodes in its 'neighbourhood', are then modified to make them slightly closer matches to the input pattern. The modified weights, w'_{ij}, are derived as follows:

$$w'_{ij} = w_{ij} + \alpha_1 \alpha_2 (x_i - w_{ij}) \quad (8)$$

where: α_1 = learning-rate coefficient; and
α_2 = neighbourhood coefficient.

Both of these coefficients are less than or equal to unity and decay with training time. In addition, the neighbourhood coefficient decreases with distance from the winning node, thus the winning node's distant neighbours are modified less than its close neighbours. The neighbourhood coefficient is typically defined by a bell-shaped function with its maximum value (i.e. unity) centred on the winning node. Initially the bell is very wide, covering a large neighbourhood, but as training time proceeds its diameter gradually shrinks, thus confining the neighbourhood to an ever tightening circle around the winning node. Equation (8) therefore has the effect of gradually reducing both the size and spatial extent of the modifications as training proceeds. The final result, when training is complete, is that neighbouring nodes represent very similar patterns and well-separated nodes represent very different patterns. Thus, any individual element of the patterns (e.g. w_{2j},), when plotted on the output array as a contoured map, will produce a well-defined feature map, provided it is an important discriminating factor. If it is not, its 'feature map' will be poorly defined, appearing more like random noise.

4.4 Development of a General SOM Classifier of River Quality

4.4.1 Training and testing

In view of the encouraging results from SOM10, a second attempt was made to develop a general (i.e. all-site-types-in-one) classifier of river quality using a 20x20 output array (SOM20). It was trained over 60,000 cycles using an input vector consisting of the 76 BMWP and 13 environmental variables, with the Euclidean distance as the similarity metric.

Table 4.2 compares the *noise* levels on the feature maps of the 13 key attributes produced by SOM10 and SOM20. The overall average ratio between the two sets of data (i.e. 1.009) indicates that SOM20 produced marginally higher *noise* levels than SOM10. However, it performed noticeably better on ASPT and NFAM which, being composites of the whole taxonomic list, were considered the most important of the 13 key attributes. Thus it was concluded that the single classifier, SOM20, achieved at least the same level of performance as the five site-specific classifiers making up SOM10. An interesting feature of the results shown in Table 4.2 is the change in importance of the environmental variables ALK and SILT from SOM10 to SOM20. ALK is the most important environmental factor in SOM10, because this network was based on the site-type classifications in which ALK was the primary determining factor. However, in SOM20 the importance of SILT was much increased at the expense of ALK. Note the ratios of 1.259 for ALK and 0.727 for SILT, indicating that SOM20 fits the SILT data much better than SOM10, and *vice versa* for the ALK data.

Table 4.2 Average standard deviations across the SOM10 and SOM20 feature maps of 13 important attributes. The SOM10 values are the average across the five site types, as per Table 4.1.

Attribute	SOM10 Avg.(S1)	SOM20 Avg.(S2)	Ratio S2/S1
ASPT	0.313	0.305	0.974
NFAM	3.0238	2.929	0.969
ALK	36.938	46.523	1.259
ALT	35.3286	35.449	1.003
SILT	12.8026	9.307	0.727
Leptoceridae	0.4892	0.498	1.018
Gammaridae	0.6398	0.634	0.991
Elmidae	0.4742	0.486	1.025
Baetidae	0.5642	0.573	1.016
Caenidae	0.4932	0.511	1.036
Hydrobiidae	0.6626	0.692	1.044
Limnephilidae	0.5512	0.562	1.020
Hydropsychidae	0.5346	0.55	1.029
		Average	1.009

The benefit of being able to classify the river quality of a site directly from its biological and environmental data, without first having to classify its site type, was seen as a major advantage of SOM20. Consequently, the remainder of the work on the development of an unsupervised classifier of river quality was focused on the SOM20 network.

4.4.2 Production of feature maps

The final step in the development of a SOM is the allocation of meaningful labels to each of its output bins. Not all bins need to be given unique labels, since it may not be possible or even desirable to give separate labels to all bins, since the difference between some, in terms of river quality, may be negligible. Indeed, if SOM20's 400 bins are to be divided into just six GQA classes, then some classes will clearly be represented by 100 or more bins. Nevertheless, careful examination of the common characteristics of the sites allocated to each bin may make it possible to identify bins that represent particular types of pollution. If this can be achieved, then the network will serve not only as a classifier of biological GQA class, but also as a diagnostic tool capable of identifying problems due to specific pollutants. In an attempt to facilitate the interpretation of SOM20's output array, and thus to aid the labelling of its nodes, feature maps were produced of 97 attributes. These are presented as coloured contour maps in Appendix D. The 97 attributes include: site type, 13 environmental variables; three chemical variables (BOD, DO and ammonia); ASPT and NFAM; GQA class, as derived by RIVPACS; GQA class, as derived from the neural network predictions of ASPT and NFAM based on revised scores; and the 76 BMWP families.

4.4.3 Analysis and interpretation of feature maps

Before attempting to analyse the feature maps it is necessary to have a clear understanding of what the output array represents. Basically, it provides a means of visualising multi-dimensional data in two dimensions. Each node in the output array represents a cluster of data points, but not necessarily on the bases of one node per 'natural' cluster, because the pre-defined number of nodes (e.g. 400 in SOM20) may differ from the number of 'natural' clusters in the data. Thus a large cluster may be allocated to several nodes, each one representing a particular region of the cluster. However, the SOM's training algorithm arranges the nodes in the two-dimensional output array such that neighbouring nodes represent closely related clusters or sub-clusters. That is, clusters that are near neighbours in data space (i.e. data having very similar patterns) are near neighbours in the output array. Nodes that are well-separated in the output array represent very dissimilar patterns, or clusters that are well-separated in data space.

In our case the patterns in the combined biological and environmental data represent different combinations of biological communities and site types, that can be interpreted as symptomatic of different states of health of the river. The network has sorted out the patterns into an ordered two dimensional arrangement, but has left us to define the meaning of each pattern in river quality terms. The feature maps help us to understand the logic behind the arrangement of the patterns, by allowing us to see how various attributes of the patterns are distributed across the array.

The feature maps shown in Appendix D are represented as coloured topographical maps. For the sake of clarity of the map they do not show the 400 output nodes from which the contours are derived. The nodes are in fact located at the intersections of the set of grid lines that would be formed if lines were drawn vertically through the 20 points on the x axis and horizontally through the 20 points on the y axis. The numerical values represented by each node are the particular attribute's mean values derived from the samples that were allocated to that node.

The most informative feature map shown in Appendix D is the one on page D-2 showing the distribution of Silt. This clearly shows that the SOM has allocated the very silty sites (i.e. sites that one would associate with pools) to a distinct group of nodes in a triangular region on the right hand side of the output array. The maps for Altitude, Boulders and Slope, also on page D-2, clearly show that the upland, rocky, steep sloping sites (i.e. sites normally associated with riffles) are represented by the nodes in the top left hand portion of the array. The map of Slope shows that the nodes representing the gently sloping sites occupy the bottom-left to top-right diagonal of the array. These findings are confirmed by the Site Type feature map on page D-1 which shows that:

a) Type 1 sites (upland riffles) occupy a compact area in the top-left of the map; and
b) Type 5 sites (lowland pools) occupy three regions of the map, a triangular area on the right, a smaller area on the bottom-left and an isolated area towards the centre.

The other site types form contour bands between site types 1 and 5, with a diagonal tendency roughly parallel to the low slope band.

On page D-3, the maps for Discharge Category, Distance from Source, Width and Depth show distributions that are clearly related to one another and to the distribution of Slope on

page D-2. The relationship between the distributions of Width and Depth indicate that high depth to width ratios are associated with silty substrates, as one might expect.

On page D-4, the map of Alkalinity shows a close inverse relationship with that of Altitude on page D-2. The distributions of BOD, Ammonia, ASPT and NFAM provide a good indication of which nodes represent the good river quality sites and which represent the poor ones. There is a large group of nodes in the upper left hand region of the array that represent good quality sites with high percentages of boulders and pebbles (i.e. riffles), and there is a small group of nodes located around grid points (14, 7) to (14, 11) in the pool region of the array that have relatively high ASPTs and NFAMs, thus indicating good quality pools. Poor quality sites appear to be represented by a large group of nodes in the bottom right, and a smaller group at the top right.

The first two maps on page D-5 show the distribution of biological GQA classes as defined by: a) RIVPACS; and b) the neural networks predictors of ASPT and NFAM based on revised scores. These confirm the conclusion about the distribution of river quality drawn from the maps on the previous page, but also indicate that there is a third region of 'poorer' quality represented by a small group of nodes in the top left hand corner of the array. The fact that these represent low alkalinity upland streams having low NFAM values implies that they are probably sites affected by acidification. The two GQA maps provide a first approximation of which nodes represent the six GQA classes, and a visual representation of the differences between the 'RIVPACS' and 'Neural Network (RScrs)' methods of classification. The distribution of the classes appears complex because different site types are represented by different regions of the map, and all six GQA classes have to be represented within each site type.

The remaining maps on page D-5, together with those on the following 12 pages, show the distributions of the 76 BMWP taxa in terms of their average abundance levels. It is interesting to compare the distributions of the top indicator taxa (as determined in Section 2.1) with the distribution of GQA classes. The close similarity between the distribution of Elmidae (the top indicator taxon) with that of the GQA classes is quite astounding. Other very similar distributions are displayed by Hydropsychidae, Ephemerellidae and to a lesser degree Baetidae and Leptoceridae. Several taxa exhibit distributions that relate closely to the distribution of good quality riffles, for example Heptageniidae, Leuctridae, Perlodidae, Lepidostomatidae and Sericostomatidae.

Some taxa show close relationships to physical or chemical attributes. For example, many taxa have distributions that are closely related to the distribution of sites with low Slope. These include Neritidae, Viviparidae, Unionidae, Platycnemidae, Coenagriidae, Aphelocheiridae, Notonectidae, Phryganeidae and Molannidae. Others, like Physidae, Planorbidae, Sphaeriidae, Glossiphoniidae, Erpobdellidae and Asellidae appear to be more closely associated with Alkalinity than any other single physical or chemical attribute.

There are also some interesting similarities and differences to be found in the distributions of taxa both within and between groups, like the mayflies, stoneflies and caddis flies. For example, note the difference between Baetidae, Heptageniidae and Caenidae, and the similarity between Ephemerellidae and Sericostomatidae.

There are many other interesting features to be found in these pages - too many to be listed here. Suffice to say that the feature maps of SOMs provide a powerful means of visualising multi-dimensional data.

4.4.4 SOM viewer on the Web

To enable readers to compare any two feature maps, the authors have provided a SOM viewer on the Web (http://www.soc.staffs.ac.uk/research/groups/cies/somview/somview.htm). This enables any two of the 97 feature maps to be viewed alongside each other. Please note that this Web page uses frames and that not all Web browsers are frame capable. Frame capable browsers include all recent versions of Netscape and MS Internet Explorer.

5. NAIVE BAYESIAN MODELS

5.1 Introduction

A brief introduction to methods of probabilistic reasoning, including the naive Bayesian approach, was given in Section 1.5. The first probabilistic classifier of biological river quality was a naive Bayesian model developed by Walley *et al.* (1992b). This model performed well in tests, albeit on a relatively small dataset of 300 sites. A later version of the same model was shown to out-perform a range of other AI-based models (Walley and Džeroski, 1995). It was for this reason that it was decided to develop a naive Bayesian classifier in this project.

5.2 The Mathematics of Naive Bayesian Inference

For the purpose of this section, and in the interests of mathematical simplicity, let us for the moment ignore the effects of site type and season on the biological community. Under these circumstances the naive Bayesian approach defines our six river qualities (a-f) as a set of exhaustive and mutually exclusive classes, Q_i (i = 1 to 6), and uses the evidence provided by the 76 BMWP taxa (the witnesses), each having five possible states of existence, k (i.e. absent plus abundance categories 1 to 4). The method is based upon the assumption that the states of existence of the taxa are conditionally independent (i.e. independent within each quality class, but not necessarily between them). Evidence is combined using the standard equation for naive Bayesian inference:

$$P(Q_i|e_1,..e_j,..e_{76}) = \frac{P(Q_i)}{P(E_{76})}\prod_{j=1}^{76} P(e_j|Q_i) \qquad (3)$$

where:

$P(Q_i|e_1,..e_j,..e_{76})$ = probability that river quality (Q_i) is equal to the i th quality class given the evidence (e_1,..e_j,..e_{76}) provided by the sample;

$P(e_j|Q_i)$ = probability of evidence (e_j) given river quality class (Q_i);

$P(Q_i)$ = prior probability of river quality class Q_i;

$P(E_{76})$ = prior probability of evidence (e_1,..e_j,..e_{76});

e_j = evidence provided by taxon j (i.e. its current state of existence k)

The prior probabilities, $P(Q_i)$, differ considerably from class to class, since class Q_2 (i.e. class 'b') occurs far more frequently than class Q_6 (i.e. class 'f'). This has the effect of causing the model's predictions to favour commonly occurring classes at the expense of infrequently occurring ones. Although this may give slightly better overall performance in terms of the percentage of correct classifications, it produces non-uniform accuracy across the classes, resulting in poor performance on less commonly occurring classes (i.e. 'd', 'e' and 'f'). On the other hand, if one applies the Principle of Indifference, by assuming that the prior probability of each class is the same (i.e. $P(Q_i) = 1/6$ for all i), this has negligible difference on overall performance (at least for our particular problem with its 76 witnesses) but results in approximately uniform precision across the classes. Quite apart from this obvious benefit, there is good reason to judge each case solely on the merit of the evidence presented, totally

ignoring expectations from prior experience. This is the same principle that tries to overcome prejudice in our judicial and appointments processes.

If the Principle of Indifference is applied to equation (3) it gives:

$$P(Q_i|e_1,..e_j,..e_{76}) = \frac{1}{N}\prod_{j=1}^{76} P(e_j|Q_i) \tag{4}$$

where: $N = 6P(E_{76})$.

The apparently difficult task of estimating N is made trivial by the fact that the sum of the probabilities of the river quality being in each of the six classes is unity. That is:

$$\sum_{i=1}^{6} P(Q_i|e_1,..e_j,..e_{76}) = 1 \tag{5}$$

By summing equation (4) over all six quality classes, it follows that:

$$N = \sum_{i=1}^{6}\left\{\prod_{j=1}^{76} P(e_j|Q_i)\right\} \tag{6}$$

N is then substituted into equation (4), thus permitting the evaluation of the probability, $P(Q_i|e_1,..e_j,..e_{76})$, of each river quality class i, given the evidence from the taxa.

Although this looks a complex procedure, N is simply a normalisation constant that ensures that the predicted probabilities across the classes sum to unity. Thus, the whole process is quite straightforward once the conditional probabilities, $P(e_j|Q_i)$, have been derived.

The conditional probabilities, like the family scores in the BMWP system and the saprobic indices in the saprobic system, are constants associated with each taxon, except that in this case there are 30 for each taxon (i.e. one probability per quality class and state of existence).

5.3 A Simple Example

Figure 5.1 illustrates, in diagrammatic form, the process by which evidence is combined. In the interest of simplicity, this example is based upon just five indicator taxa, each having two states of existence (i.e. present and absent). In the example, four of the taxa are present (Gammaridae, Heptageniidae, Baetidae and Erpobdellidae) and one absent (Asellidae). The conditional probability distributions of the five taxa are shown stacked upon each other. In each case the distribution shown is the one appropriate to the taxon's state of existence, hence the one for Asellidae shows the probability of it being absent from each of the six quality classes. The height of each shaded column indicates the probability of the absence of Asellidae from the particular class. In the other cases, it indicates the probability of the taxon's presence in the class. Note that the probabilities in these distributions do not sum to unity. This is because they are distributions of the probability of the taxon's state given the river quality class (e.g. the probability of Gammaridae being present in class 'b' is shown as about 0.90).

The process of combining the evidence (i.e. the probability distributions) proceeds as follows.

1) For each quality class, multiply the probabilities corresponding to each of the five taxa (i.e. the heights of the five columns standing immediately above each other). Since these are all less than unity, their product may be a very small fraction, as shown in the distribution labelled "Products of Columns".

2) If we apply the Principle of Indifference, as suggested earlier, the "Products of Columns" distribution is identical in shape to the required distribution, except that it does not sum to unity. Thus all that is now required is to normalise the distribution to ensure that it does sum to unity. That is, sum the six products to determine the value of the normalisation constant N, then divide the product of each column by N to determine the probabilities of the six classes, $P(Q_i|e_1,..e_j,..e_{76})$, shown as $P(Q|E)$ in Figure 5.1 for brevity.

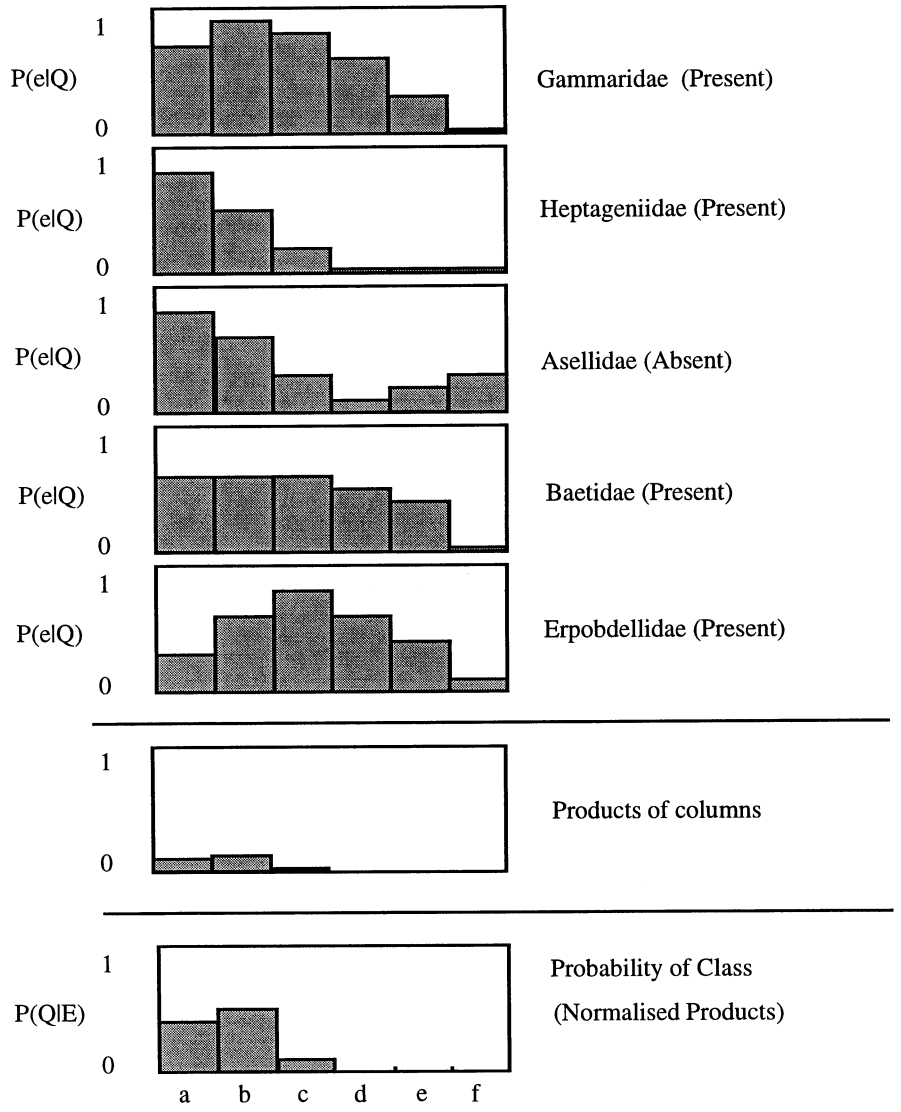


Figure 5.1 A simple example of mechanics of naive Bayesian classification of river quality using evidence from indicator taxa.

In this particular example the final conclusion is that the river quality is most probably class 'b' (50%), but perhaps class 'a' (40%) or even class 'c' (10%). When only a few indicator taxa are used the conclusion is generally not very conclusive, as in this case, but when many more are used the result is generally quite conclusive.

5.4 Avoiding Brittle Behaviour

If the conditional probability matrices $P(e_j|Q_i)$ contain some elements that are zero, then the system is likely to behave in a brittle way. That is, on occasions it will 'crash' and fail to give a result. This happens when something occurs which the probability matrices believe cannot occur. That is, when each of the six columns being multiplied out contains at least one zero probability. For example, suppose a sample has 20 taxa, 19 of which indicate a river quality of class 'a' (80%) or class 'b' (20%), but the twentieth taxon (Chironomidae in abundance category 4) indicates class 'f'. If the probability distribution for Chironomidae shows that the probability of it occurring in class 'a' or class 'b' in abundance category 4 is zero, then the system will crash because an apparent impossibility has occurred. If, however, its probability distribution indicates very small probabilities of Chironomidae occurring in classes 'a' and 'b', then the system will most probably conclude that the river quality is class 'a', or possibly class 'b', because the weight of evidence from the nineteen taxa overpowers that of Chironomidae. This is a very extreme example, and one that would certainly require further investigation if it occurred in reality, but highlights the importance of never allowing zero probabilities to occur in the conditional probability matrices. They should be replaced by small positive values. This is similar to adopting a sceptical attitude to miracles, ghost stories and great money making schemes. It pays to be sceptical. Indeed, Walley and Džeroski (1995) showed that the elimination of zeros from conditional probability matrices resulted in an improvement in overall performance of naive Bayesian classifiers of river quality. They tested two methods of elimination, details of which are given in their paper, and found that they resulted in very similar improvements in performance.

5.5 Conformity Indices

Walley *et al.* (1992b) showed that a benefit of the Bayesian approach is that it permits the evaluation of a probability-based conformity index (C_j) for each taxon (j), defined as:

$$C_j = \frac{\text{Probability of taxon being in its current state, given the nature of the rest of the sample}}{\text{Probability of taxon being in its current state, given no evidence at all}}$$

That is:

$$C_j = \frac{P(e_j|e_1,\ldots e_{j-1},e_{j+1},\ldots..e_{76})}{P(e_j)} \tag{7}$$

where:

$$P(e_j) = \sum_{i=1}^{6} P(Q_i)P(e_j|Q_i)$$

$$P(e_j|e_j,\ldots e_{j-1},e_{j+1},\ldots e_{76}) = \frac{P(E_{76})}{P(E_{75}^j)}$$

and

$$P(E_{75}^j) = P(e_j,\ldots e_{j-1},e_{j+1},\ldots e_{76})$$

The probability of the evidence, $P(E_{76})$, can be derived from $N = 6P(E_{76})$, as stated in equation (4). $P(E_{75}^{j})$ is the probability of the evidence minus the j th item (e_j). It can be derived by undoing the contribution of $P(e_j|Q_i)$ in equation (6).

The index is greater than unity when the evidence given by the taxon conforms with that given by the rest of the sample, and is less than unity when the two conflict. If the index is very low (say < 0.4), the taxon's state is sufficiently inconsistent with the rest of the sample to be considered exceptional. Thus, by deriving conformity indices for each taxon (i.e. present or absent) it is possible to identify unexpected presences and absences for any given quality class, site type and season. Furthermore, a conformity index for the sample as a whole can be derived by averaging the conformity indices of all the taxa. Averages greater than about 1.4 indicate that the sample has a consistent community composition, whereas those less than about 1.1 indicate that the sample has a fair degree of inconsistency in its biological composition. It is important to note that the unexpected absences and presences are not determined relative to a predicted reference-state community, as in RIVPACS, but relative to the rest of the actual community in its actual river quality state. Such information can prove invaluable when trying to identify specific pollutants. The system thus identifies oddities in the actual sample, not the deviation of the samples from some predicted 'ideal' or reference-state community. The latter could readily be determined by the Bayesian system, but it would contribute nothing to the classification process.

5.6 The Models Tested

Naive Bayesian models were developed to classify 'organic' river quality class (i.e. as defined in Section 2.4) for each of the five site types. 'Organic' river quality was used instead of the biological GQA class, because the models required reliable target classifications from which to derive their conditional probabilities. The GQA classifications were not considered sufficiently reliable for this task (section 2.4), owing to the unreliability of NFAM predictions (Section 3.3.7), and hence of the EQI(NFAM) component of the biological GQA classification.

The first model developed was based upon combined spring and autumn biological data, but models were later developed for spring and autumn separately. This was done because the combined samples did not represent true biological communities and were therefore not necessarily consistent in terms of their community compositions. Any such inconsistency would undermine the ability of the conformity index to correctly identify unexpected occurrences and absences.

The probability matrices required for the combined-season models, $P(e_j|Q,T)$, were conditioned on river quality class (Q) and site type (T), whereas those required for the spring and autumn models, $P(e_j|Q,T,S)$, were further conditioned on season (S). All of these conditional probabilities were derived from the database of exemplar sites as described in Section 2.6.

Once developed, the models were used to classify the 'organic' river quality of each of the 6038 validated sites, and the results compared with the 'organic' classification derived from EQI(ASPT)s derived by RIVPACS. Since neither of these classifications can be considered as

absolutely correct, the results have been expressed in terms of the percentage agreement between the two. Table 5.1 gives the results for the combined-season model.

Table 5.1 Percentage agreement between 'organic' river quality classifications given by the naive Bayesian (combined-season) model and RIVPACS III

Site Type	Percentage agreement by 'organic' river quality class						Average per class	Overall rate
	A	b	c	d	e	f		
1	83.9	60.0	79.8	88.7	88.4	100.0	83.5	75.0
2	82.5	66.9	75.9	77.1	85.6	91.7	80.0	76.8
3	81.0	64.5	72.9	75.6	82.1	100.0	79.4	74.8
4	77.2	71.6	71.5	80.4	79.6	87.5	78.0	74.9
5	78.9	67.3	64.4	63.5	67.9	85.7	71.3	68.8
Avg	81.1	65.6	71.4	75.5	81.5	92.7	78.0	74.1

The benefits of applying the Principle of Indifference are apparent in the relative uniformity of 'accuracy' across all classes, giving an overall average per class of 78%. Compare this with the distribution produced by the neural network classifier given in Table 3.16, where a very similar overall rate (i.e. 75%) was achieved but with a much lower overall average per class (i.e. 61.2%). Although these two cases are not strictly comparable, they do serve to illustrate the point that higher overall performance does not necessarily give an acceptable distribution of accuracy across the classes.

Table 5.2 and 5.3 compare the Bayesian (spring and autumn) and RIVPACS classifications of the 6038 validated sites. The overall levels of agreement between the two systems were 67.1% in spring and 65.7% in autumn. The tables reveal that 96 of the spring and 192 of the autumn Bayesian classifications were two or more classes below their RIVPACS classification, whereas only 29 and 20 respectively were two or more classes above their RIVPACS classification. This imbalance between downgradings and upgradings is due to the fact that the RIVPACS classification, being based on combined samples, tends to give an optimistic view of the year as a whole. That is, the combined samples are more representative of the better of the two individual samples and therefore tend to dominate the classification.

Table 5.2 Comparison between the Bayesian (spring) and RIVPACS classifications of 'organic' river quality of the 6038 validated sites.

RIVPACS 'Organic' Class	Bayesian 'Organic' Class (Spring)						Total
	a	b	c	d	e	f	
a	1492	403	51	0	1	0	1947
b	419	1000	319	19	5	1	1763
c	21	216	767	194	14	1	1213
d	0	5	111	471	106	4	697
e	0	0	3	66	285	24	378
f	0	0	0	0	4	36	40
Total	1932	1624	1251	750	415	66	6038

Table 5.3 Comparison between the Bayesian (autumn) and RIVPACS classifications of 'organic' river quality of the 6038 validated sites.

RIVPACS 'Organic' Class	Bayesian 'Organic' Class (Autumn)						Total
	a	b	c	d	e	f	
a	1477	370	91	8	1	0	1947
b	418	990	307	40	4	4	1763
c	20	242	704	211	35	1	1213
d	0	8	127	471	83	8	697
e	0	0	4	68	290	16	378
f	0	0	0	0	3	37	40
Total	1915	1610	1233	798	416	66	6038

Table 5.4 compares the Bayesian spring and autumn classifications. The overall level of agreement between the two seasons was 64.0%. There distributions across the classes were very similar, although autumn produced slightly fewer class 'a' and 'b' sites than spring, but in percentage terms the difference was negligible. There were, however, some major changes in site classification between the two seasons. The table shows that 98 spring samples were classified as two or more classes below their autumn classification, and 170 autumn samples were classified as two or more classes below their spring classification. The most extreme case was Aller Brook (30m d/s ft br Aller Orchard) in South West Region (Devon and Cornwall Areas) which was classified as 'f' in spring and 'a' in autumn.

There is clearly advantage to be gained from having separate spring and autumn classifiers, if only for operational reasons.

Table 5.4 Comparison between the Bayesian (spring) and Bayesian (autumn) classifications of 'organic' river quality of the 6038 validated sites.

Bayesian 'Organic' Class (Autumn)	Bayesian 'Organic' Class (Spring)						Total
	a	b	c	d	e	f	
a	1472	386	54	0	2	1	1915
b	366	955	270	17	2	0	1610
c	85	254	694	182	17	1	1233
d	9	27	196	445	117	4	798
e	0	2	33	96	263	22	416
f	0	0	4	10	14	38	66
Total	1932	1624	1251	750	415	66	6038

The Bayesian classifiers not only give the 'organic' river quality class of the site in terms of a probability distribution, but also the conformity index of the sample and details of any unexpected occurrences or absences of taxa in relation to the rest of the community. Taxa with conformity indices of less than 0.4 were considered to be in an unexpected state of existence. That might mean unexpectedly absent, unexpectedly present, or an unexpectedly high or low level of abundance.

Table 5.5 gives sets of such results for two sites surveyed in 1995, Pearl Brook and Thurgarton Beck. Details are also given of the spring and autumn samples and the site characteristics. These sites were selected to illustrate the value of the conformity index and the probability-based classification, hence they are presented as unusual cases, not typical ones. Five taxa were identified as unexpectedly present, all having conformity indices of less than 0.04, except Ancylidae with 0.36. Just one taxon was identified as unexpectedly absent, Gammaridae at Thurgarton Beck in both spring (C_j = 0.39) and autumn (C_j = 0.36). Since Ancylidae and Gammaridae were only just below the 0.40 threshold used for the identification of exceptional cases, they are not as obviously exceptional as the other taxa. In fact, it is difficult to see why the occurrence of Ancylidae was identified as exceptional. It may be the result of using conditional probabilities derived directly from field data, without smoothing the distributions first.

Table 5.5 Selected results from the naive Bayesian classifier. Note that the number in front of each taxon is its abundance category based upon a $\log_{10}$ scale. Taxa shown in bold type had conformity indices of less than 0.40.

Pearl Brook		Thurgarton Beck	
Altit. 100 m, Dist. from source 3.6 km,		Altit. 20 m, Dist from source 5.0 km,	
Alk. 88.5 mg/l, Width 5 m, Depth 20 cm,		Alk. 262 mg/l, Width 1 m, Depth 8 cm.	
Blds 3%, Pebls 30%, Sand 0%, Silt 67%.		Blds 30%, Pebls 40%, Sand 20%, Silt 10%	
Biological GQA class 'e'		Biological GQA class 'c'	
Spring Sample	*Autumn Sample*	*Spring Sample*	*Autumn Sample*
1 Hydrobiidae	2 Oligochaeta	2 Hydrobiidae	2 Hydrobiidae
1 Sphaeriidae	1 Erpobdellidae	2 Lymnaeidae	2 Lymnaeidae
1 Oligochaeta	1 Asellidae	**1 Ancylidae**	1 Ancylidae
1 Glossiphoniidae	1 Baetidae	1 Sphaeriidae	1 Sphaeriidae
1 Erpobdellidae	**1 Leuctridae**	4 Oligochaeta	2 Oligochaeta
1 Asellidae	**1 Limnephilidae**	1 Glossiphoniidae	1 Glossiphoniidae
1 Baetidae	2 Chironomidae	1 Baetidae	1 Baetidae
2 Hydropsychidae		**1 Leptophlebiidae**	2 Caenidae
2 Chironomidae		1 Dytiscidae	1 Haliplidae
		1 Tipulidae	2 Dytiscidae
		3 Chironomidae	1 Elmidae
			1 Sialidae
			1 Hydroptilidae
			2 Tipulidae
			2 Chironomidae
		Missing Taxon	*Missing Taxon*
		Gammaridae	**Gammaridae**
Classification (Prob)	*Classification (Prob)*	*Classification (Prob)*	*Classification (Prob)*
Class 'a' (0.00)	Class 'a' (0.00)	Class 'a' (0.00)	Class 'a' (0.04)
Class 'b' (0.00)	Class 'b' (0.00)	Class 'b' (0.01)	Class 'b' (0.25)
Class 'c' (0.02)	Class 'c' (0.01)	Class 'c' (0.80)	Class 'c' (0.71)
Class 'd' (0.91)	Class 'd' (0.42)	Class 'd' (0.19)	Class 'd' (0.00)
Class 'e' (0.07)	Class 'e' (0.37)	Class 'e' (0.00)	Class 'e' (0.00)
Class 'f' (0.00)	Class 'f' (0.20)	Class 'f' (0.00)	Class 'f' (0.00)
Avg. Conf. Ind. 1.05	Avg. Conf. Ind. 1.05	Avg. Conf. Ind. 0.99	Avg. Conf. Ind. 1.00

The degree of certainty that the model attaches to its classification is easily seen from its probabilistic output. Samples with a high average conformity index generally result in a classification in which the probability of the predicted class is greater than 0.9. Inconsistent samples, however, tend to produce less conclusive classifications, their distributions often span two or more classes, as can be seen in one of the four cases given in Table 5.5.

6. SUMMARY AND GENERAL DISCUSSION

The performance of the various AI models developed during the project were discussed in the appropriate sections, together with the results of their predictions/classifications and any issues arising therefrom. This section highlights the key findings and discusses the issues that they raise. It also provides pointers to those sections where fuller discussion can be found.

6.1 Key Findings

6.1.1 Prior knowledge

The key findings from earlier AI studies (Section 1.1), that are not incorporated in any existing biological monitoring systems, are as follows.

- Expert river ecologists use two complimentary mental processes when directly interpreting biological data: plausible reasoning, based on their scientific knowledge; and pattern recognition, based on their experience.
- Bioindicator data are inherently uncertain in what they imply about river quality (Walley, 1994; Walley and Fontama, in press). For example, the presence of a given taxon does not indicate a specific quality, but a range of possible qualities (Walley and Martin, 1997). This uncertainty in meaning has important implications for the design of interpretation systems (1.3). Its elimination by the use of representative values (e.g. BMWP family scores) and averages (e.g. ASPT) simply results in loss of information.
- Two AI techniques are particularly well suited to the modelling of the mental processes used by experts, namely Bayesian reasoning (1.5) and neural networks (1.4).
- The absence of a taxon from a sample often provides valuable information, thus biomonitoring systems should incorporate data on the absence of bioindicator taxa as well as their presence (Walley and Fontama, in press).
- Different levels of abundance (including the absence) of a bioindicator can differ noticeably in both the weight (1.6.5 and 2.1), and meaning of the evidence they provide (Walley and Fontama, in press).
- River quality is a concept having so many complex dimensions that it can only be conceived and defined in terms that are essentially subjective, thus all river quality classification systems, whether biological or chemical, have their subjective components. The issue to be addressed is how to minimise any detrimental impact of this component on system performance (1.6.4).
- There are no absolute standards for river quality, so existing classifications do not provide ideal examples on which to base the development of new systems.

6.1.2 Indicator values

Information theory was used to define the indicator values of the 76 BMWP families for both presence/absence and abundance-level (including absent) data. The key findings of the study were as follows.

- The term 'indictor value' is used fairly loosely, despite the fact that it can have two very different meanings (2.1.1), a conditional value and an unconditional value. The first has value as a weighting coefficient in systems like the saprobic system, whereas the second has a much more general meaning in that it indicates the taxon's overall value as a sensor of river quality, irrespective of any specific site or sample. This value has significance in relation to the overall design and optimisation of monitoring systems.
- Information theory (2.1.2) provides a means of evaluating indicator values (conditional or unconditional) of bioindicators in terms of the value of the information they provide in relation to some desired classification. Unconditional indicator values were derived using national presence/absence and abundance data from *Riffles* and *Pools* (2.1.3 and Appendix A).
- Indicator values differed between *Riffles* and *Pools*, and only six families were common to the top 20 from each site type, namely (in rank order) Elmidae, Leptoceridae, Baetidae, Caenidae, Gammaridae and Ephemeridae (Appendix A).
- Some families gave much higher indicator values when based on abundance data rather than just presence/absence data, notably Oligochaeta, Chironomidae, Asellidae and Erpobdellidae (2.1.3 and Appendix A).
- Two non-BMWP families, Hydracarina and Ceratopogonidae, gave indicator values that placed them in the top 50 indicator families (2.1.3), and it is suggested that they should be added to the list of BMWP families.
- Indicator values based upon regional subsets of the national data showed a high degree of similarity between the rank orders of families in most regions, with the exception of Anglia and Thames which, although similar in themselves, were noticeably different from the rest (2.1.4).
- The rank order of families based on the average of their regional indicator values was different from that based on the national data, in that some families moved up a several places and others move down. The reason for this appeared to be that some families showed marked differences in indicator value from region to region, which when averaged resulted in a change in their national ranking position (2.1.4). This was most probably due to spatial variations in the species mix within the families in question.

6.1.3 Supervised-learning networks

Supervised-learning networks were developed to: classify site type; predict 'unpolluted' ASPT and NFAM; predict BOD, DO and ammonia; and classify 'organic' river quality. However, the first two were investigated in greater depth than the other two. It should be noted that the predictors of ASPT and NFAM were not developed as part of the mainstream AI approach to biomonitoring, but as a means of comparing the capabilities of neural networks with a recognised statistical model (RIVPACS) in a like-for-like way using the same dataset. The key outcomes from these studies were as follows.

Site Type Classifiers (3.2)

These were developed to enable the validated sites to be partitioned into approximately equal subsets on the basis of their environmental characteristics.

- The site types were classified into five categories, labelled 1 to 5, by banding predicted 'unpolluted' ASPTs such that each site type had an approximately equal number of sites. Three neural network classifiers were developed, each based on a different training set but the same three-variable input vector consisting of alkalinity, altitude and substrate (percentage sand plus silt).
- The 6038 validated biological sites were classified into site types using a consensus of neural network classifiers and a classifier based upon RIVPACS predictions of ASPT. The geographic distributions (Figures B1-B5)) of the five site types indicate that they vary from upland riffles (type 1) to lowland pools (type 5). The precise nature of these five types is best illustrated graphically in terms of their altitudes, alkalinities and substrate compositions (Figures B6-B9).

Predictors of ASPT and NFAM (3.3)

- Preliminary tests carried out on several different types of network showed that the most suitable network for the task was the standard back-propagation network (3.3.3).
- Irrelevant environmental variables were removed from the input vector using impact analyses (3.3.4). These showed that the ASPT predictor required just five inputs (i.e. alkalinity, altitude, percentage silt, $\log_{10}$ of slope and discharge category) and that the NFAM predictor required seven inputs (i.e. distance north, altitude, percentage sand, $\log_{10}$ of distance from source, $\log_{10}$ of slope, river depth and distance east).
- Two-fold cross validation analysis permitted the development and testing of both independent and dependent models (3.3.2). Tests on these models (3.3.7) showed that:
  - the ASPT predictors performed markedly better than the NFAM predictors on both dependent and independent data;
  - the NFAM predictors performed markedly worse on independent data than on dependent data, thus casting further doubt on the reliability of NFAM predictions in practice.
- Comparisons made between the neural networks and RIVPACS III showed that the networks slightly out-performed RIVPACS, and achieved in one non-linear mathematical step what RIVPACS achieved in three steps (3.3.7)
- An ASPT predictor based upon the revised BMWP family scores derived by Walley and Hawkes (1996, 1997) produced a noticeable improvement in the correlation between the predicted and target ASPTs, thus indicating that the revised scores provided a better fit than the original scores (3.3.9).
- An investigation into possible causes of error and bias in the predictions of ASPT and NFAM indicated that:
  - the most likely cause of an apparent spatial bias on a river basin scale (3.3.8 and Appendix C) was the lack of one or more relevant environmental variables in the input data, especially catchment (as opposed to site) variables (e.g. geology).
  - a significant component of the apparent prediction errors may not be errors at all but simply stochastic variations of natural processes (3.3.8).
  - errors were greatest when one or more of the key predictor variable (e.g. alkalinity, altitude and silt) were close to the upper or lower limits of their range (i.e. at or near the edge of data space).

Accuracy of GQA Classifications

- Biological GQA classifications of the 6038 validated sites based upon the neural network predictors of ASPT and NFAM showed 79.8% agreement with the RIVPACS classification when based upon original BMWP scores and 72.5% agreement when based on revised scores, thus indicating that up to about 25% of existing biological GQA classifications may be of questionable accuracy (3.3.11).
- Biological GQA classifications are sensitive to the mathematical method used to predict ASPT and NFAM, even when the methods have the same overall level of accuracy (3.3.11).

Other Supervised-learning Networks

- Networks were trained to predict BOD, DO and ammonia from biological data for each of the five site types (3.4). The correlation coefficients achieved were 0.497, 0.578 and 0.660 for BOD, DO and ammonia respectively.
- Networks were trained to classify 'organic' river quality (3.5) using target classifications derived by two different methods, one based on RIVPACS and one based on a neural network predictor of ASPT (2.4). Separate networks were trained for each of the five site types. These achieved an overall 'correct' classification rate of 75%, but performed poorly on class 'f' due to its under-representation in the data (3.5).

6.1.4 Self-Organising Maps (unsupervised networks)

Initial tests carried out on three different unsupervised networks (i.e. SOM, GTM and ART2) to determine which was the most suitable for the task, showed that SOM and GTM performed better than ART2 and that there was very little to choose between the two (4.1 and 4.2). It was decided to proceed with SOM on the basis that it was well established and had a long record on successful applications. The key findings were as follows.

- The principal advantage of unsupervised networks is that they do not require target values, but this has a cost in that their output categories have no meaning until they have been labelled by experts, which is not a trivial task. SOM and GTM have the advantage that they permit visual interpretation of multi-variate data via feature maps (Appendix D) based upon their output arrays.
- Five SOMs were developed using 10x10 output arrays, one for each site type (4.2). Their input vectors consisted of the 76 BMWP taxa only, since it was assumed that the site types had accounted for the environmental factors.
- A single SOM with a 20x20 output array was developed to cover all site types (4.4). Its input vector consisted of 76 BMWP taxa plus 13 environmental variables. Tests showed that there was very little difference in performance between the five 10x10 site-specific SOMs and the single 20x20 SOM covering all site types. Thus it was decided to proceed with the development of the latter since it had the advantage of not requiring site type classifications.
- Feature maps of 97 attributes of the data were produced (Appendix D), together with a software package, SOMVIEW, that permits any two feature maps to be compared on screen. Examination of the feature maps has highlighted some interesting relationships

(4.4.3) and SOMVIEW clearly has potential as a research tool. A limited version of SOMVIEW is freely available on the Web (4.4.4).

- No attempt was made to label the 400 output categories, because there are improvements that can be made to the SOM's topology and function before attempting this task. These objectives now form part of a new project (National R&D Project E1-056). Preliminary investigations have shown that different parts of the map represent different types of pollution, a property that clearly opens up the possibility of using the SOM for diagnosis, not just classification.

6.1.5 Bayesian classifiers

Although this project was primarily concerned with neural networks, it did include the development of classifiers of river quality based on naive Bayesian inference. This was done to explore the potential of Bayesian methods and to provide an independent measure of the performance of the neural network classifiers. The principal outcomes were as follows.

- Naive Bayesian classifiers of 'organic' river quality were developed for each of the five site types using combined-season data (5.5). They incorporated conformity indices (5.4) to highlight families that are unexpectedly absent or over/under abundant in relation to the composition of the rest of the community. Unfortunately, the lack of biological consistency in the combined-season samples (i.e. they were not true communities) undermined the ability of the conformity indices to identify anomalous states of existence. Nevertheless, the models achieved 74.1% agreement with the 'desired' combined-season values, compared to 75.0% for the equivalent neural network, but produced much better uniformity of accuracy across the classes than the network.

- Naive Bayesian classifiers of 'organic' river quality were developed for each of the five site types using spring and autumn data separately (5.5). This was done to overcome the problem arising from the lack of biological consistency in the combined-season model. These models classify 'organic' river quality in probabilistic terms using the spring and autumn samples separately, and in both cases provide lists of families that are considered to be misfits in the observed community (not relative to a 'reference-state' community), including unexpected absences.

- The spring and autumn models achieved 67.1% and 65.7% agreement with the 'desired' combined-season classification (5.6). The lower figures achieved here (i.e. compared with the 74.1% of the combined-season Bayesian models) are a natural consequence of producing two separate classifications in place of one annual classification, and do not indicate lower performance. The level of agreement between the Bayesian spring and autumn classifications was 64.0%.

- The Bayesian approach provides two distinct advantages. Firstly, it not only handles the inherent uncertainty in the meaning of the data in a mathematically sound way, but also carries it through to its conclusion, thus providing a measure of the reliability of the conclusion. Secondly, the ability of conformity indices to identify unexpected states of existence has potential for use in quality assurance and the diagnosis of specific pollutants.

6.2 Discussion of Main Issues

There are several important issues that arise out of this project. These are mainly concerned with fundamental differences between the AI approach and approaches used by existing methods, but they also include issues relating to: inherent subjectivity; the scope and meaning of quality classes; the best ways forward for classification systems; and the development of diagnostic/ prognostic systems.

6.2.1 Basis of approach

The AI approach is fundamentally different from the RIVPACS approach, in that it is not based upon the concept of a 'clean' reference state, but takes a more holistic approach to 'clean' and 'dirty' water biology. That is, RIVPACS is based only on the biology of 'unpolluted' waters, leaving the BMWP system to provide the basis for the river quality classification via EQIs, whereas the AI approaches is based on the biology of all waters, polluted or unpolluted. No single river quality condition is considered more important than any other, since each is treated as a sort of 'reference' condition in its own right. Thus the problem of how to measure the *absolute* difference between a river's actual state and its 'unpolluted' (or reference) state, in order to define its degree of environmental stress (i.e. as via EQIs in the RIVPACS approach) simply does not arise in the AI approach. The river's actual state is judged relative to several possible 'unpolluted' and polluted (or unstressed / stressed) states, based on the known or recorded biology of those states. The principal reason for taking this approach was that it is basically what an expert with knowledge and experience of 'dirty' water ecology would do, but there are also good analytical reasons for rejecting the metric (i.e. EQI) approach, as will be explained later.

The AI approach also incorporates some key characteristics of biological data that hitherto have been ignored or inadequately represented, namely inherent uncertainty in its meaning and the relevance of absent taxa. The absence of a taxon, like its other states of 'existence', is allowed to contribute evidence about the river's state of health. In all cases, the evidence presented is inherently uncertain, but is more so in the case of absence. It is not absolute in its meaning but vague, and is best thought of in terms of probability distributions. These distributions are non-linear, sometimes bi-modal, and differ in shape between abundance levels (Walley and Fontama, in press). Consequently, the relationship between community composition and river quality turns out to be non-linear and not necessarily uni-modal. These facts severely undermine any reference state approach that is based upon simple metrics of environmental stress (e.g. EQIs). The AI techniques used in the study were able to accommodate both the uncertainty and the variations in meaning in the data.

The Bayesian models represented them explicitly in the form of probability distributions and combined the evidence using probability theory, thus producing conclusions expressed in probabilistic terms.

The neural networks handled the uncertainty implicitly via their inherent ability to generalise by recognising patterns as a whole, and modelled the variations in meaning with respect to abundance level via their ability to perform complex non-linear mappings. One network produced outputs in probabilistic terms, and the others could be made to do so if required.

6.2.2 Subjectivity

The only models developed in the study that could be termed 'objective' were the neural network predictors of BOD, DO and ammonia, because they were based on chemically derived target outputs, but even these contained some subjectivity (e.g. choice of output topology, choice of sampling site and time). However, general river quality by its very nature (1.6.4) dictates that any classification system will contain a significant subjective component. We need to recognise this fact and concentrate on maximising the quality of the subjective component and minimising any detrimental impact it may have on the system's performance. Walley and Hawkes (1996, 1997) demonstrated how the quality of the subjectively derived BMWP scores could be enhanced by data analysis. Even so, the BMWP system, on which the biological GQA classifications are based, remains founded on subjectivity. One interesting feature of the SOM classifier is that it transfers the subjective input from the beginning to the end of the classification process, by 'objectively' categorising the data into different patterns and then leaving experts to define the river quality that each pattern represents. The only subjectivity remaining at the start of the process is the choice of output topology and similarity metric. Intuitively, this transfer of subjectivity to the end of the process seems desirable, since it at least makes any impact on final classifications more transparent.

6.2.3 Scope and meaning of quality classes

Many different terms have been used in relation to the quality classes, including water quality, environmental quality, biological quality, biological condition and river quality. These have all been used rather loosely or at least without a clear definition of meaning. River quality has been used throughout this report in recognition that man-made impacts on community composition do not occur only as a result of pollution of the river water, but also through factors such as contamination of the river bed, destruction of habitats by river engineering works, and the regulation of river flows. However, some natural events and processes can have very similar effects on community composition, but being natural they do not constitute pollution and should not result in a degradation of river quality class. This presents quite a challenge and highlights the need for the collection of relevant environmental data across a wider range of variables than at present. The study has highlighted some current deficiencies in this respect (3.3.8), with reference to bias in predicted ASPTs and NFAMs on a river-basin scale. If the effects of bed contamination, engineering works and flow regulation are to be separated out from natural effects, then urgent action is needed to acquire the data necessary for the construction of the models.

6.2.4 The future of classification systems

The results of this project show that the present biological GQA classification system is sensitive to the type of model used to predict the unpolluted ASPTs and NFAMs (even given the same overall accuracy), and that a change from RIVPACS to neural network predictions would result in 20.1% of sites changing their classification by one class. Furthermore, a revision of BMWP scores to the site-related values derived by Walley and Hawkes (1997) would result in a further 7.3% of sites changing their classification by one class. Thus, if we assume the GQA classification system is perfect except for possible improvements to the BMWP scores and predictions of ASPTs and NFAMs, then it follows that such improvements could result in up to 27% of sites changing their class. However, the GQA classification system is not perfect for reasons relating to its EQI metrics mentioned earlier. Thus the

Agency has to judge whether this situation is acceptable and, if not, to what extent it wishes to improve the system. Does it wish to revise the existing system to gain whatever improvements can be had without embarking on wholesale change, or does it wish to embrace a new approach with potential for greater accuracy and reliability ?

There are several ways in which the existing system can be improved.

- A revision of BMWP family scores would improve the system's ability to discriminate between different organic qualities. This could be done either by adopting the scores derived by Walley and Hawkes (1996, 1997) or, preferably, by carrying out a reappraisal of scores based on the 1995 validated biological data.
- Improvements could be made to the prediction of EQI metrics. In the case of ASPT, this could be achieved using a combination of neural network models and RIVPACS to form a 'committee of experts' that would provide a prediction by consensus. In the case of EQI(NFAM), the less reliable of the two metrics, it is difficult to see how any worthwhile improvement could be made without a fundamental change to its basis. For example, a weighted NFAM could be defined by weighting families according to their sensitivity to toxic pollution. This would remove some of the coarseness of NFAM but do little to reduce its sensitivity to sampling effort. In addition, or alternatively, a 'toxic' equivalent to ASPT could be developed based on a set of family scores that reflect each family's overall sensitivity to 'toxic' pollution.
- The collection of additional environmental data that are relevant to the prediction of ASPT and NFAM would help to improve the predictions, and hence the accuracy of the EQI metrics.
- If family scores were expressed in the form of probability distributions (i.e. similar to saprobic valencies), then the naive Bayesian approach could be used to combine the evidence (including that given by absent taxa) to derive a pseudo-ASPT (i.e. an exact Bayesian equivalent of the average score per taxon). This would have the advantage of improving accuracy and providing a measure of reliability for each individual case.
- In the longer term, once the AI systems have been fully proven under operational conditions, it would be beneficial to replace the existing metric-based classification system by an AI system that takes a more holistic approach to river ecology by incorporating the biology of both 'unpolluted' and polluted waters. The thousands of samples that are collected annually by the Agency contain a wealth of information on the biology of polluted and environmentally stressed waters. The present system inadequately models this biology using a simple metric based upon predicted ASPTs and NFAMs. This represents a considerable waste of valuable information that has been collected at great expense.

6.2.5 Development of diagnostic / prognostic systems

The project has demonstrated that AI techniques have considerable potential for the development of operational tools for the diagnosis of river quality problems. In particular, the results of tests on the Self-Organising Maps (SOM10 and SOM20) showed that these networks have the potential of relating different patterns in the biological/environmental data to specific types of pollution. However, the links between the patterns and the types of pollution are not made by the SOM. It requires experts to create these links by identifying and labelling the river condition represented by each pattern. This is because SOMs are

unsupervised networks that learn in the absence of target outputs. The outputs categories of the SOMs that were developed in this project were not labelled by experts, because it was recommended that further improvements be made to the topology and function of the SOMs before commencing this exercise. These improvements are now being implemented through National R&D Project E1-056.

If the targets are known, as was the case for BOD, DO and ammonia (despite their known inadequacies), then supervised networks can be trained to predict concentrations of specific pollutants. This, however, requires the matching of biological sites to chemical sites having data on the range of pollutants of interest. Machine learning techniques have also been used to infer chemical parameters from biological data (Džeroski *et al.*, 1997a). Unlike neural networks, machine learning attempts to derive knowledge from the data by inducing rules. This has the advantage of making their conclusions more transparent, but they have yet to demonstrate the same level of performance as neural networks.

The ability of the naive Bayesian classifiers to identify anomalies in community composition offers a possible way of identifying specific pollutants. It also provides a powerful means of rapidly screening data for quality assurance purposes. These benefits, together with the ability of Bayesian Belief Networks (BBNs) to model dependencies within complex systems and to reason predictively and diagnostically under conditions of uncertainty, make Bayesian methods the most promising knowledge-based tools available. These systems are so flexible that they can be used for diagnosis or prognosis as the need arises. Furthermore, they are non-monotonic, which means that they can 'change their minds' when new evidence 'explains away' earlier evidence.

Furthermore, the SOM and BBN techniques are complimentary: one making good use of available databases and replicating the pattern recognition skills of experts; and the other making good use of existing scientific knowledge and replicating expert reasoning. It is envisaged that it will be possible in future to combine these two approaches to produce a single overall conclusion, weighted in proportion to the certainty in the separate conclusions drawn by the SOM and BBN.

7. RECOMMENDATIONS AND CONCLUSION

7.1 Recommendations

- That consideration be given to ways of improving the reliability of the biological GQA classification system. The suggested improvements are:
  - the use of neural networks in conjunction with RIVPACS to predict unpolluted ASPTs and NFAMs;
  - to collect additional environmental variables to improve the predictive capabilities of the models;
  - to revise the BMWP family scores through an analysis of the 1995 National Survey data, including the derivation of scores for different site types and abundance levels;
  - to add Hydracarina and Ceratopogonidae to the list of BMWP taxa;
  - to apply a Bayesian approach to the combination of the evidence given by the BMWP taxa (including absent taxa) so as to produce a pseudo-ASPT (i.e. an exact Bayesian equivalent of ASPT); and
  - to introduce a 'toxic' equivalent of ASPT in an attempt to overcome the problems resulting from the unreliability of the EQI(NFAM) metric.

- That action be taken to compile databases of matched biological and chemical sites, incorporating as many relevant environmental and chemical variables as possible, but especially the principal pollutants and other main causes of environmental stress.

- That the work on SOM networks be continued to further improve their topology and internal functions, and to develop and test an operational SOM for use on the diagnosis of river quality problems.

- To further explore the potential of Bayesian methods through a feasibility study involving the development of a prototype Bayesian Belief Network for the diagnosis and prognosis of specific river quality problems.

- In the longer term, when AI methods have proven their worth as diagnostic and prognostic tools, consideration should be given to using them for river quality classification.

7.2 Conclusion

Two techniques from the field of Artificial Intelligence have been shown to offer considerable potential for use in river quality classification and as operational tools for the diagnosis and prognosis of river quality problems. These techniques are neural networks and Bayesian methods of reasoning under conditions of uncertainty. Together they offer the possibility of modelling the mental processes used by expert river ecologists when directly interpreting biological data.

8. ACKNOWLEDGEMENTS

The project was made possible by earlier studies, to which freshwater ecologist Dr. H. A. Hawkes made a major contribution. It was his wealth of knowledge and experience of biomonitoring that provided the ecological basis of the AI approach ultimately developed. Thanks are due to him for those past contributions and for his continuing interest in the project. Thanks are also due to Dr J. Murray-Bligh (Project Manager) from the Environment Agency and Dr. M. Furse of the Institute of Freshwater Ecology for their assistance on various aspects of the project.

9. REFERENCES

Beale R. and Jackson T. (1990) *Neural Computing: An Introduction.* Adam Hilger, Bristol.

Bishop C. M. (1995) *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford.

Boyd M. (1995) *The application of methods of uncertain reasoning to the biological classification of river water quality.* PhD thesis, Aston University, Birmingham.

Boyd M., Walley W. J. and Hawkes H. A. (1993) Dempster-Shafer Reasoning for the Biological Surveillance of River Water Quality. In *Water Pollution II: Modelling, Measuring and Prediction*, (eds. Wrobel L. C. and Brebbia C. A.). CMP, Southampton.

Chong H. G. and Walley W. J. (1996) Rule-based versus probabilistic approaches to the diagnosis of faults in wastewater processes. *Artificial Intelligence in Engineering*, **10**(3), 265-273.

De Pauw N. and Hawkes H. A. (1993) Biological monitoring of river water quality. In *Proceedings of the Freshwater Europe Symposium on River Water Quality Monitoring and Control*, (eds. Walley W. J. and Judd S.), 87-111. Aston University, Birmingham.

Džeroski S., Dehaspe L., Ruck B. M. and Walley W. J. (1994) Classification of river water quality using machine learning. In *Computer Techniques to Environmental Studies, Vol I: Pollution Modeling*, (ed. Zanetti P.), 129-137. CMP, Southampton.

Džeroski S., Demšar D., Grbović J. and Walley W. J. (1997a) Learning to infer chemical parameters of river water quality from bioindicator data. *Proceedings of the 6th Electrotechnical and Computer Science Conference ERK'97*, 129-132. Slovenia Section IEEE, Ljubljana.

Džeroski S., Grbović J, and Walley W. J. (1998) Machine learning applications in biological classification of river water quality. In *Methods and Applications of Machine Learning, Data Mining and Knowledge Discovery*, (eds. Michalski R. S., Bratko I. and Kubat M.), 429-448. John Wiley & Son, Chichester.

Džeroski S., Grbović J, Walley W. J. and Demšar D. (1997b) Reappraisal of bioindicator saprobic values and weights using data from river quality surveys in Slovenia. In *Water Pollution IV*, (eds. Rajar and C. A. Brebbia). CMP, Southampton.

Džeroski S., Grbović J., Walley W. J. and Kompare B. (1996) Using machine learning techniques in the construction of models: II Data analysis and rule induction. *Ecological Modelling*, **95**, 95-111

Giarratano, J., and Riley, G. (1989) *Expert Systems - Principles and Programming.* PWS-Kent, Boston.

Hawkes H. A. (1998) Origin and development of the Biological Monitoring Working Party score system. *Wat. Res.* **32**(3), 964-968.

Haykin S. (1994) *Neural Networks: A Comprehensive Foundation.* Macmillan, New York.

Jenson F. V. (1996) *An Introduction to Bayesian Networks.* UCL Press, London.

Kohonen T. (1995) *Self-Organising Maps.* Springer-Verlag, Berlin.

Lauritzen S. L. and Spiegelhalter D. J. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, **50**(2), 157-224.

Moss D., Furse M. T, Wright J. F. and Armitage P. D. (1987) The prediction of the macroinvertebrate fauna of unpolluted running water sites in Great Britain using environmental data. *Freshwater Biology*, **17**, 41-52.

Neapolitan, R. E. (1990) *Probabilistic Reasoning in Expert Systems*, Wiley, New York.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*, Morgan Kaufmann, San Mateo, California.

Ruck B. M. (1995) *The application of artificial neural networks to the interpretation and classification of freshwater benthic invertebrate communities.* PhD thesis, Aston University, Birmingham.

Ruck B. M., Reynoldson T. B., Day K. E. and Walley W. J. (1996) *Classification and prediction of benthic community structure in the Great Lakes using neural networks.* NWRI Contribution No. 96-56, Environment Canada, Burlington, Ontario.

Ruck B. M., Walley W. J. and Hawkes H. A. (1993a) Biological Classification of River Water Quality using Neural Networks. In *Applications of Artificial Intelligence in Engineering VIII, Vol 2: Applications and Techniques,* (eds. Rzevski G., Pastor J. and Adey R. A.), 361-372. CMP/Elsevier, Southampton.

Ruck B. M., Walley W. J. and Reynoldson T. B. (1993b) A Neural Network Predictor of Benthic Community Structures in the Canadian Waters of the Laurentian Great Lakes. In *Water Pollution II: Modelling, Measuring and Prediction*, (eds. Wrobel L. C. and Brebbia C. A.), CMP, Southampton.

Ruse L. P. (1996) Multivariate techniques relating macroinvertebrate and environmental data from a river catchment. *Wat. Res.* **30** (12), 3017-3024.

Shafer G and Pearl J. (1990) *Readings in Uncertain Reasoning.* Morgan Kaufmann, San Mateo, California.

Sládeček V. (1964) Zur biologischen Gliederung der höheren Saprobitätsstufen. *Archiv für Hydrobiologie*, **58**(1), 103-121.

Walley W. J. (1994) New approaches to the interpretation and classification of water quality data based on techniques from the field of artificial intelligence. In *Proceedings of Monitoring Tailor-made*, (eds. Adriaanse M., Kraats J., Stoks P. G. and Ward R. C.), 195-210. RIZA, Lelystad, The Netherlands.

Walley W. J. (1993) Artificial Intelligence in River Water Quality Monitoring and Control. In *Proceedings of the Freshwater Europe Symposium on River Water Quality Monitoring and Control*, (eds. Walley W. J. and Judd S.), 179-193. Aston University, Birmingham.

Walley W. J., Boyd M. and Hawkes H. (1992a) An Expert System for the Biological Monitoring of River Pollution. In *Computer Techniques to Environmental Studies IV*, (ed. Zanetti P.), 721-736. CMP/Elsevier, Southampton.

Walley W. J. and Džeroski S. (1995) Biological monitoring: a comparison between Bayesian, neural and machine learning methods of water quality classification. In *Environmental Software Systems*, (eds. Denzer R., Schimak G. and Russell D.), 229-240. IFIP Conference Series, Chapman & Hall, London.

Walley W. J. and Fontama V. N. (in press) New approaches to the biological classification of river quality based upon artificial intelligence. In *RIVPACS International Workshop, Oxford, 16-18 September 1997*, (eds. Furse M. T. and Wright J. F.), Freshwater Biological Association, Ambleside.

Walley W. J. and Fontama V. N. (1998) Neural network predictors of average score per taxon and number of families at unpolluted river sites in Great Britain. *Wat. Res.* **32**(3), 613-622.

Walley W. J. and Fontama V. N. (1997) Bio-monitoring of rivers: an AI approach to data interpretation. *Ecotoxicology and Chemistry,* **4**(6), 183-185.

Walley W. J. and Hawkes H. A. (1997) A computer-based development of the Biological Monitoring Working Party score system incorporating abundance rating, biotope type and indicator value. *Wat. Res.* **31**(2), 201-210.

Walley W. J. and Hawkes H. A. (1996) A computer-based reappraisal of Biological Monitoring Working Party scores using data from the 1990 River Quality Survey of England and Wales. *Wat. Res.* **30**(9), 2086-2094.

Walley W. J., Hawkes H. A. and Boyd M. (1992b) Application of Bayesian Inference to River Water Quality Surveillance, In *Applications of Artificial Intelligence in Engineering VII*, (eds. Grierson D. E., Rzevski G. and Adey R. A.), 1030-1047. CMP/Elsevier, Southampton.

Walley W. J. and Martin R. W. (1997) *Distribution of Macroinvertebrates in English and Welsh Rivers based on the 1995 Survey*. R&D Technical Report E12, Environment Agency, Bristol.

Walley W. J. and Martin R. W. (1998) *Applications of Artificial Inteligence in River Quality Surveys.* R&D Project Record E1/i621/6, Environment Agency, Bristol.

Wright J. F., Furse M. T., Clarke R. T., Moss D., Gunn R. J. M., Blackburn J. H., Symes K. L., Winder J. M., Grieve N. J. and Bass J. A. B. (1995) *Testing and Further Development of RIVPACS.* R&D Note 453, National Rivers Authority (now Environment Agency), Bristol.

Appendix A

Information Values of 76 BMWP Taxa

Table A1 Information values of 76 BMWP taxa for 'Riffle' sites listed in order of their $M'(C, X)$ values derived from abundance data.

		$M(C, X)$		$M'(C, X)$		
		Mutual Inform.		Indiff. Mut. Inform.		Improv.
Rnk	Taxon	(Pres)	(Abun)	(Pres)	(Abun)	Ratio
1	ELMIDAE	0.2551	0.3158	0.3809	0.4415	1.16
2	HYDROPSYCHIDAE	0.1591	0.1839	0.3044	0.3309	1.09
3	HEPTAGENIIDAE	0.1769	0.1958	0.2481	0.2688	1.08
4	SERICOSTOMATIDAE	0.1844	0.1894	0.2520	0.2575	1.02
5	RHYACOPHILIDAE	0.1478	0.1571	0.2287	0.2390	1.05
6	LEUCTRIDAE	0.1465	0.1514	0.1955	0.2030	1.04
7	LEPTOCERIDAE	0.1127	0.1149	0.1899	0.1917	1.01
8	LEPIDOSTOMATIDAE	0.1451	0.1458	0.1845	0.1849	1.00
9	BAETIDAE	0.0576	0.0741	0.1544	0.1845	1.20
10	GYRINIDAE	0.1347	0.1368	0.1721	0.1740	1.01
11	PERLODIDAE	0.1220	0.1236	0.1706	0.1728	1.01
12	GAMMARIDAE	0.0307	0.0383	0.1502	0.1714	1.14
13	TIPULIDAE	0.0766	0.0916	0.1345	0.1561	1.16
14	CAENIDAE	0.0800	0.0822	0.1457	0.1484	1.02
15	EPHEMERELLIDAE	0.0945	0.0963	0.1320	0.1337	1.01
16	LIMNEPHILIDAE	0.0495	0.0537	0.1280	0.1336	1.04
17	HYDROPHILIDAE	0.0872	0.0937	0.1262	0.1336	1.06
18	GOERIDAE	0.1018	0.1038	0.1304	0.1323	1.01
19	SIMULIIDAE	0.0528	0.0605	0.1125	0.1210	1.08
20	EPHEMERIDAE	0.0997	0.1005	0.1166	0.1173	1.01
21	HYDROBI/BITHINIIDAE	0.0224	0.0292	0.0995	0.1173	1.18
22	NEMOURIDAE	0.0663	0.0691	0.1093	0.1132	1.04
23	ANCYLIDAE	0.0460	0.0501	0.0944	0.1055	1.12
24	LEPTOPHLEBIIDAE	0.0704	0.0709	0.0987	0.0996	1.01
25	ASELLIDAE	0.0560	0.0803	0.0586	0.0951	1.62
26	SPHAERIIDAE	0.0125	0.0154	0.0719	0.0812	1.13
27	CHLOROPERLIDAE	0.0505	0.0510	0.0791	0.0796	1.01
28	HYDROPTILIDAE	0.0390	0.0405	0.0719	0.0754	1.05
29	POLYCENTROPODIDAE	0.0474	0.0488	0.0725	0.0744	1.03
30	DYTISCIDAE	0.0226	0.0232	0.0688	0.0701	1.02
31	TAENIOPTERYGIDAE	0.0438	0.0452	0.0670	0.0688	1.03
32	PLANARIIDAE	0.0151	0.0166	0.0661	0.0685	1.04
33	GLOSSIPHONIIDAE	0.0118	0.0132	0.0542	0.0566	1.04
34	BRACHYCENTRIDAE	0.0386	0.0388	0.0492	0.0493	1.00
35	ODONTOCERIDAE	0.0466	0.0470	0.0483	0.0485	1.00
36	OLIGOCHAETA	0.0022	0.0250	0.0052	0.0465	8.94
37	PSYCHOMYIIDAE	0.0150	0.0155	0.0440	0.0446	1.01
38	CHIRONOMIDAE	0.0067	0.0205	0.0250	0.0423	1.69
39	PERLIDAE	0.0363	0.0365	0.0410	0.0411	1.00

Table A1 (cont.) **Information values of 76 BMWP taxa for 'Riffle' sites listed in order of their *M'(C, X)* values derived from abundance data.**

Rnk	Taxon	*M(C, X)* Mutual Inform. (Pres)	*M(C, X)* Mutual Inform. (Abun)	*M'(C, X)* Indiff. Mut. Inform. (Pres)	*M'(C, X)* Indiff. Mut. Inform. (Abun)	Improv. Ratio
40	CALOPTERYGIDAE	0.0174	0.0177	0.0352	0.0356	1.01
41	PISCICOLIDAE	0.0214	0.0215	0.0345	0.0347	1.00
42	HALIPLIDAE	0.0097	0.0102	0.0322	0.0328	1.02
43	ERPOBDELLIDAE	0.0116	0.0191	0.0199	0.0301	1.52
44	PLANORBIDAE	0.0049	0.0063	0.0241	0.0256	1.06
45	SIALIDAE	0.0102	0.0105	0.0236	0.0241	1.02
46	LYMNAEIDAE	0.0040	0.0060	0.0192	0.0219	1.14
47	NERITIDAE	0.0147	0.0151	0.0206	0.0212	1.03
48	SCIRTIDAE	0.0089	0.0092	0.0184	0.0188	1.02
49	APHELOCHEIRIDAE	0.0131	0.0140	0.0175	0.0179	1.03
50	DENDROCOELIDAE	0.0042	0.0046	0.0153	0.0157	1.03
51	VALVATIDAE	0.0032	0.0045	0.0120	0.0142	1.18
52	PHILOPOTAMIDAE	0.0091	0.0094	0.0130	0.0133	1.02
53	CORIXIDAE	0.0030	0.0047	0.0101	0.0119	1.19
54	COENAGRIIDAE	0.0029	0.0033	0.0086	0.0090	1.05
55	GERRIDAE	0.0042	0.0045	0.0075	0.0079	1.05
56	UNIONIDAE	0.0040	0.0040	0.0074	0.0074	1.01
57	PHYSIDAE	0.0017	0.0029	0.0032	0.0060	1.88
58	BERAEIDAE	0.0031	0.0031	0.0050	0.0051	1.00
59	MOLANNIDAE	0.0034	0.0034	0.0050	0.0050	1.00
60	CAPNIIDAE	0.0029	0.0036	0.0039	0.0046	1.18
61	ASTACIDAE	0.0022	0.0027	0.0036	0.0041	1.11
62	HYDROMETRIDAE	0.0012	0.0014	0.0029	0.0031	1.06
63	HIRUDINIDAE	0.0030	0.0031	0.0027	0.0031	1.13
64	CORDULEGASTERIDAE	0.0020	0.0020	0.0030	0.0030	1.00
65	PLATYCNEMIDIDAE	0.0010	0.0014	0.0024	0.0028	1.15
66	PHRYGANEIDAE	0.0012	0.0012	0.0027	0.0027	1.00
67	NOTONECTIDAE	0.0009	0.0009	0.0026	0.0026	1.00
68	VIVIPARIDAE	0.0017	0.0020	0.0021	0.0023	1.08
69	DRYOPIDAE	0.0010	0.0010	0.0016	0.0016	1.03
70	COROPHIIDAE	0.0007	0.0011	0.0013	0.0015	1.16
71	SIPHLONURIDAE	0.0004	0.0009	0.0007	0.0010	1.53
72	POTAMANTHIDAE	0.0006	0.0008	0.0009	0.0010	1.14
73	NEPIDAE	0.0004	0.0004	0.0009	0.0009	1.00
74	LIBELLULIDAE	0.0004	0.0004	0.0006	0.0006	1.00
75	AESHNIDAE	0.0005	0.0005	0.0005	0.0005	1.00
76	NAUCORIDAE	0.0001	0.0001	0.0003	0.0003	1.00

Table A2 Information values of 76 BMWP taxa for 'Pool' sites listed in order of their $M'(C, X)$ values derived from abundance data.

		$M(C, X)$		$M'(C, X)$		
		Mutual Inform.		Indiff. Mut. Inform.		Improv.
Rnk	Taxon	(Pres)	(Abun)	(Pres)	(Abun)	Ratio
1	LEPTOCERIDAE	0.1927	0.1989	0.2469	0.2531	1.02
2	CAENIDAE	0.1689	0.1831	0.2127	0.2243	1.05
3	ELMIDAE	0.1398	0.1511	0.1989	0.2107	1.06
4	BAETIDAE	0.0959	0.1067	0.1624	0.1747	1.08
5	HYDROBI/BITHINIIDAE	0.1073	0.1126	0.1645	0.1728	1.05
6	GAMMARIDAE	0.0883	0.1060	0.1487	0.1705	1.15
7	SPHAERIIDAE	0.0659	0.0771	0.1240	0.1501	1.21
8	PLANORBIDAE	0.0809	0.0841	0.1362	0.1418	1.04
9	COENAGRIIDAE	0.0893	0.0975	0.1284	0.1377	1.07
10	SIALIDAE	0.0752	0.0767	0.1300	0.1318	1.01
11	LIMNEPHILIDAE	0.0746	0.0831	0.1167	0.1289	1.10
12	CALOPTERYGIDAE	0.0945	0.0970	0.1256	0.1280	1.02
13	DYTISCIDAE	0.0616	0.0653	0.1134	0.1198	1.06
14	CORIXIDAE	0.0691	0.0748	0.1088	0.1176	1.08
15	VALVATIDAE	0.0730	0.0757	0.1106	0.1146	1.04
16	EPHEMERIDAE	0.0904	0.0922	0.1080	0.1098	1.02
17	HALIPLIDAE	0.0661	0.0691	0.1059	0.1097	1.04
18	PLANARIIDAE	0.0596	0.0610	0.1009	0.1032	1.02
19	HYDROPTILIDAE	0.0646	0.0689	0.0855	0.0892	1.04
20	LYMNAEIDAE	0.0305	0.0404	0.0723	0.0840	1.16
21	ANCYLIDAE	0.0608	0.0634	0.0812	0.0838	1.03
22	GLOSSIPHONIIDAE	0.0337	0.0368	0.0779	0.0836	1.07
23	POLYCENTROPODIDAE	0.0568	0.0595	0.0742	0.0769	1.04
24	PSYCHOMYIIDAE	0.0555	0.0577	0.0752	0.0768	1.02
25	HYDROPSYCHIDAE	0.0422	0.0457	0.0693	0.0721	1.04
26	MOLANNIDAE	0.0629	0.0633	0.0717	0.0720	1.00
27	LEPTOPHLEBIIDAE	0.0519	0.0545	0.0686	0.0707	1.03
28	PISCICOLIDAE	0.0540	0.0552	0.0646	0.0656	1.02
29	UNIONIDAE	0.0492	0.0502	0.0615	0.0622	1.01
30	NOTONECTIDAE	0.0409	0.0429	0.0589	0.0609	1.03
31	ERPOBDELLIDAE	0.0277	0.0379	0.0431	0.0541	1.26
32	ASELLIDAE	0.0174	0.0314	0.0311	0.0520	1.67
33	OLIGOCHAETA	0.0020	0.0337	0.0023	0.0484	21.21
34	HYDROPHILIDAE	0.0214	0.0256	0.0444	0.0483	1.09
35	SERICOSTOMATIDAE	0.0368	0.0381	0.0454	0.0464	1.02
36	CHIRONOMIDAE	0.0103	0.0260	0.0222	0.0463	2.09
37	EPHEMERELLIDAE	0.0354	0.0379	0.0436	0.0456	1.05
38	PHRYGANEIDAE	0.0367	0.0367	0.0455	0.0455	1.00
39	NERITIDAE	0.0368	0.0385	0.0439	0.0453	1.03

Table A2 (cont.) **Information values of 76 BMWP taxa for 'Pool' sites listed in order of their $M'(C, X)$ values derived from abundance data.**

Rnk	Taxon	$M(C, X)$ Mutual Inform. (Pres)	$M(C, X)$ Mutual Inform. (Abun)	$M'(C, X)$ Indiff. Mut. Inform. (Pres)	$M'(C, X)$ Indiff. Mut. Inform. (Abun)	Improv. Ratio
40	PHYSIDAE	0.0247	0.0277	0.0400	0.0439	1.10
41	PLATYCNEMIDIDAE	0.0345	0.0375	0.0417	0.0437	1.05
42	TIPULIDAE	0.0176	0.0201	0.0385	0.0422	1.10
43	GYRINIDAE	0.0255	0.0279	0.0360	0.0382	1.06
44	GOERIDAE	0.0313	0.0327	0.0371	0.0381	1.03
45	GERRIDAE	0.0191	0.0196	0.0283	0.0286	1.01
46	LEPIDOSTOMATIDAE	0.0260	0.0265	0.0279	0.0281	1.01
47	SIMULIIDAE	0.0114	0.0133	0.0261	0.0278	1.07
48	HEPTAGENIIDAE	0.0223	0.0229	0.0270	0.0273	1.01
49	RHYACOPHILIDAE	0.0179	0.0204	0.0247	0.0264	1.07
50	VIVIPARIDAE	0.0173	0.0194	0.0234	0.0250	1.07
51	NEMOURIDAE	0.0127	0.0151	0.0223	0.0248	1.12
52	BERAEIDAE	0.0173	0.0181	0.0201	0.0207	1.03
53	DENDROCOELIDAE	0.0122	0.0125	0.0196	0.0199	1.02
54	COROPHIIDAE	0.0106	0.0151	0.0145	0.0190	1.31
55	NAUCORIDAE	0.0138	0.0151	0.0169	0.0179	1.06
56	BRACHYCENTRIDAE	0.0118	0.0125	0.0151	0.0157	1.04
57	SCIRTIDAE	0.0071	0.0087	0.0138	0.0154	1.12
58	APHELOCHEIRIDAE	0.0132	0.0137	0.0146	0.0149	1.02
59	HYDROMETRIDAE	0.0083	0.0089	0.0134	0.0139	1.04
60	PERLODIDAE	0.0099	0.0117	0.0114	0.0136	1.20
61	LEUCTRIDAE	0.0082	0.0096	0.0115	0.0131	1.14
62	AESHNIDAE	0.0059	0.0059	0.0087	0.0087	1.00
63	NEPIDAE	0.0055	0.0055	0.0083	0.0083	1.00
64	LIBELLULIDAE	0.0044	0.0044	0.0068	0.0068	1.00
65	TAENIOPTERYGIDAE	0.0036	0.0047	0.0042	0.0056	1.31
66	ASTACIDAE	0.0013	0.0026	0.0024	0.0036	1.51
67	CORDULEGASTERIDAE	0.0030	0.0030	0.0034	0.0034	1.00
68	ODONTOCERIDAE	0.0022	0.0031	0.0027	0.0031	1.18
69	CHLOROPERLIDAE	0.0026	0.0026	0.0026	0.0026	1.00
70	HIRUDINIDAE	0.0013	0.0013	0.0018	0.0018	1.00
71	SIPHLONURIDAE	0.0007	0.0007	0.0012	0.0012	1.00
72	PERLIDAE	0.0013	0.0013	0.0011	0.0011	1.00
73	DRYOPIDAE	0.0006	0.0006	0.0010	0.0010	1.00
74	CAPNIIDAE	0.0009	0.0009	0.0009	0.0009	1.00
75	POTAMANTHIDAE	0.0007	0.0007	0.0006	0.0006	1.00
76	PHILOPOTAMIDAE	0.0000	0.0000	0.0000	0.0000	1.00

Table A3 Distribution by region of 'Riffle' $M'(C, X)$ values based on abundance data for 76 BMWP taxa. The taxa are listed in order of their average $M'(C, X)$ value.

		Region (as defined in Table 1.2)													
		1	2	3	4	5	6	7	8	9	10				
Rank	Taxon	Ang	Nrthm	NWest	Mid	Sthn	SWest	Thms	Welsh	Wssx	York	MAX	MIN	ST DEV	AVG
1	ELMIDAE	0.411	0.469	0.403	0.435	0.342	0.424	0.343	0.452	0.509	0.481	0.509	0.342	0.055	0.427
2	HYDROPSYCHIDAE	0.352	0.363	0.354	0.342	0.173	0.407	0.369	0.348	0.340	0.390	0.407	0.173	0.064	0.344
3	HEPTAGENIIDAE		0.441	0.397	0.251	0.159	0.397	0.070	0.276	0.185	0.370	0.441	0.070	0.127	0.283
4	GAMMARIDAE	0.513	0.160	0.145	0.151	0.269	0.276	0.463	0.212	0.196	0.237	0.513	0.145	0.128	0.262
5	LEPTOCERIDAE	0.388	0.205	0.109	0.230	0.293	0.186	0.347	0.116	0.316	0.221	0.388	0.109	0.094	0.241
6	RHYACOPHILIDAE	0.092	0.333	0.308	0.213	0.139	0.201	0.188	0.279	0.271	0.323	0.333	0.092	0.081	0.235
7	BAETIDAE	0.235	0.170	0.206	0.282	0.202	0.281	0.321	0.153	0.171	0.257	0.321	0.153	0.056	0.228
8	SERICOSTOMATIDAE	0.069	0.300	0.204	0.208	0.197	0.405	0.134	0.252	0.224	0.258	0.405	0.069	0.091	0.225
9	CAENIDAE	0.430	0.188	0.147	0.142	0.207	0.127	0.213	0.136	0.205	0.144	0.430	0.127	0.089	0.194
10	SIMULIIDAE	0.234	0.207	0.119	0.117	0.133	0.347	0.217	0.255	0.185	0.124	0.347	0.117	0.074	0.194
11	HYDROBI/BITHINIIDAE	0.280	0.065	0.127	0.155	0.119	0.219	0.294	0.158	0.365	0.155	0.365	0.065	0.093	0.194
12	ASELLIDAE	0.337	0.134	0.192	0.228	0.100	0.107	0.253	0.142	0.254	0.174	0.337	0.100	0.076	0.192
13	LEUCTRIDAE	0.039	0.262	0.268	0.131	0.115	0.329	0.058	0.226	0.141	0.220	0.329	0.039	0.096	0.179
14	ANCYLIDAE	0.226	0.204	0.114	0.214	0.110	0.274	0.167	0.066	0.253	0.110	0.274	0.066	0.071	0.174
15	TIPULIDAE	0.175	0.240	0.124	0.148	0.090	0.253	0.145	0.143	0.147	0.265	0.265	0.090	0.059	0.173
16	OLIGOCHAETA	0.229	0.270	0.019	0.102	0.081	0.273	0.131	0.187	0.308	0.084	0.308	0.019	0.099	0.168
17	LIMNEPHILIDAE	0.117	0.171	0.148	0.148	0.207	0.119	0.258	0.124	0.215	0.168	0.258	0.117	0.047	0.167
18	GLOSSIPHONIIDAE	0.394	0.125	0.066	0.049	0.139	0.131	0.280	0.071	0.314	0.084	0.394	0.049	0.120	0.165
19	LEPIDOSTOMATIDAE	0.041	0.210	0.135	0.109	0.194	0.345	0.060	0.187	0.142	0.216	0.345	0.041	0.088	0.164
20	SPHAERIIDAE	0.167	0.117	0.094	0.117	0.115	0.118	0.423	0.105	0.178	0.113	0.423	0.094	0.098	0.155
21	PERLODIDAE	0.007	0.303	0.283	0.132	0.075	0.216	0.043	0.193	0.077	0.194	0.303	0.007	0.101	0.152
22	EPHEMERIDAE	0.174	0.109	0.027	0.182	0.251	0.104	0.232	0.076	0.198	0.149	0.251	0.027	0.071	0.150
23	GYRINIDAE	0.101	0.124	0.103	0.143	0.129	0.255	0.087	0.194	0.204	0.127	0.255	0.087	0.054	0.147
24	CHIRONOMIDAE	0.131	0.074	0.044	0.073	0.086	0.253	0.306	0.155	0.227	0.040	0.306	0.040	0.094	0.139
25	EPHEMERELLIDAE	0.109	0.077	0.059	0.115	0.154	0.203	0.156	0.135	0.216	0.089	0.216	0.059	0.052	0.131
26	GOERIDAE	0.142	0.035	0.096	0.092	0.113	0.199	0.203	0.111	0.185	0.119	0.203	0.035	0.054	0.129

Table A2 (cont.) **Information values of 76 BMWP taxa for 'Pool' sites listed in order of their $M'(C, X)$ values derived from abundance data.**

		$M(C, X)$		$M'(C, X)$		
		Mutual Inform.		Indiff. Mut. Inform.		Improv.
Rnk	Taxon	(Pres)	(Abun)	(Pres)	(Abun)	Ratio
40	PHYSIDAE	0.0247	0.0277	0.0400	0.0439	1.10
41	PLATYCNEMIDIDAE	0.0345	0.0375	0.0417	0.0437	1.05
42	TIPULIDAE	0.0176	0.0201	0.0385	0.0422	1.10
43	GYRINIDAE	0.0255	0.0279	0.0360	0.0382	1.06
44	GOERIDAE	0.0313	0.0327	0.0371	0.0381	1.03
45	GERRIDAE	0.0191	0.0196	0.0283	0.0286	1.01
46	LEPIDOSTOMATIDAE	0.0260	0.0265	0.0279	0.0281	1.01
47	SIMULIIDAE	0.0114	0.0133	0.0261	0.0278	1.07
48	HEPTAGENIIDAE	0.0223	0.0229	0.0270	0.0273	1.01
49	RHYACOPHILIDAE	0.0179	0.0204	0.0247	0.0264	1.07
50	VIVIPARIDAE	0.0173	0.0194	0.0234	0.0250	1.07
51	NEMOURIDAE	0.0127	0.0151	0.0223	0.0248	1.12
52	BERAEIDAE	0.0173	0.0181	0.0201	0.0207	1.03
53	DENDROCOELIDAE	0.0122	0.0125	0.0196	0.0199	1.02
54	COROPHIIDAE	0.0106	0.0151	0.0145	0.0190	1.31
55	NAUCORIDAE	0.0138	0.0151	0.0169	0.0179	1.06
56	BRACHYCENTRIDAE	0.0118	0.0125	0.0151	0.0157	1.04
57	SCIRTIDAE	0.0071	0.0087	0.0138	0.0154	1.12
58	APHELOCHEIRIDAE	0.0132	0.0137	0.0146	0.0149	1.02
59	HYDROMETRIDAE	0.0083	0.0089	0.0134	0.0139	1.04
60	PERLODIDAE	0.0099	0.0117	0.0114	0.0136	1.20
61	LEUCTRIDAE	0.0082	0.0096	0.0115	0.0131	1.14
62	AESHNIDAE	0.0059	0.0059	0.0087	0.0087	1.00
63	NEPIDAE	0.0055	0.0055	0.0083	0.0083	1.00
64	LIBELLULIDAE	0.0044	0.0044	0.0068	0.0068	1.00
65	TAENIOPTERYGIDAE	0.0036	0.0047	0.0042	0.0056	1.31
66	ASTACIDAE	0.0013	0.0026	0.0024	0.0036	1.51
67	CORDULEGASTERIDAE	0.0030	0.0030	0.0034	0.0034	1.00
68	ODONTOCERIDAE	0.0022	0.0031	0.0027	0.0031	1.18
69	CHLOROPERLIDAE	0.0026	0.0026	0.0026	0.0026	1.00
70	HIRUDINIDAE	0.0013	0.0013	0.0018	0.0018	1.00
71	SIPHLONURIDAE	0.0007	0.0007	0.0012	0.0012	1.00
72	PERLIDAE	0.0013	0.0013	0.0011	0.0011	1.00
73	DRYOPIDAE	0.0006	0.0006	0.0010	0.0010	1.00
74	CAPNIIDAE	0.0009	0.0009	0.0009	0.0009	1.00
75	POTAMANTHIDAE	0.0007	0.0007	0.0006	0.0006	1.00
76	PHILOPOTAMIDAE	0.0000	0.0000	0.0000	0.0000	1.00

Table A3 Distribution by region of 'Riffle' $M'(C, X)$ values based on abundance data for 76 BMWP taxa. The taxa are listed in order of their average $M'(C, X)$ value.

		Region (as defined in Table 1.2)													
		1	2	3	4	5	6	7	8	9	10				
Rank	Taxon	Ang	Nrthm	NWest	Mid	Sthn	SWest	Thms	Welsh	Wssx	York	MAX	MIN	ST DEV	AVG
1	ELMIDAE	0.411	0.469	0.403	0.435	0.342	0.424	0.343	0.452	0.509	0.481	0.509	0.342	0.055	0.427
2	HYDROPSYCHIDAE	0.352	0.363	0.354	0.342	0.173	0.407	0.369	0.348	0.340	0.390	0.407	0.173	0.064	0.344
3	HEPTAGENIIDAE		0.441	0.397	0.251	0.159	0.397	0.070	0.276	0.185	0.370	0.441	0.070	0.127	0.283
4	GAMMARIDAE	0.513	0.160	0.145	0.151	0.269	0.276	0.463	0.212	0.196	0.237	0.513	0.145	0.128	0.262
5	LEPTOCERIDAE	0.388	0.205	0.109	0.230	0.293	0.186	0.347	0.116	0.316	0.221	0.388	0.109	0.094	0.241
6	RHYACOPHILIDAE	0.092	0.333	0.308	0.213	0.139	0.201	0.188	0.279	0.271	0.323	0.333	0.092	0.081	0.235
7	BAETIDAE	0.235	0.170	0.206	0.282	0.202	0.281	0.321	0.153	0.171	0.257	0.321	0.153	0.056	0.228
8	SERICOSTOMATIDAE	0.069	0.300	0.204	0.208	0.197	0.405	0.134	0.252	0.224	0.258	0.405	0.069	0.091	0.225
9	CAENIDAE	0.430	0.188	0.147	0.142	0.207	0.127	0.213	0.136	0.205	0.144	0.430	0.127	0.089	0.194
10	SIMULIIDAE	0.234	0.207	0.119	0.117	0.133	0.347	0.217	0.255	0.185	0.124	0.347	0.117	0.074	0.194
11	HYDROBI/BITHINIIDAE	0.280	0.065	0.127	0.155	0.119	0.219	0.294	0.158	0.365	0.155	0.365	0.065	0.093	0.194
12	ASELLIDAE	0.337	0.134	0.192	0.228	0.100	0.107	0.253	0.142	0.254	0.174	0.337	0.100	0.076	0.192
13	LEUCTRIDAE	0.039	0.262	0.268	0.131	0.115	0.329	0.058	0.226	0.141	0.220	0.329	0.039	0.096	0.179
14	ANCYLIDAE	0.226	0.204	0.114	0.214	0.110	0.274	0.167	0.066	0.253	0.110	0.274	0.066	0.071	0.174
15	TIPULIDAE	0.175	0.240	0.124	0.148	0.090	0.253	0.145	0.143	0.147	0.265	0.265	0.090	0.059	0.173
16	OLIGOCHAETA	0.229	0.270	0.019	0.102	0.081	0.273	0.131	0.187	0.308	0.084	0.308	0.019	0.099	0.168
17	LIMNEPHILIDAE	0.117	0.171	0.148	0.148	0.207	0.119	0.258	0.124	0.215	0.168	0.258	0.117	0.047	0.167
18	GLOSSIPHONIIDAE	0.394	0.125	0.066	0.049	0.139	0.131	0.280	0.071	0.314	0.084	0.394	0.049	0.120	0.165
19	LEPIDOSTOMATIDAE	0.041	0.210	0.135	0.109	0.194	0.345	0.060	0.187	0.142	0.216	0.345	0.041	0.088	0.164
20	SPHAERIIDAE	0.167	0.117	0.094	0.117	0.115	0.118	0.423	0.105	0.178	0.113	0.423	0.094	0.098	0.155
21	PERLODIDAE	0.007	0.303	0.283	0.132	0.075	0.216	0.043	0.193	0.077	0.194	0.303	0.007	0.101	0.152
22	EPHEMERIDAE	0.174	0.109	0.027	0.182	0.251	0.104	0.232	0.076	0.198	0.149	0.251	0.027	0.071	0.150
23	GYRINIDAE	0.101	0.124	0.103	0.143	0.129	0.255	0.087	0.194	0.204	0.127	0.255	0.087	0.054	0.147
24	CHIRONOMIDAE	0.131	0.074	0.044	0.073	0.086	0.253	0.306	0.155	0.227	0.040	0.306	0.040	0.094	0.139
25	EPHEMERELLIDAE	0.109	0.077	0.059	0.115	0.154	0.203	0.156	0.135	0.216	0.089	0.216	0.059	0.052	0.131
26	GOERIDAE	0.142	0.035	0.096	0.092	0.113	0.199	0.203	0.111	0.185	0.119	0.203	0.035	0.054	0.129

Table A3 cont. **Distribution by region of 'Riffle' $M'(C, X)$ values based on abundance data for 76 BMWP taxa. The taxa are listed in order of their average $M'(C, X)$ value.**

		Region (as defined in Table 1.2)													
		1	2	3	4	5	6	7	8	9	10				
Rank	Taxon	Ang	Nrthm	NWest	Mid	Sthn	SWest	Thms	Welsh	Wssx	York	MAX	MIN	ST DEV	AVG
27	PLANARIIDAE	0.113	0.163	0.060	0.050	0.171	0.094	0.129	0.163	0.245	0.056	0.245	0.050	0.062	0.124
28	NEMOURIDAE	0.033	0.278	0.221	0.093	0.070	0.086	0.031	0.162	0.064	0.152	0.278	0.031	0.082	0.119
29	LEPTOPHLEBIIDAE	0.089	0.136	0.090	0.116	0.195	0.136	0.089	0.084	0.122	0.123	0.195	0.084	0.034	0.118
30	ERPOBDELLIDAE	0.182	0.051	0.042	0.048	0.061	0.111	0.374	0.102	0.134	0.064	0.374	0.042	0.101	0.117
31	HYDROPHILIDAE	0.047	0.130	0.176	0.068	0.082	0.261	0.026	0.195	0.055	0.090	0.261	0.026	0.076	0.113
32	HYDROPTILIDAE	0.164	0.107	0.045	0.100	0.141	0.078	0.182	0.088	0.115	0.092	0.182	0.045	0.041	0.111
33	DYTISCIDAE	0.188	0.136	0.095	0.074	0.017	0.041	0.147	0.137	0.205	0.067	0.205	0.017	0.062	0.111
34	POLYCENTROPODIDAE	0.099	0.191	0.099	0.046	0.058	0.099	0.131	0.076	0.106	0.128	0.191	0.046	0.041	0.103
35	PLANORBIDAE	0.212	0.039	0.014	0.060	0.042	0.036	0.193	0.029	0.189	0.026	0.212	0.014	0.080	0.084
36	PSYCHOMYIIDAE	0.102	0.076	0.034	0.076	0.146	0.031	0.106	0.040	0.073	0.064	0.146	0.031	0.036	0.075
37	CHLOROPERLIDAE		0.127	0.141	0.050	0.031	0.114	0.009	0.104	0.019	0.060	0.141	0.009	0.050	0.073
38	LYMNAEIDAE	0.145	0.073	0.027	0.033	0.080	0.095	0.134	0.055	0.054	0.021	0.145	0.021	0.043	0.071
39	CALOPTERYGIDAE	0.167	0.004	0.001	0.032	0.094	0.056	0.175	0.014	0.099	0.003	0.175	0.001	0.066	0.065
40	SIALIDAE	0.119	0.052	0.016	0.043	0.056	0.024	0.178	0.022	0.093	0.023	0.178	0.016	0.053	0.063
41	HALIPLIDAE	0.137	0.068	0.022	0.033	0.040	0.021	0.094	0.036	0.114	0.045	0.137	0.021	0.041	0.061
42	PHYSIDAE	0.208	0.045	0.020	0.044	0.028	0.045	0.079	0.021	0.075	0.036	0.208	0.020	0.056	0.060
43	TAENIOPTERYGIDAE	0.011	0.116	0.072	0.051	0.005	0.123	0.004	0.084	0.027	0.093	0.123	0.004	0.045	0.059
44	BRACHYCENTRIDAE		0.018	0.110	0.048	0.023	0.069	0.072	0.024	0.081	0.065	0.110	0.018	0.031	0.057
45	PISCICOLIDAE	0.113	0.005	0.003	0.046	0.048	0.044	0.097	0.022	0.113	0.009	0.113	0.003	0.043	0.050
46	VALVATIDAE	0.108	0.028	0.007	0.013	0.038	0.021	0.143	0.006	0.114	0.006	0.143	0.006	0.052	0.048
47	PERLIDAE		0.094	0.099	0.022	0.002	0.032		0.062	0.006	0.052	0.099	0.002	0.037	0.046
48	ODONTOCERIDAE		0.064	0.040	0.012	0.059	0.078	0.036	0.045	0.049	0.030	0.078	0.012	0.020	0.046
49	CORIXIDAE	0.113	0.022	0.010	0.017	0.048	0.026	0.089	0.012	0.070	0.016	0.113	0.010	0.037	0.042
50	NERITIDAE	0.134	0.002	0.003	0.017	0.034	0.007	0.087	0.007	0.086	0.036	0.134	0.002	0.046	0.041
51	COENAGRIIDAE	0.103	0.008	0.001	0.009	0.036	0.039	0.086	0.005	0.031	0.008	0.103	0.001	0.036	0.033
52	DENDROCOELIDAE	0.031	0.017	0.010	0.040	0.053	0.004	0.091	0.013	0.037	0.010	0.091	0.004	0.026	0.030

Table A3 cont. **Distribution by region of 'Riffle' M'(C, X) values based on abundance data for 76 BMWP taxa. The taxa are listed in order of their average M'(C, X) value.**

		Region (as defined in Table 1.2)													
		1	2	3	4	5	6	7	8	9	10				
Rank	Taxon	Ang	Nrthm	NWest	Mid	Sthn	SWest	Thms	Welsh	Wssx	York	MAX	MIN	ST DEV	AVG
53	ASTACIDAE	0.037	0.008		0.008	0.007	0.001	0.005	0.001	0.162	0.019	0.162	0.001	0.052	0.028
54	APHELOCHEIRIDAE	0.041		0.003	0.020	0.035	0.019	0.066	0.019	0.042	0.005	0.066	0.003	0.020	0.028
55	UNIONIDAE	0.055		0.005	0.015	0.013		0.077		0.013	0.012	0.077	0.005	0.028	0.027
56	SCIRTIDAE	0.013	0.038	0.033	0.013	0.034	0.026	0.014	0.020	0.038	0.035	0.038	0.013	0.011	0.026
57	MOLANNIDAE	0.028			0.005	0.010	0.001	0.064		0.015	0.002	0.064	0.001	0.023	0.018
58	VIVIPARIDAE	0.016				0.012		0.033		0.010		0.033	0.010	0.010	0.018
59	COROPHIIDAE				0.001			0.042		0.001		0.042	0.001	0.024	0.015
60	HYDROMETRIDAE	0.064		0.002	0.001	0.005	0.006	0.022	0.003	0.013		0.064	0.001	0.021	0.014
61	CORDULEGASTERIDAE			0.002	0.001	0.006	0.007		0.055	0.008		0.055	0.001	0.021	0.013
62	GERRIDAE	0.022		0.003	0.002	0.011	0.019	0.027	0.014	0.006		0.027	0.002	0.009	0.013
63	PHILOPOTAMIDAE		0.006	0.020	0.009	0.005	0.031		0.019	0.008	0.003	0.031	0.003	0.010	0.012
64	CAPNIIDAE		0.026	0.016	0.002	0.022			0.001	0.001		0.026	0.001	0.011	0.011
65	PLATYCNEMIDIDAE	0.008			0.002	0.017	0.001	0.021	0.001	0.023		0.023	0.001	0.010	0.010
66	BERAEIDAE		0.011	0.004	0.001	0.016	0.008	0.020	0.004	0.012	0.010	0.020	0.001	0.006	0.010
67	HIRUDINIDAE		0.008	0.007	0.006			0.020		0.006		0.020	0.006	0.006	0.009
68	NOTONECTIDAE	0.031		0.001	0.002	0.017	0.001	0.013	0.005	0.011	0.003	0.031	0.001	0.010	0.009
69	PHRYGANEIDAE	0.016	0.009	0.001	0.004	0.005	0.001	0.022		0.017	0.004	0.022	0.001	0.008	0.009
70	NEPIDAE	0.012			0.001	0.002	0.001	0.010		0.003	0.003	0.012	0.001	0.004	0.005
71	POTAMANTHIDAE								0.004			0.004	0.004	N/A	0.004
72	DRYOPIDAE		0.003			0.005		0.004	0.005	0.001	0.003	0.005	0.001	0.002	0.003
73	LIBELLULIDAE	0.005			0.002			0.005	0.002	0.001		0.005	0.001	0.002	0.003
74	AESHNIDAE	0.007				0.002	0.001	0.004		0.001		0.007	0.001	N/A	0.003
75	NAUCORIDAE	0.004			0.001	0.006	0.001					0.006	0.001	N/A	0.003
76	SIPHLONURIDAE		0.002	0.003			0.001		0.002		0.003	0.003	0.001	N/A	0.002

Table A4 Distribution by region of 'Pool' M'(C, X) values based on abundance data for 76 BMWP taxa. The taxa are listed in order of their average M'(C, X) value

		Region (as defined in Table 1.2)										Statistics (excl. Regions 2 and 6)			
		1	*2*	3	4	5	*6*	7	8	9	10				
Rank	Taxon	Ang	*Nrthm*	NWest	Mid	Sthn	*SWest*	Thms	Welsh	Wssx	York	MAX	MIN	ST DEV	AVG
1	GAMMARIDAE	0.316	*0.592*	0.255	0.284	0.411	*0.442*	0.228	0.187	0.410	0.331	0.411	0.187	0.081	0.303
2	LEPTOCERIDAE	0.286	*0.163*	0.135	0.435	0.266	*0.587*	0.306	0.046	0.244	0.245	0.435	0.046	0.116	0.245
3	SPHAERIIDAE	0.362	*0.449*	0.182	0.284	0.327	*0.665*	0.306	0.079	0.119	0.302	0.362	0.079	0.104	0.245
4	HYDROBI/BITHINIIDAE	0.295	*0.540*	0.200	0.386	0.198	*0.719*	0.276	0.146	0.145	0.301	0.386	0.145	0.085	0.243
5	BAETIDAE	0.199	*0.441*	0.294	0.317	0.268	*0.492*	0.184	0.210	0.274	0.182	0.317	0.182	0.053	0.241
6	ELMIDAE	0.130	*0.566*	0.151	0.355	0.233	*0.621*	0.250	0.218	0.197	0.231	0.355	0.130	0.069	0.221
7	ASELLIDAE	0.378	*0.407*	0.053	0.304	0.199	*0.510*	0.211	0.095	0.335	0.165	0.378	0.053	0.115	0.218
8	LIMNEPHILIDAE	0.132	*0.347*	0.168	0.324	0.309	*0.507*	0.119	0.269	0.188	0.181	0.324	0.119	0.079	0.211
9	OLIGOCHAETA	0.207	*0.779*	0.139	0.331	0.223	*0.440*	0.134	0.193	0.087	0.178	0.331	0.087	0.073	0.186
10	LYMNAEIDAE	0.244	*0.461*	0.159	0.283	0.152	*0.294*	0.195	0.144	0.109	0.195	0.283	0.109	0.057	0.185
11	PLANORBIDAE	0.270	*0.243*	0.101	0.142	0.270	*0.266*	0.199	0.182	0.199	0.108	0.270	0.101	0.065	0.184
12	CAENIDAE	0.300	*0.186*	0.028	0.145	0.241	*0.570*	0.185	0.036	0.267	0.254	0.300	0.028	0.104	0.182
13	CHIRONOMIDAE	0.040	*0.438*	0.105	0.348	0.218	*0.358*	0.318	0.121	0.141	0.147	0.348	0.040	0.107	0.180
14	PLANARIIDAE	0.175	*0.164*	0.209	0.135	0.227	*0.513*	0.042	0.171	0.191	0.193	0.227	0.042	0.058	0.168
15	SIALIDAE	0.185	*0.357*	0.086	0.129	0.176	*0.398*	0.190	0.142	0.201	0.213	0.213	0.086	0.043	0.165
16	ERPOBDELLIDAE	0.204	*0.325*	0.079	0.198	0.070	*0.637*	0.216	0.137	0.263	0.078	0.263	0.070	0.074	0.156
17	CORIXIDAE	0.122	*0.235*	0.119	0.138	0.147	*0.240*	0.154	0.147	0.202	0.205	0.205	0.119	0.033	0.154
18	VALVATIDAE	0.281	*0.500*	0.074	0.152	0.130	*0.428*	0.120	0.125	0.189	0.138	0.281	0.074	0.061	0.151
19	HYDROPSYCHIDAE	0.048	*0.492*	0.235	0.327	0.113	*0.305*	0.160	0.127	0.138	0.046	0.327	0.046	0.094	0.149
20	ANCYLIDAE	0.104	*0.458*	0.141	0.337	0.096	*0.089*	0.114	0.159	0.142	0.091	0.337	0.091	0.080	0.148
21	EPHEMERIDAE	0.087	*0.500*	0.165	0.040	0.137	*0.455*	0.212	0.212	0.092	0.141	0.212	0.040	0.061	0.136
22	DYTISCIDAE	0.190	*0.412*	0.148	0.063	0.183	*0.105*	0.072	0.090	0.138	0.184	0.190	0.063	0.052	0.133
23	COENAGRIIDAE	0.255	*0.278*	0.029	0.155	0.201	*0.362*	0.114	0.053	0.158	0.088	0.255	0.029	0.076	0.132
24	GLOSSIPHONIIDAE	0.187	*0.520*	0.103	0.155	0.231	*0.493*	0.067	0.085	0.078	0.142	0.231	0.067	0.058	0.131
25	TIPULIDAE	0.046	*0.429*	0.090	0.288	0.075	*0.186*	0.104	0.246	0.050	0.108	0.288	0.046	0.091	0.126
26	PHYSIDAE	0.166	*0.205*	0.069	0.145	0.092	*0.226*	0.076	0.187	0.094	0.154	0.187	0.069	0.045	0.123

Table A4 cont. **Distribution by region of 'Pool' M'(C, X) values based on abundance data for 76 BMWP taxa. The taxa are listed in order of their average M'(C, X) value**

		Region (as defined in Table 1.2)													
		1	*2*	3	4	5	*6*	7	8	9	10				
Rank	Taxon	Ang	*Nrthm*	NWest	Mid	Sthn	*SWest*	Thms	Welsh	Wssx	York	MAX	MIN	ST DEV	AVG
27	HALIPLIDAE	0.156	*0.272*	0.099	0.086	0.126	*0.144*	0.127	0.114	0.098	0.120	0.156	0.086	0.022	0.116
28	CALOPTERYGIDAE	0.098	*0.186*	0.009	0.139	0.158	*0.534*	0.171	0.094	0.188	0.030	0.188	0.009	0.065	0.111
29	HYDROPTILIDAE	0.055	*0.186*	0.009	0.144	0.108	*0.432*	0.095	0.047	0.180	0.202	0.202	0.009	0.067	0.105
30	PSYCHOMYIIDAE	0.058	*0.186*	0.026	0.124	0.136	*0.187*	0.090	0.054	0.051	0.157	0.157	0.026	0.047	0.087
31	POLYCENTROPODIDAE	0.106	*0.042*	0.066	0.045	0.092	*0.481*	0.100	0.062	0.087	0.109	0.109	0.045	0.023	0.083
32	LEPTOPHLEBIIDAE	0.018	*0.172*	0.063	0.130	0.150	*0.174*	0.076	0.057	0.102	0.072	0.150	0.018	0.042	0.083
33	SIMULIIDAE	0.048	*0.532*	0.078	0.089	0.062	*0.365*	0.089	0.151	0.095	0.050	0.151	0.048	0.033	0.083
34	PISCICOLIDAE	0.093	*0.042*	0.035	0.062	0.119	*0.079*	0.066	0.042	0.101	0.075	0.119	0.035	0.029	0.074
35	MOLANNIDAE	0.143		0.026	0.036	0.047		0.077	0.056	0.077	0.114	0.143	0.026	0.040	0.072
36	HYDROPHILIDAE	0.039	*0.137*	0.091	0.051	0.092	*0.195*	0.036	0.126	0.094	0.038	0.126	0.036	0.034	0.071
37	NOTONECTIDAE	0.107		0.063	0.129	0.052	*0.220*	0.062	0.039	0.067	0.040	0.129	0.039	0.032	0.070
38	NERITIDAE	0.060			0.048	0.005		0.061	0.012	0.101	0.179	0.179	0.005	0.059	0.067
39	UNIONIDAE	0.123		0.012	0.045	0.040		0.094	0.011	0.076	0.112	0.123	0.011	0.044	0.064
40	GYRINIDAE	0.045	*0.042*	0.037	0.071	0.064	*0.030*	0.079	0.068	0.077	0.065	0.079	0.037	0.015	0.063
41	NEMOURIDAE	0.010	*0.242*	0.151	0.015	0.046	*0.253*	0.027	0.085	0.054	0.110	0.151	0.010	0.050	0.062
42	HEPTAGENIIDAE			0.077	0.011	0.036		0.041	0.115	0.020	0.122	0.122	0.011	0.045	0.060
43	GOERIDAE	0.010		0.086	0.036	0.025	*0.079*	0.123	0.106	0.053	0.020	0.123	0.010	0.043	0.057
44	SERICOSTOMATIDAE	0.004		0.008	0.005	0.085	*0.405*	0.070	0.138	0.047	0.093	0.138	0.004	0.049	0.056
45	EPHEMERELLIDAE	0.008	*0.186*		0.007	0.047	*0.175*	0.064	0.086	0.110	0.051	0.110	0.007	0.038	0.053
46	PHRYGANEIDAE	0.080		0.008	0.072	0.025	*0.074*	0.053	0.037	0.042	0.083	0.083	0.008	0.027	0.050
47	VIVIPARIDAE	0.091			0.031	0.012		0.049	0.019	0.020	0.115	0.115	0.012	0.040	0.048
48	COROPHIIDAE	0.084			0.042	0.023		0.039			0.015	0.084	0.015	0.027	0.041
49	RHYACOPHILIDAE	0.008		0.075	0.010	0.022	*0.161*	0.060	0.054	0.051	0.037	0.075	0.008	0.024	0.040
50	LEUCTRIDAE		*0.186*	0.009		0.017	*0.260*	0.019	0.074	0.014	0.105	0.105	0.009	0.040	0.040
51	PERLODIDAE	0.009		0.075	0.003	0.011	*0.220*	0.021	0.133	0.030	0.031	0.133	0.003	0.044	0.039
52	PLATYCNEMIDIDAE	0.005		0.009	0.033	0.060		0.081	0.030	0.076	0.017	0.081	0.005	0.030	0.039

Table A4 cont. **Distribution by region of 'Pool' M'(C, X) values based on abundance data for 76 BMWP taxa. The taxa are listed in order of their average M'(C, X) value**

		Region (as defined in Table 1.2)													
		1	*2*	3	4	5	*6*	7	8	9	10				
Rank	Taxon	Ang	*Nrthm*	NWest	Mid	Sthn	*SWest*	Thms	Welsh	Wssx	York	MAX	MIN	ST DEV	AVG
53	GERRIDAE	0.039		0.007	0.029	0.026	*0.074*	0.069	0.049	0.031	0.030	0.069	0.007	0.018	0.035
54	DENDROCOELIDAE	0.038		0.015	0.023	0.039		0.013	0.051	0.020	0.075	0.075	0.013	0.021	0.034
55	CORDULEGASTERIDAE					0.012	*0.097*		0.056			0.056	0.012	0.031	0.034
56	BRACHYCENTRIDAE				0.019	0.015		0.050	0.025	0.022	0.062	0.062	0.015	0.019	0.032
57	LEPIDOSTOMATIDAE		*0.186*		0.005	0.052		0.023	0.056	0.033	0.017	0.056	0.005	0.020	0.031
58	SCIRTIDAE	0.008	*0.134*	0.015	0.021	0.031	*0.074*	0.033	0.042	0.030	0.053	0.053	0.008	0.014	0.029
59	APHELOCHEIRIDAE	0.004			0.005	0.005	*0.171*	0.045	0.027	0.013	0.079	0.079	0.004	0.028	0.026
60	NAUCORIDAE	0.026				0.031		0.014		0.031		0.031	0.014	0.008	0.025
61	TAENIOPTERYGIDAE		*0.186*	0.009	0.010			0.030	0.024		0.046	0.046	0.009	0.016	0.024
62	HYDROMETRIDAE	0.024		0.006	0.028	0.012		0.051	0.020	0.022		0.051	0.006	0.014	0.023
63	BERAEIDAE	0.004				0.037		0.016	0.027	0.024		0.037	0.004	0.012	0.022
64	ODONTOCERIDAE							0.025	0.012			0.025	0.012	0.010	0.018
65	AESHNIDAE	0.011			0.023	0.024	*0.031*	0.011		0.015		0.024	0.011	0.006	0.017
66	CHLOROPERLIDAE					0.002			0.027		0.015	0.027	0.002	0.012	0.015
67	ASTACIDAE	0.004						0.025				0.025	0.004	0.015	0.014
68	LIBELLULIDAE	0.009			0.010	0.019	*0.171*	0.013	0.006	0.028		0.028	0.006	0.008	0.014
69	NEPIDAE	0.012			0.007	0.008		0.018	0.006	0.018	0.023	0.023	0.006	0.007	0.013
70	SIPHLONURIDAE			0.008		0.003			0.027			0.027	0.003	0.013	0.013
71	PERLIDAE								0.013			0.013	0.013	N/A	0.013
72	HIRUDINIDAE	0.003		0.007	0.005	0.004		0.005				0.007	0.003	0.002	0.005
73	POTAMANTHIDAE							0.005				0.005	0.005	N/A	0.005
74	DRYOPIDAE	0.002		0.006	0.005	0.002		0.002				0.006	0.002	0.002	0.003
75	CAPNIIDAE					0.003						0.003	0.003	N/A	0.003
76	PHILOPOTAMIDAE											0.000	0.000	N/A	N/A

Table A5. The top twelve *Riffle* indicator taxa by region, listed in order of their indifferent mutual information value, $M'(C,X)$.

Anglian		N. East (Northumbria)		North West		Midland		Southern	
GAMMARIDAE	0.513	ELMIDAE	0.469	ELMIDAE	0.403	ELMIDAE	0.435	ELMIDAE	0.342
CAENIDAE	0.430	HEPTAGENIIDAE	0.441	HEPTAGENIIDAE	0.397	HYDROPSYCHIDAE	0.342	LEPTOCERIDAE	0.293
ELMIDAE	0.411	HYDROPSYCHIDAE	0.363	HYDROPSYCHIDAE	0.354	BAETIDAE	0.282	GAMMARIDAE	0.269
GLOSSIPHONIIDAE	0.394	RHYACOPHILIDAE	0.333	RHYACOPHILIDAE	0.308	HEPTAGENIIDAE	0.251	EPHEMERIDAE	0.251
LEPTOCERIDAE	0.388	PERLODIDAE	0.303	PERLODIDAE	0.283	LEPTOCERIDAE	0.230	LIMNEPHILIDAE	0.207
HYDROPSYCHIDAE	0.352	SERICOSTOMATIDAE	0.300	LEUCTRIDAE	0.268	ASELLIDAE	0.228	CAENIDAE	0.207
ASELLIDAE	0.337	NEMOURIDAE	0.278	NEMOURIDAE	0.221	ANCYLIDAE	0.214	BAETIDAE	0.202
HYDROBIIDAE	0.280	OLIGOCHAETA	0.270	BAETIDAE	0.206	RHYACOPHILIDAE	0.213	SERICOSTOMATIDAE	0.197
BAETIDAE	0.235	LEUCTRIDAE	0.262	SERICOSTOMATIDAE	0.204	SERICOSTOMATIDAE	0.208	LEPTOPHLEBIIDAE	0.195
SIMULIIDAE	0.234	TIPULIDAE	0.240	ASELLIDAE	0.192	EPHEMERIDAE	0.182	LEPIDOSTOMATIDAE	0.194
OLIGOCHAETA	0.229	LEPIDOSTOMATIDAE	0.210	HYDROPHILIDAE	0.176	HYDROBIIDAE	0.155	HYDROPSYCHIDAE	0.173
ANCYLIDAE	0.226	SIMULIIDAE	0.207	LIMNEPHILIDAE	0.148	GAMMARIDAE	0.151	PLANARIIDAE	0.171

S. West (Devon & Cornwall)		Thames		Welsh		S. West (N & S Wessex)		N. East (Dales & Riding)	
ELMIDAE	0.424	GAMMARIDAE	0.463	ELMIDAE	0.452	ELMIDAE	0.509	ELMIDAE	0.481
HYDROPSYCHIDAE	0.407	SPHAERIIDAE	0.423	HYDROPSYCHIDAE	0.348	HYDROBIIDAE	0.365	HYDROPSYCHIDAE	0.390
SERICOSTOMATIDAE	0.405	ERPOBDELLIDAE	0.374	RHYACOPHILIDAE	0.279	HYDROPSYCHIDAE	0.340	HEPTAGENIIDAE	0.370
HEPTAGENIIDAE	0.397	HYDROPSYCHIDAE	0.369	HEPTAGENIIDAE	0.276	LEPTOCERIDAE	0.316	RHYACOPHILIDAE	0.323
SIMULIIDAE	0.347	LEPTOCERIDAE	0.347	SIMULIIDAE	0.255	GLOSSIPHONIIDAE	0.314	TIPULIDAE	0.265
LEPIDOSTOMATIDAE	0.345	ELMIDAE	0.343	SERICOSTOMATIDAE	0.252	OLIGOCHAETA	0.308	SERICOSTOMATIDAE	0.258
LEUCTRIDAE	0.329	BAETIDAE	0.321	LEUCTRIDAE	0.226	RHYACOPHILIDAE	0.271	BAETIDAE	0.257
BAETIDAE	0.281	CHIRONOMIDAE	0.306	GAMMARIDAE	0.212	ASELLIDAE	0.254	GAMMARIDAE	0.237
GAMMARIDAE	0.276	HYDROBIIDAE	0.294	HYDROPHILIDAE	0.195	ANCYLIDAE	0.253	LEPTOCERIDAE	0.221
ANCYLIDAE	0.274	GLOSSIPHONIIDAE	0.280	GYRINIDAE	0.194	PLANARIIDAE	0.245	LEUCTRIDAE	0.220
OLIGOCHAETA	0.273	LIMNEPHILIDAE	0.258	PERLODIDAE	0.193	CHIRONOMIDAE	0.227	LEPIDOSTOMATIDAE	0.216
HYDROPHILIDAE	0.261	ASELLIDAE	0.253	OLIGOCHAETA	0.187	SERICOSTOMATIDAE	0.224	PERLODIDAE	0.194

Table A6. The top twelve *Pool* indicator taxa by region, listed in order of their indifferent mutual information value, $M'(C,X)$.

Anglian		**North West**		**Midland**		**Southern**	
ASELLIDAE	0.378	BAETIDAE	0.294	LEPTOCERIDAE	0.435	GAMMARIDAE	0.411
SPHAERIIDAE	0.362	GAMMARIDAE	0.255	HYDROBIIDAE	0.386	SPHAERIIDAE	0.327
GAMMARIDAE	0.316	HYDROPSYCHIDAE	0.235	ELMIDAE	0.355	LIMNEPHILIDAE	0.309
CAENIDAE	0.300	PLANARIIDAE	0.209	CHIRONOMIDAE	0.348	PLANORBIDAE	0.270
HYDROBIIDAE	0.295	HYDROBIIDAE	0.200	ANCYLIDAE	0.337	BAETIDAE	0.268
LEPTOCERIDAE	0.286	SPHAERIIDAE	0.182	OLIGOCHAETA	0.331	LEPTOCERIDAE	0.266
VALVATIDAE	0.281	LIMNEPHILIDAE	0.168	HYDROPSYCHIDAE	0.327	CAENIDAE	0.241
PLANORBIDAE	0.270	EPHEMERIDAE	0.165	LIMNEPHILIDAE	0.324	ELMIDAE	0.233
COENAGRIIDAE	0.255	LYMNAEIDAE	0.159	BAETIDAE	0.317	GLOSSIPHONIIDAE	0.231
LYMNAEIDAE	0.244	NEMOURIDAE	0.151	ASELLIDAE	0.304	PLANARIIDAE	0.227
OLIGOCHAETA	0.207	ELMIDAE	0.151	TIPULIDAE	0.288	OLIGOCHAETA	0.223
ERPOBDELLIDAE	0.204	DYTISCIDAE	0.148	GAMMARIDAE	0.284	CHIRONOMIDAE	0.218

Thames		**Welsh**		**S. West (N & S Wessex)**		**N. East (Dales & Riding)**	
CHIRONOMIDAE	0.318	LIMNEPHILIDAE	0.269	GAMMARIDAE	0.410	GAMMARIDAE	0.331
LEPTOCERIDAE	0.306	TIPULIDAE	0.246	ASELLIDAE	0.335	SPHAERIIDAE	0.302
SPHAERIIDAE	0.306	ELMIDAE	0.218	BAETIDAE	0.274	HYDROBIIDAE	0.301
HYDROBIIDAE	0.276	EPHEMERIDAE	0.212	CAENIDAE	0.267	CAENIDAE	0.254
ELMIDAE	0.250	BAETIDAE	0.210	ERPOBDELLIDAE	0.263	LEPTOCERIDAE	0.245
GAMMARIDAE	0.228	OLIGOCHAETA	0.193	LEPTOCERIDAE	0.244	ELMIDAE	0.231
ERPOBDELLIDAE	0.216	GAMMARIDAE	0.187	CORIXIDAE	0.202	SIALIDAE	0.213
EPHEMERIDAE	0.212	PHYSIDAE	0.187	SIALIDAE	0.201	CORIXIDAE	0.205
ASELLIDAE	0.211	PLANORBIDAE	0.182	PLANORBIDAE	0.199	HYDROPTILIDAE	0.202
PLANORBIDAE	0.199	PLANARIIDAE	0.171	ELMIDAE	0.197	LYMNAEIDAE	0.195
LYMNAEIDAE	0.195	ANCYLIDAE	0.159	PLANARIIDAE	0.191	PLANARIIDAE	0.193
SIALIDAE	0.190	SIMULIIDAE	0.151	VALVATIDAE	0.189	DYTISCIDAE	0.184

Footnote. North East (Northumbria) and South West (Devon and Cornwall) were not included in these lists because they had so few pool sites that their $M'(C,X)$ were unreliable.

Appendix B

Distribution of Site Types 1 to 5

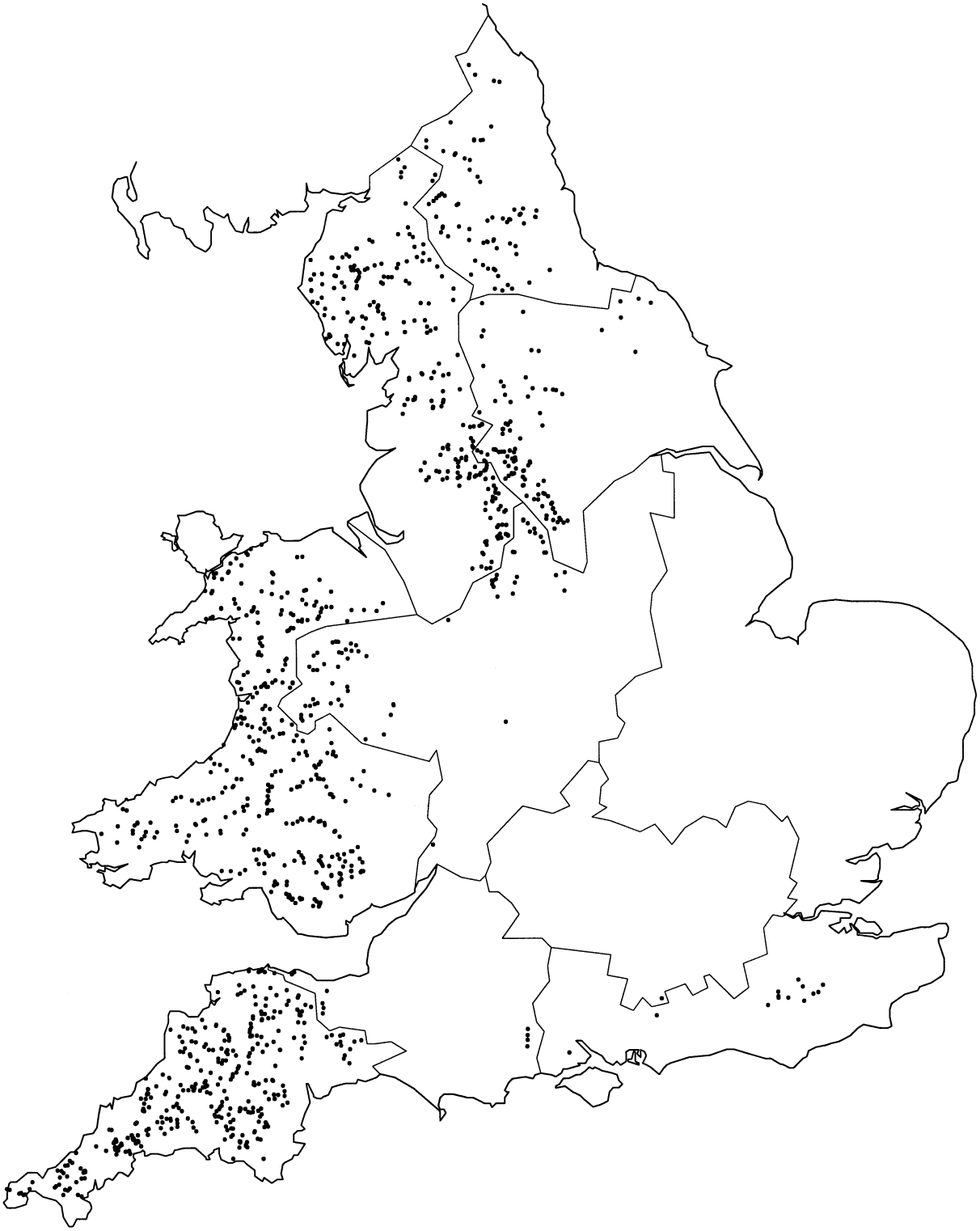


Figure B1. Geographic distribution of site type '1'.

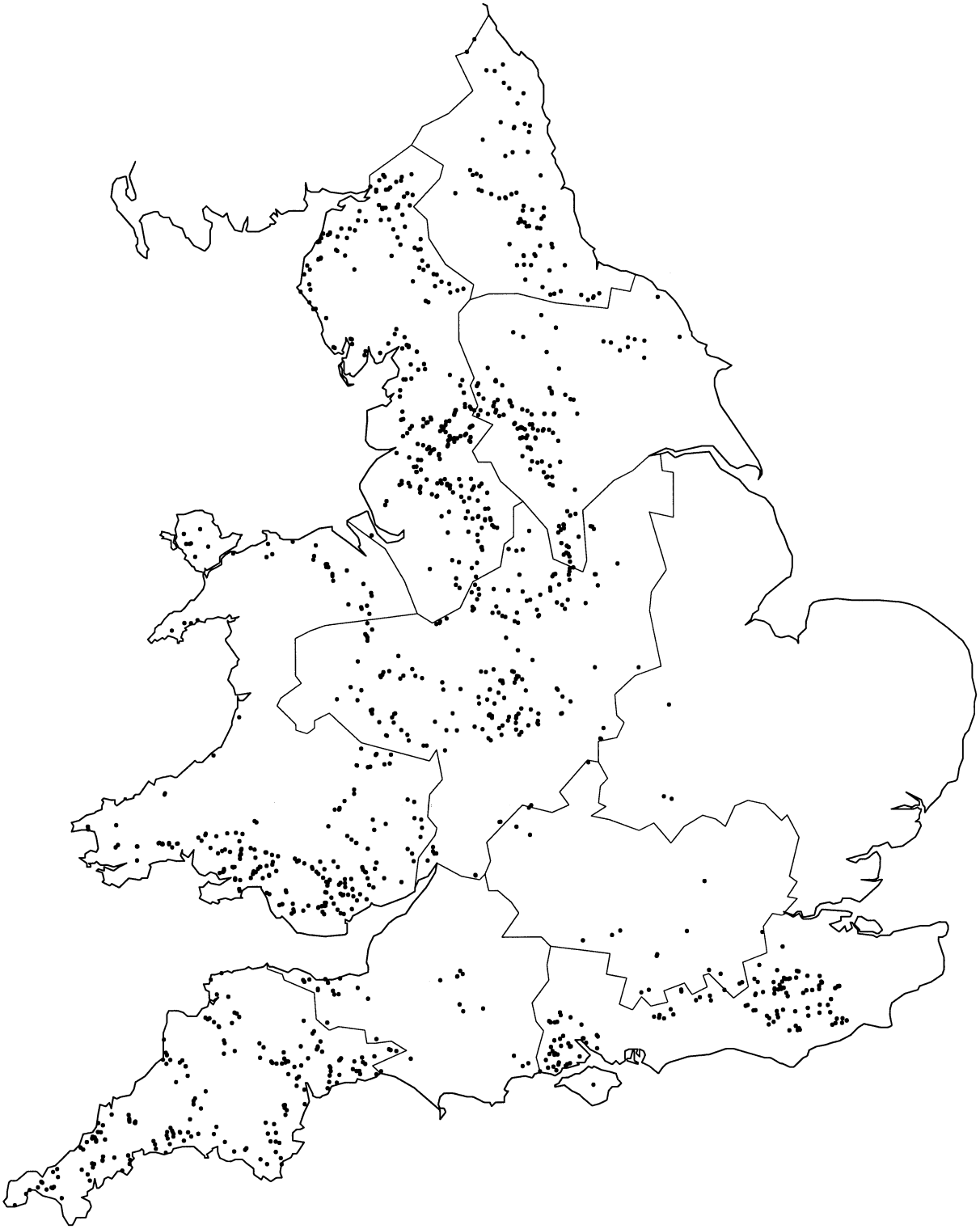


Figure B2. Geographic distribution of site type '2'.

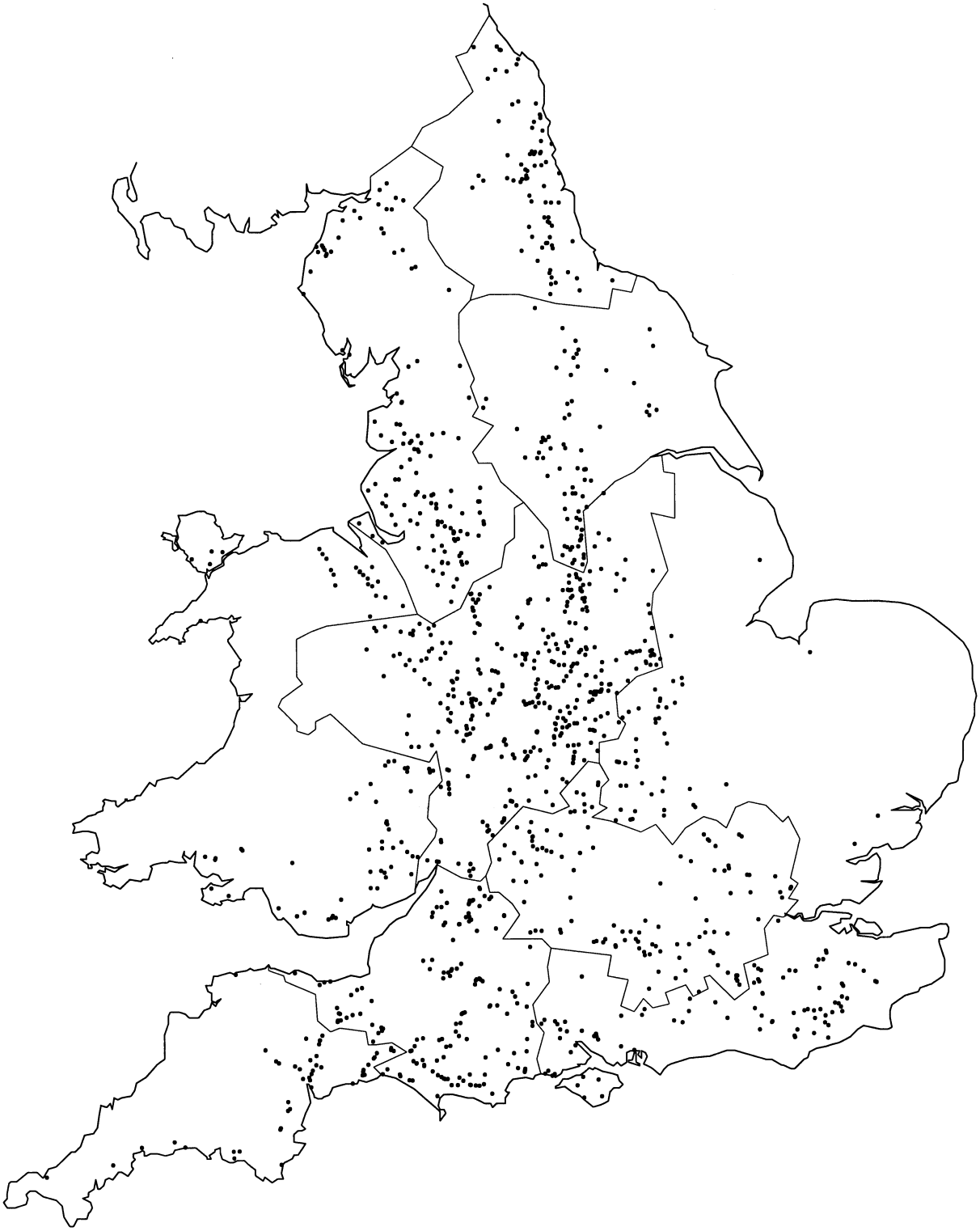


Figure B3. Geographic distribution of site type '3'.

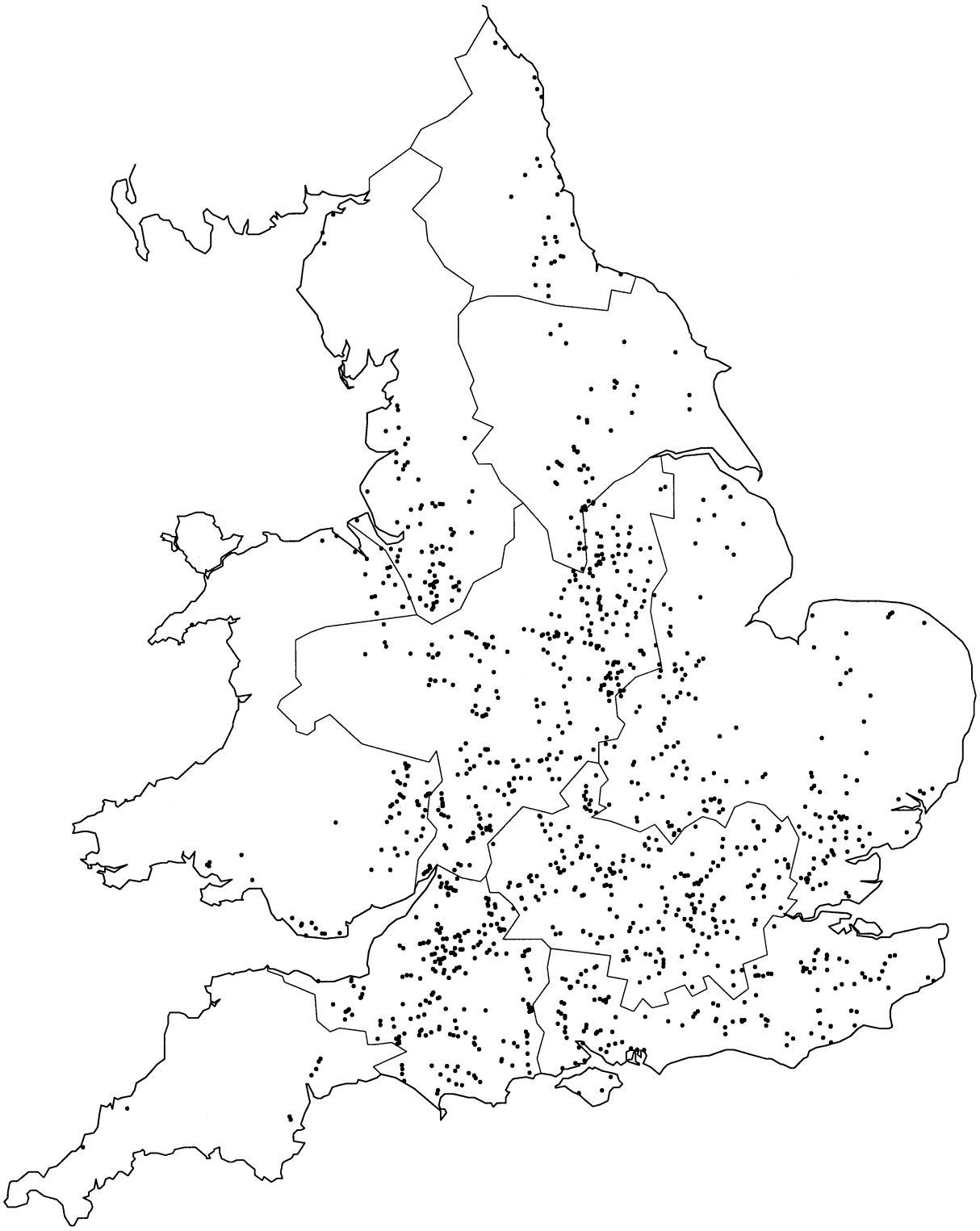


Figure B4. Geographic distribution of site type '4'.

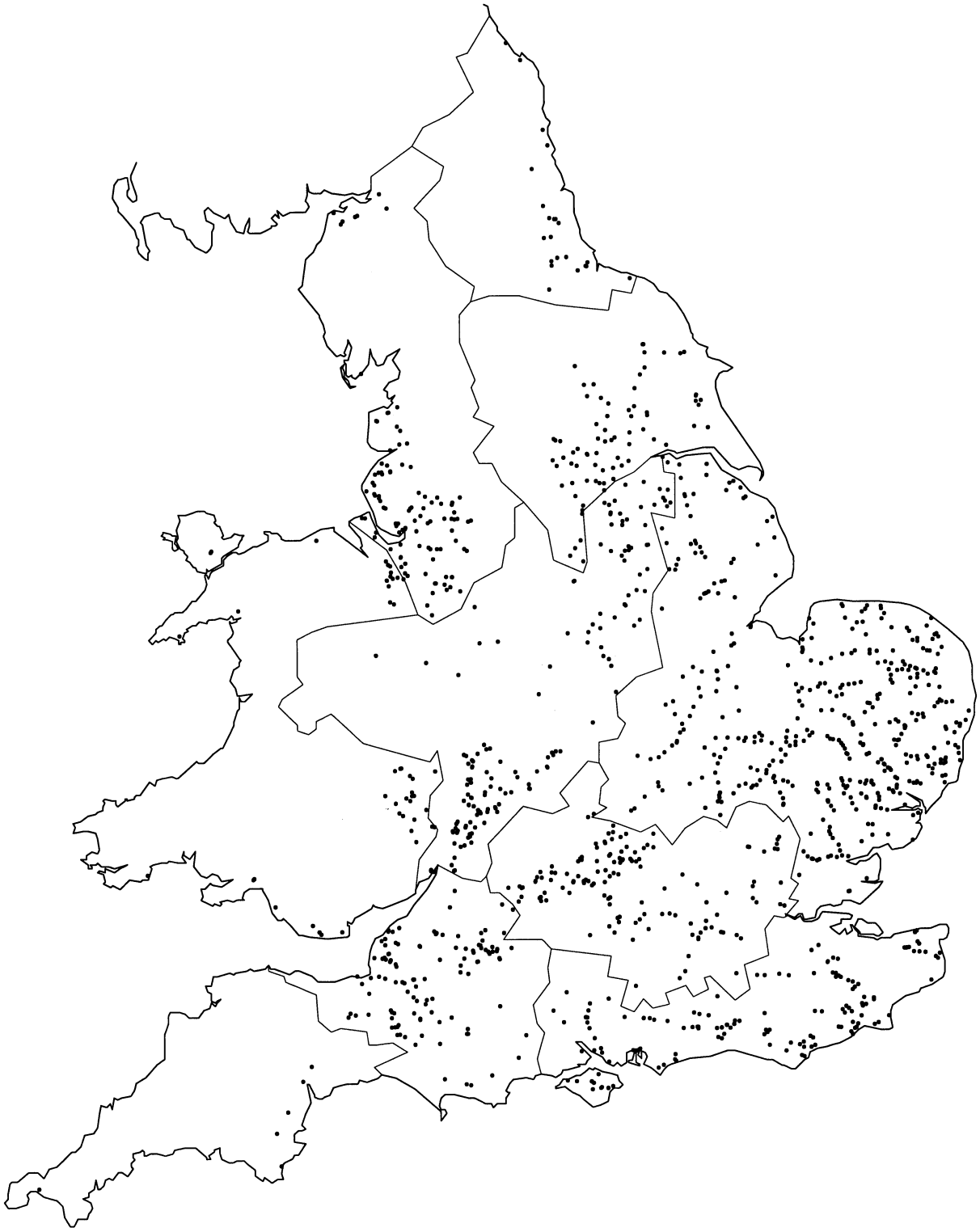


Figure B5. Geographic distribution of site type '5'.

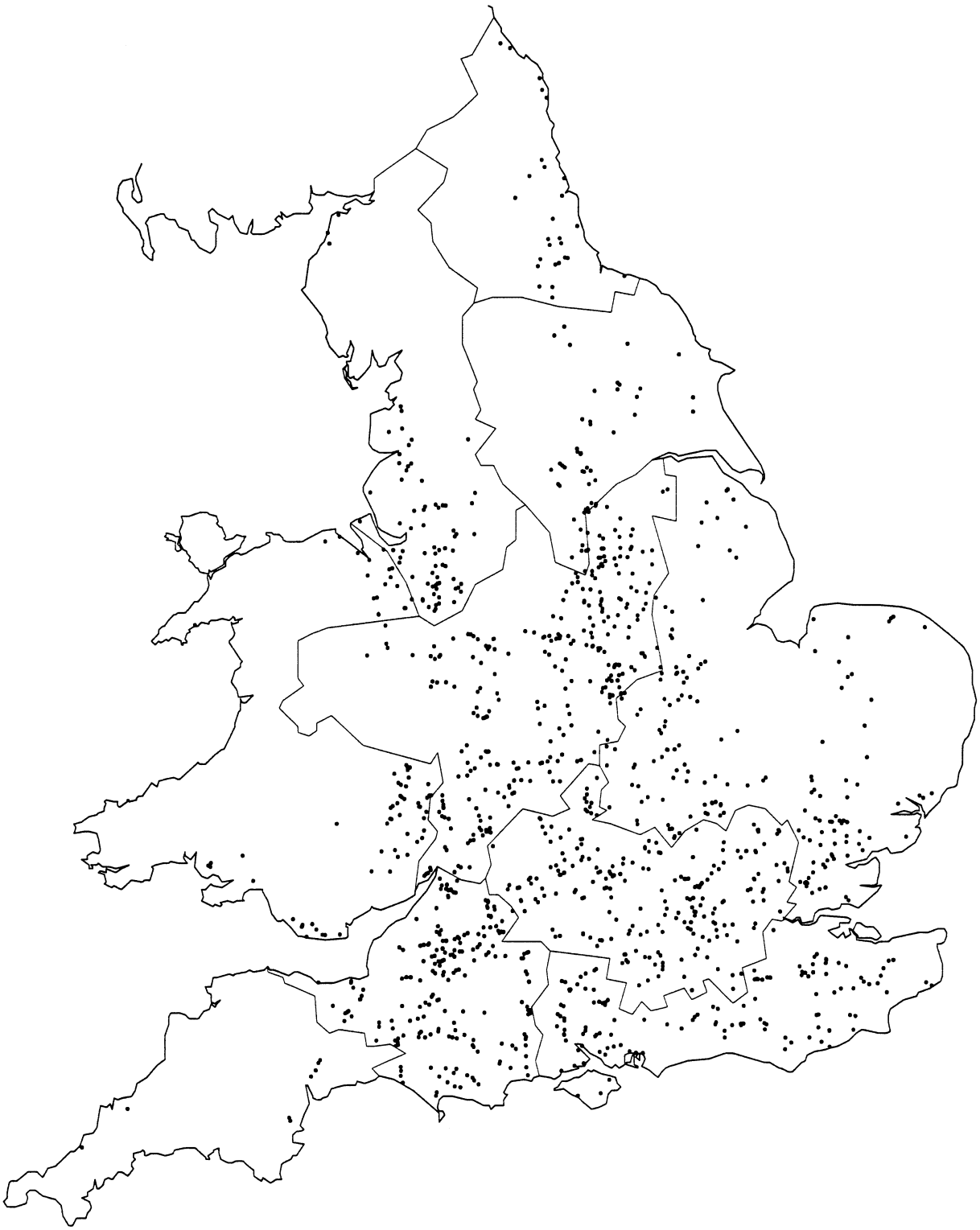


Figure B4. Geographic distribution of site type '4'.

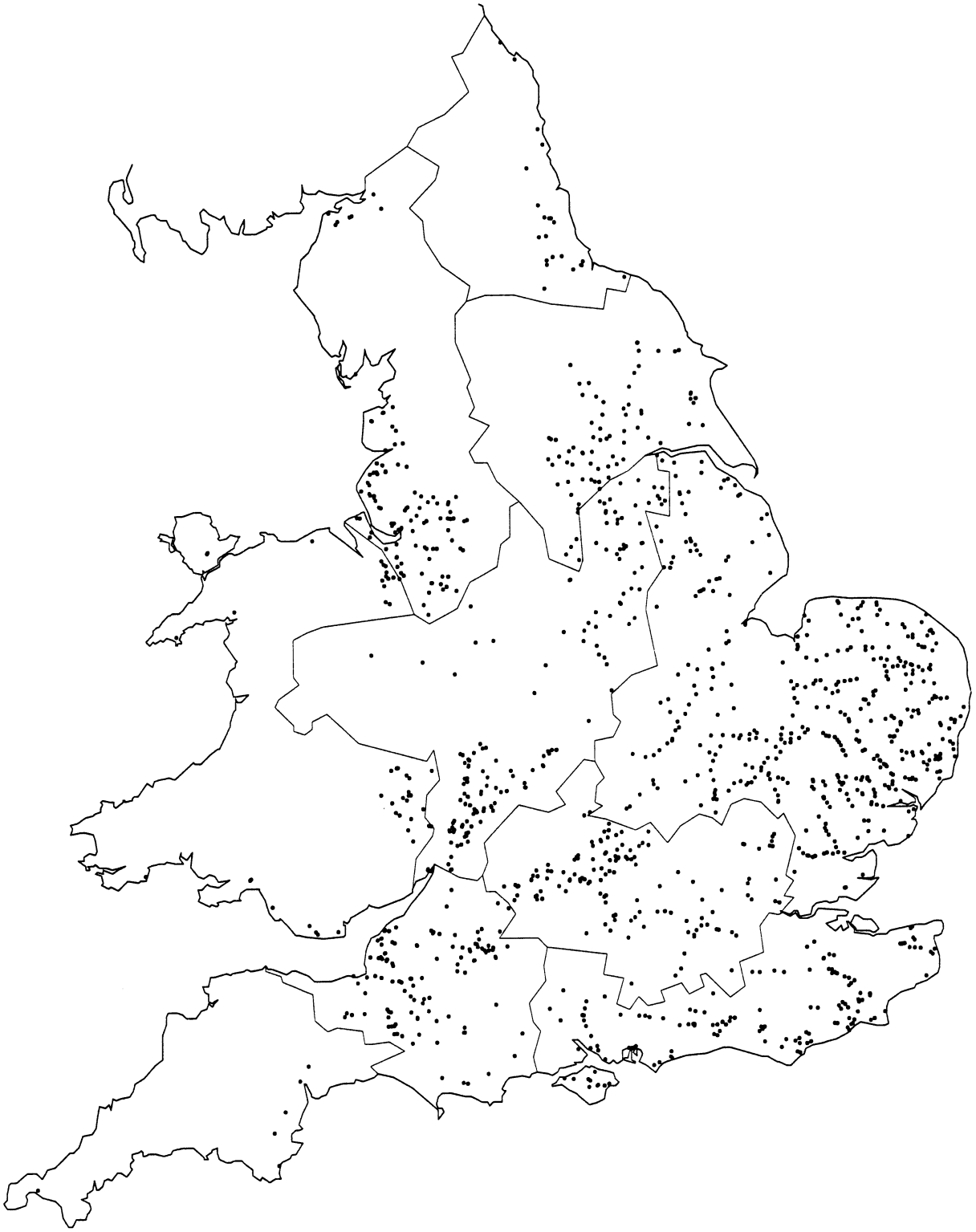


Figure B5. Geographic distribution of site type '5'.

Appendix C

Distribution of EQI(ASPT) and EQI(NFAM) over England and Wales as produced by the Neural Network and RIVPACS Predictors of ASPT and NFAM

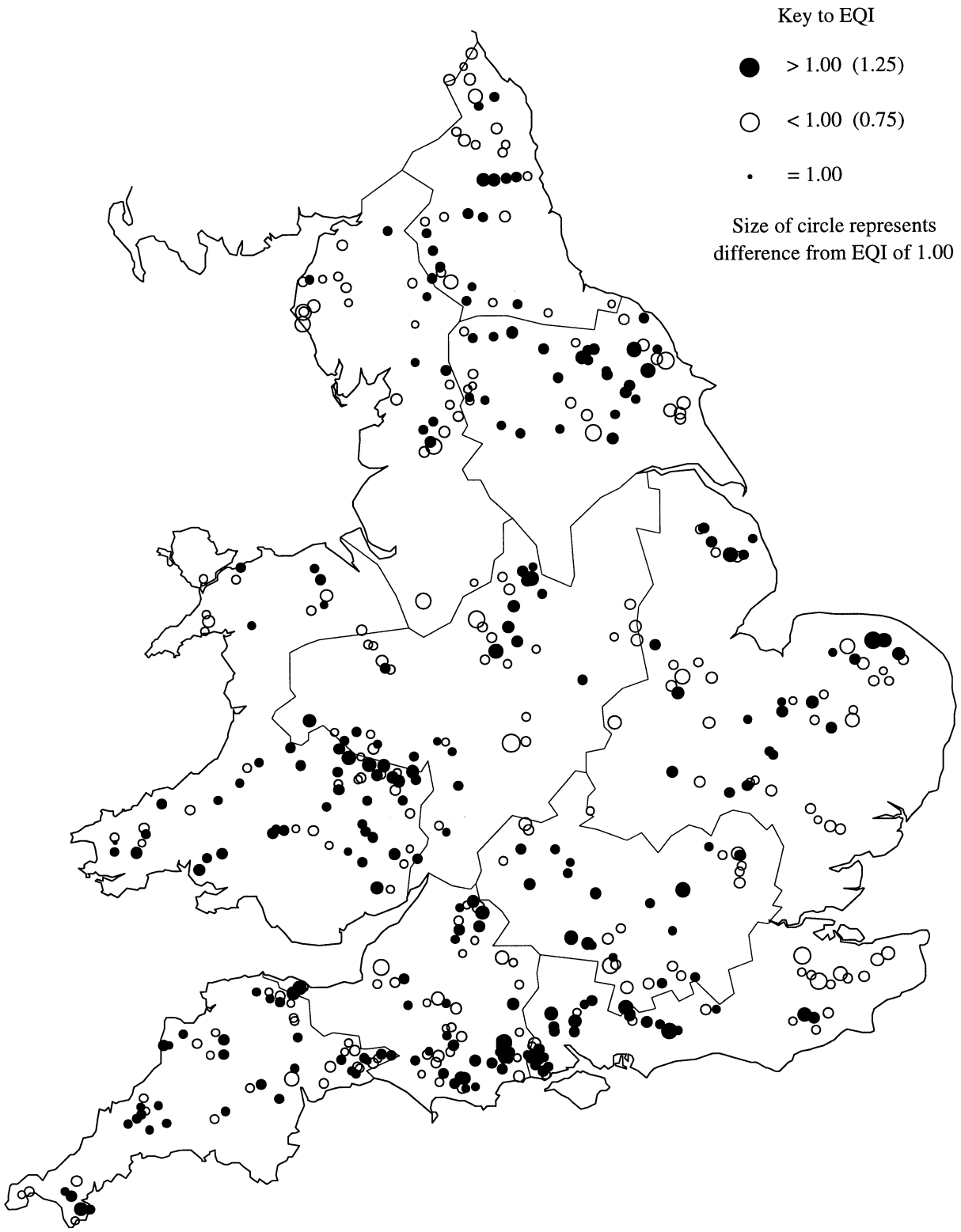


Figure C1. Distribution of EQI(ASPT) for the IFE614 sites based upon ASPT predictions produced by the neural network N5XASPT

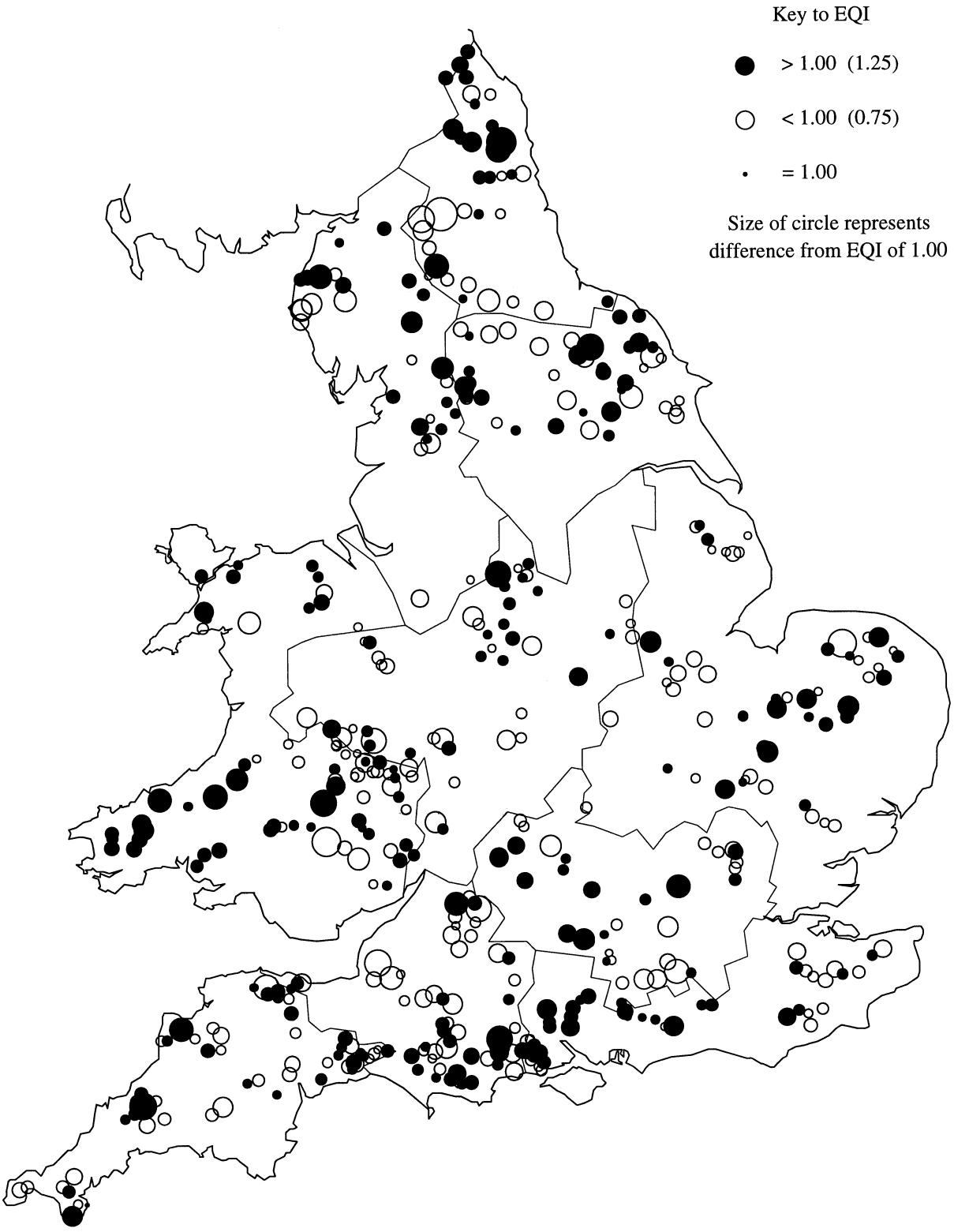


Figure C2. Distribution of EQI(NFAM) for the IFE614 sites based upon NFAM predictions produced by the neural network N5XNFAM

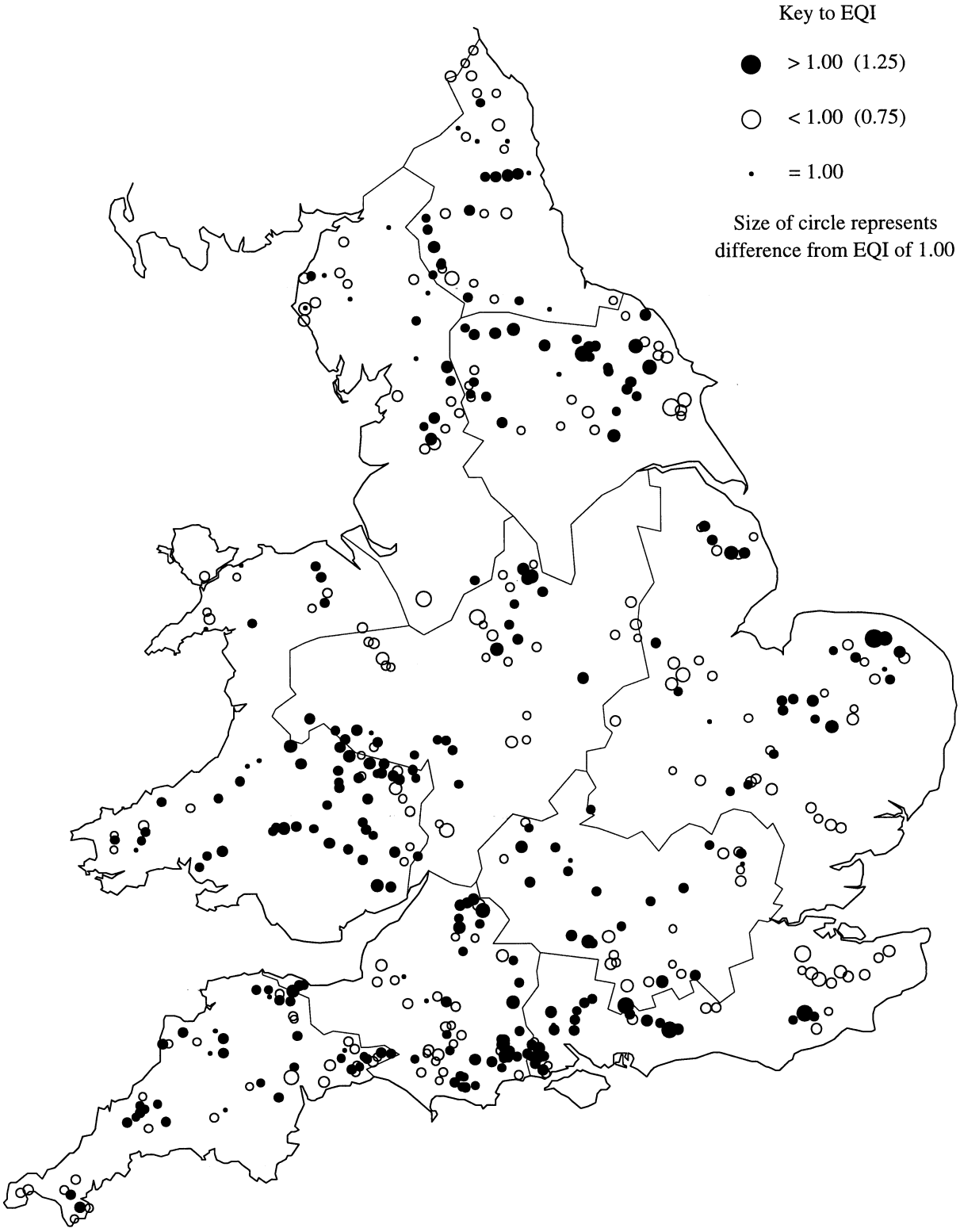


Figure C3. Distribution of EQI(ASPT) for the IFE614 sites based upon RIVPACS predictions

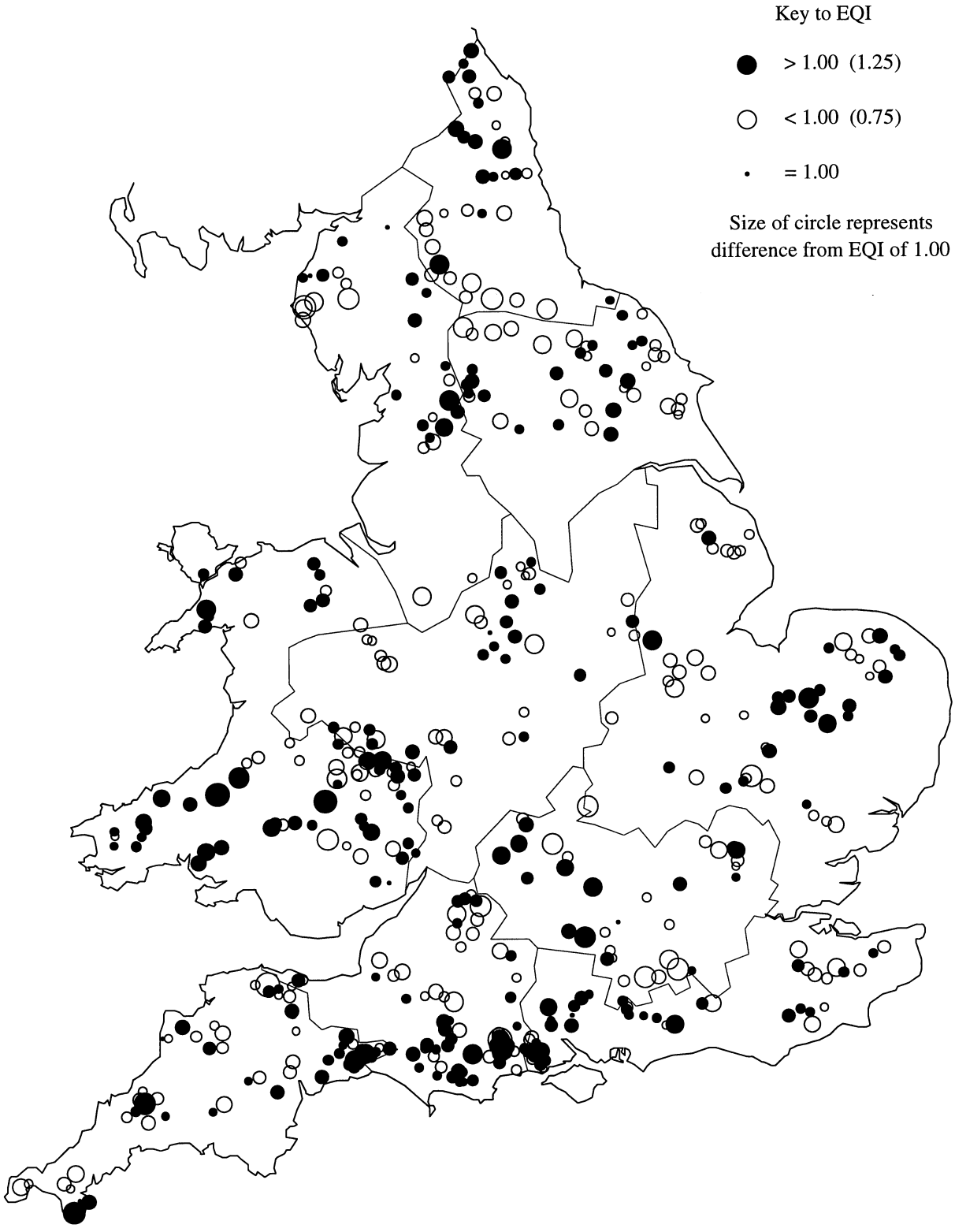
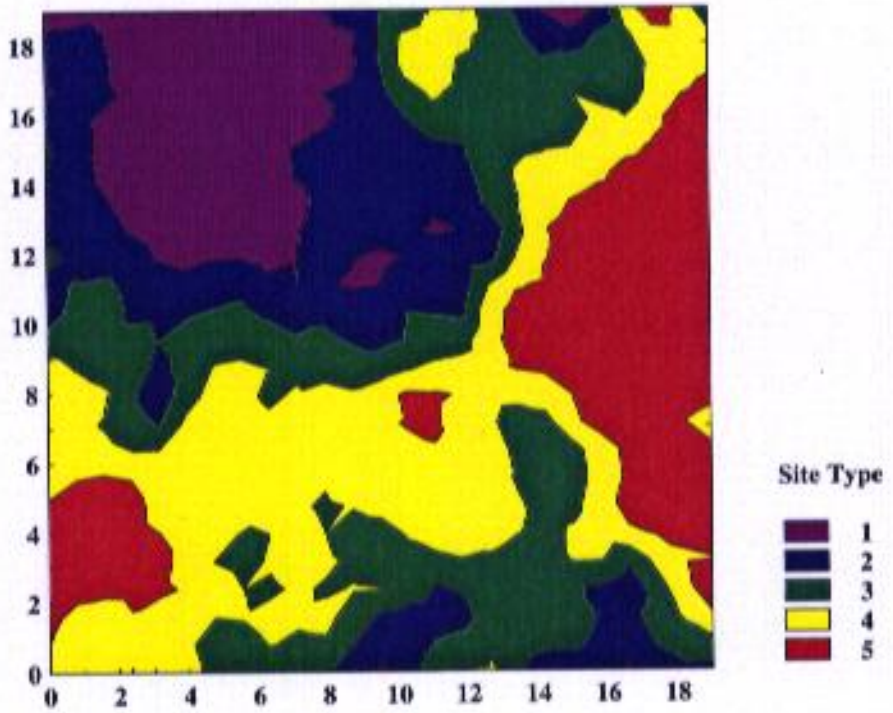


Figure C4. Distribution of EQI(NFAM) for the IFE614 sites based upon RIVPACS predictions

Appendix D

Feature Maps produced by SOM20

Site Type

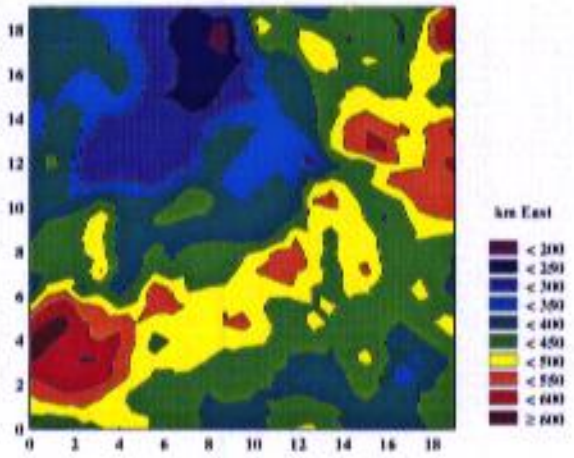


This map of Site Type is typical of the feature maps produced by SOM20. It represents the distribution of the particular attribute, Site Type, across the 20x20 output array of 'classification' bins produced by SOM20.

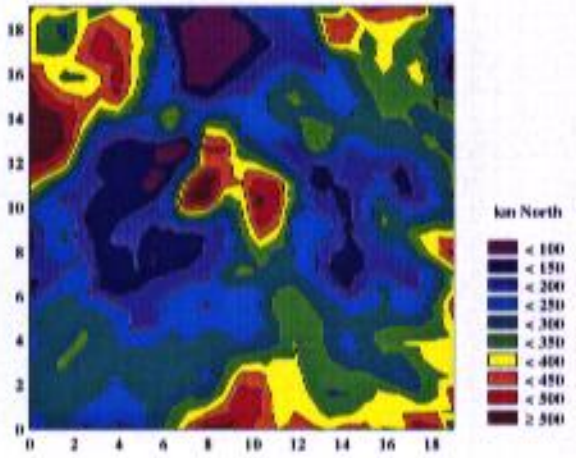
The following pages show feature maps for a further 96 attributes, including 12 physical variables, four chemical variables, ASPT, NFAM, two biological GQAs (RIVPACS and Neural Network) and 76 BMWP families.

A detailed explanation of feature maps and how to interpret them is given in Section 4.4.3.

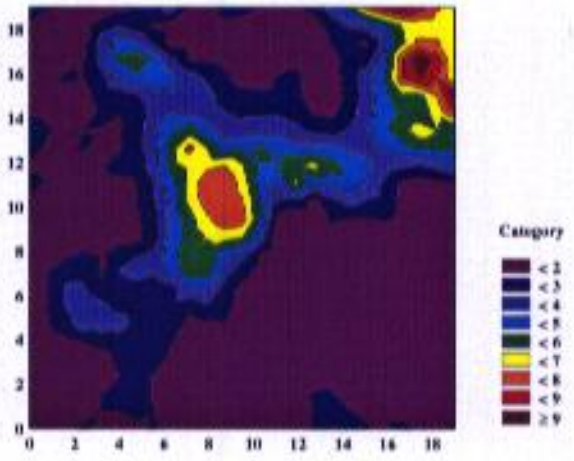
Global X (km East)



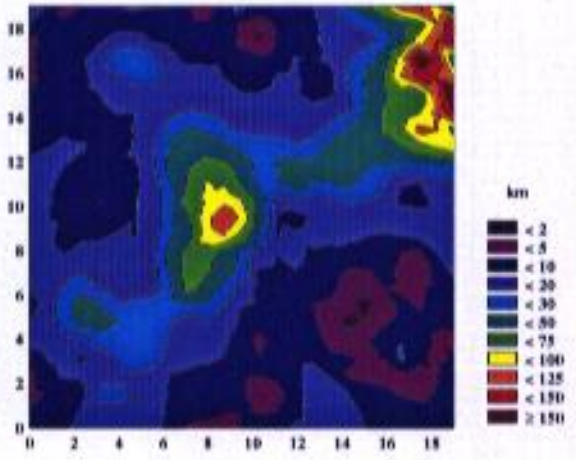
Global Y (km North)



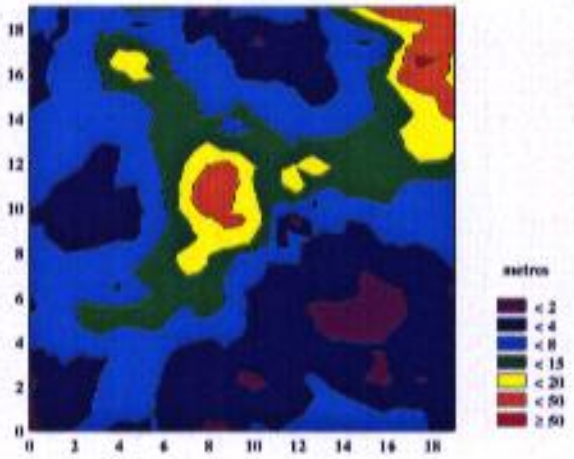
Discharge Category



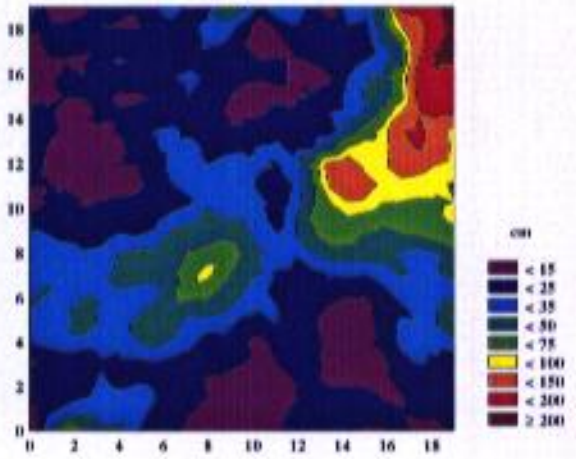
Distance from Source



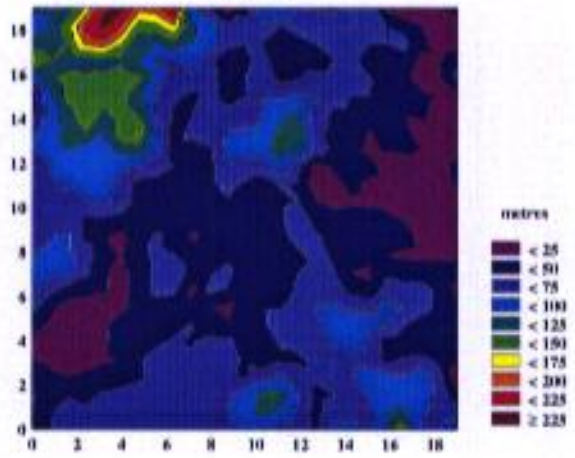
Width



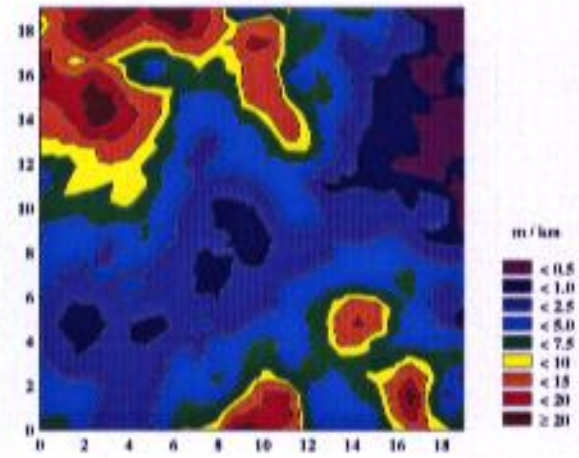
Depth



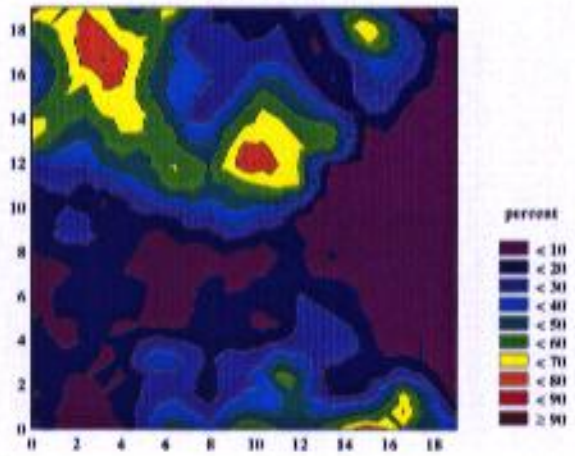
Altitude



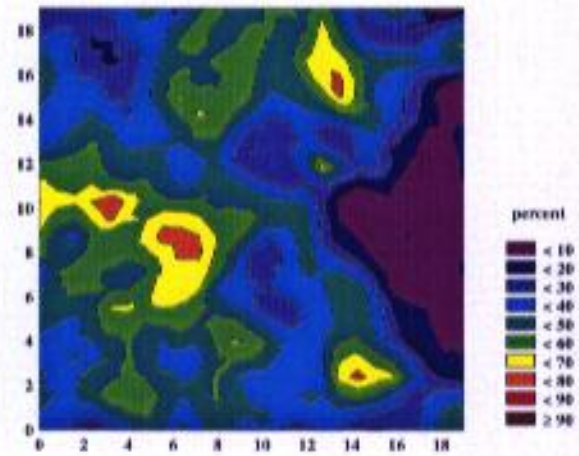
Slope



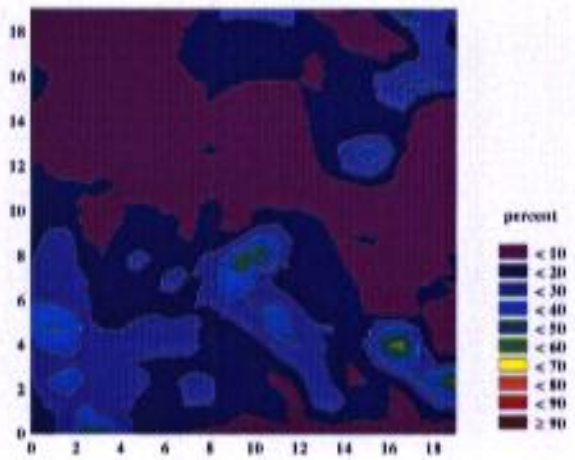
Boulders (%)



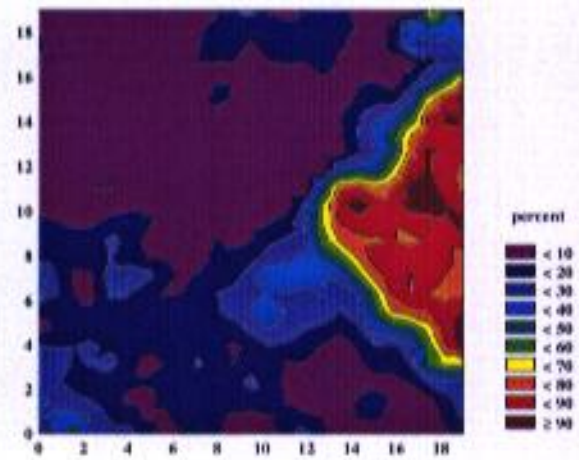
Pebbles (%)



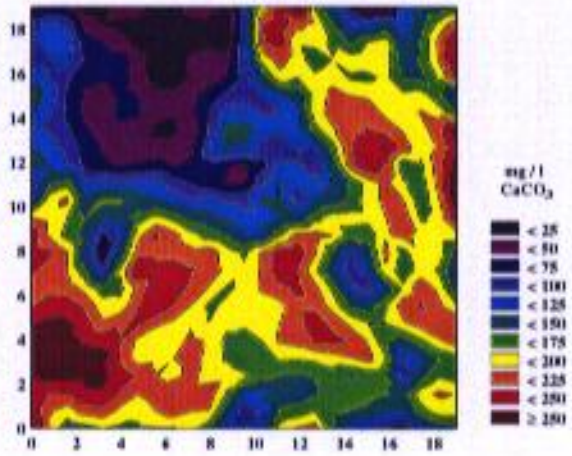
Sand (%)



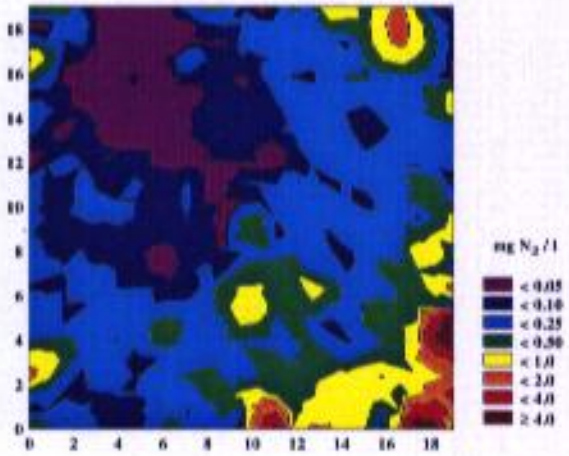
Silt (%)



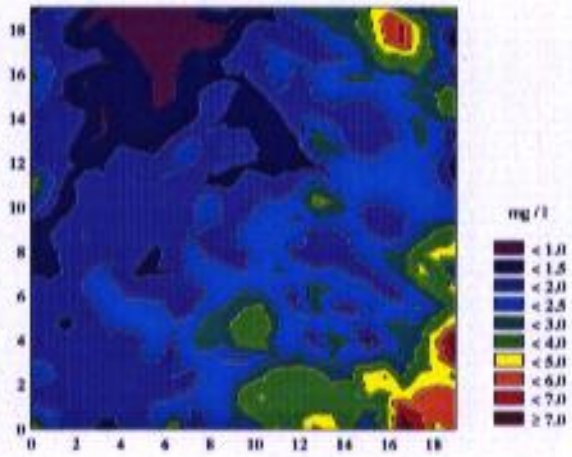
Alkalinity



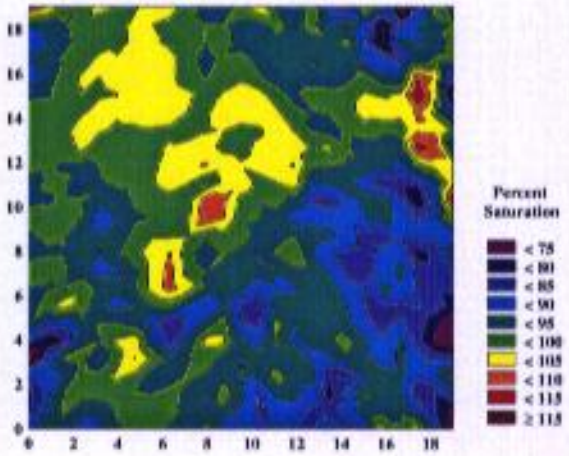
Ammonia



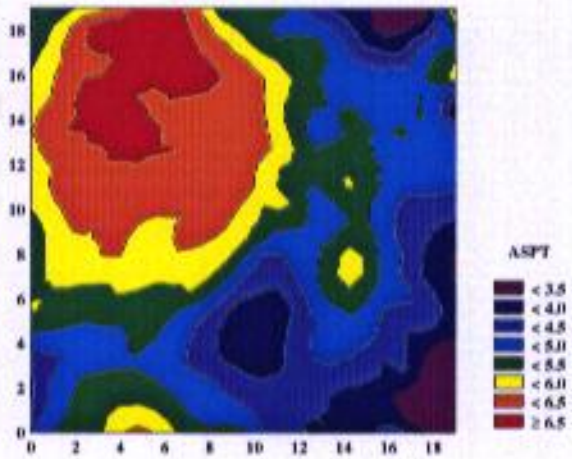
Biochemical Oxygen Demand



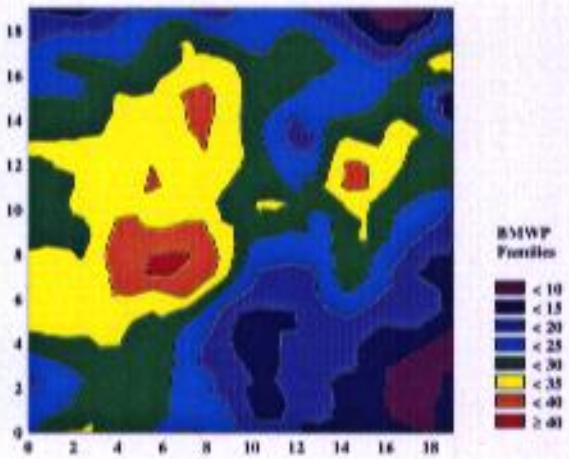
Dissolved Oxygen (%)



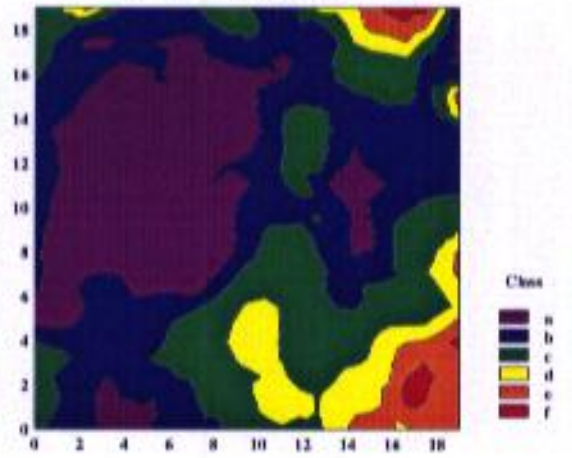
ASPT



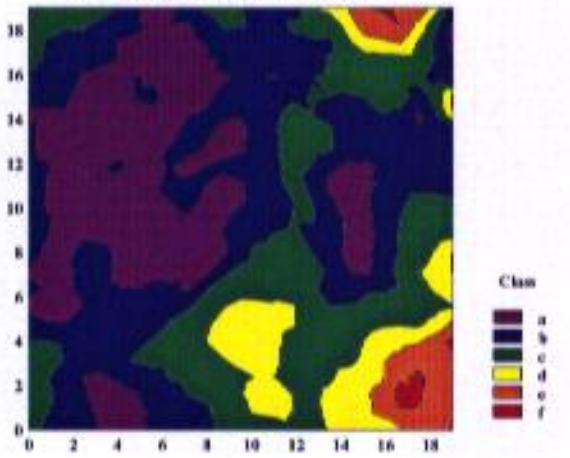
Number of Families



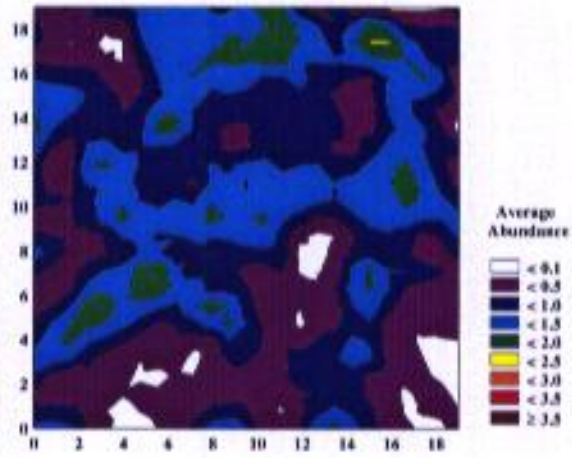
GQA - RIVPACS



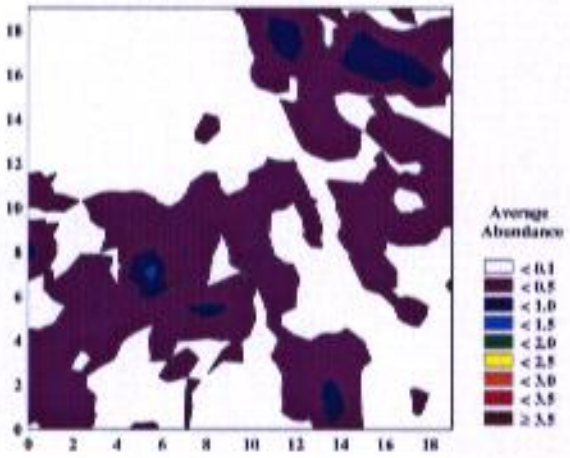
GQA - Neural Net (RScrs)



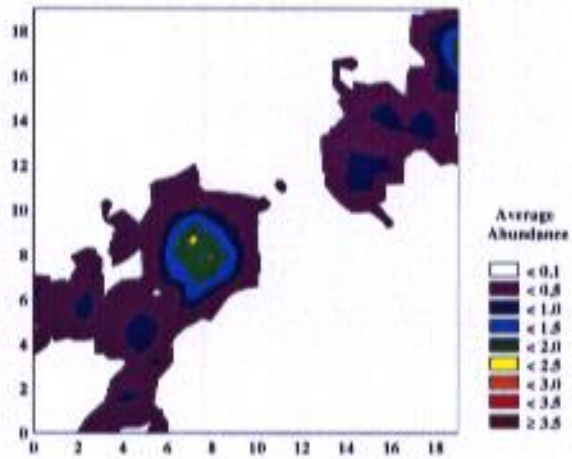
Planariidae



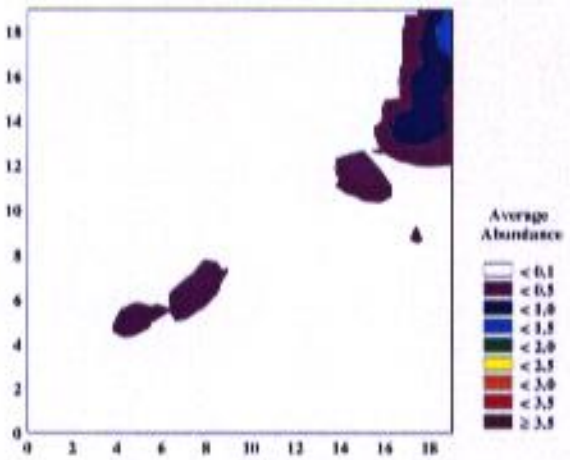
Dendrocoelidae



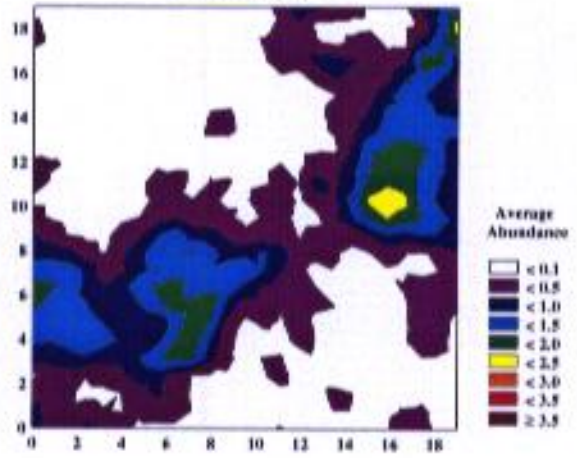
Neritidae



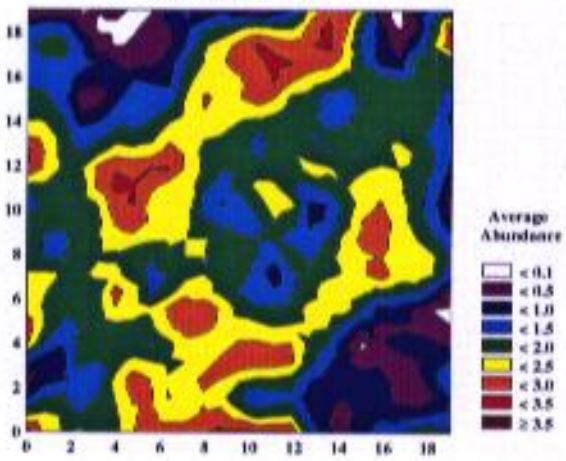
Viviparidae



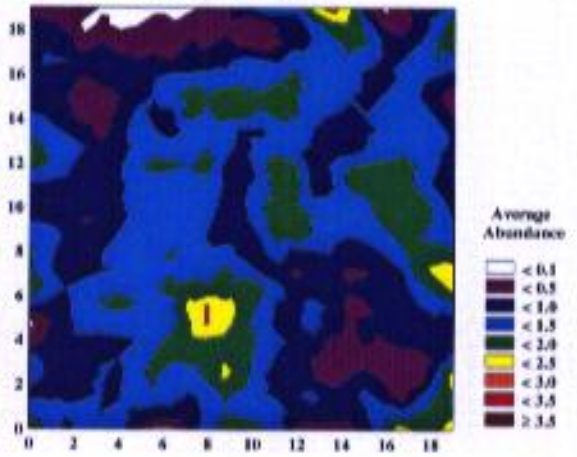
Valvatidae



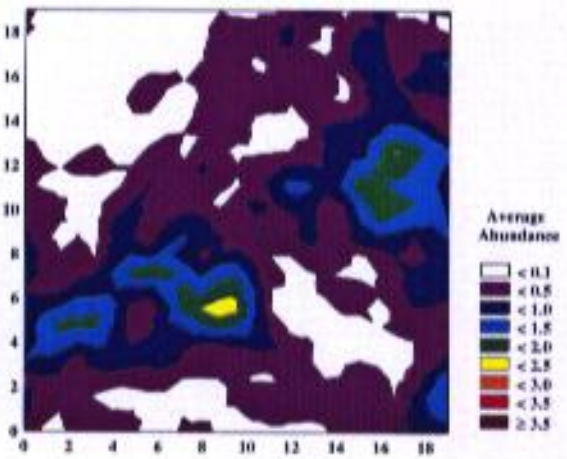
Hydrobiidae



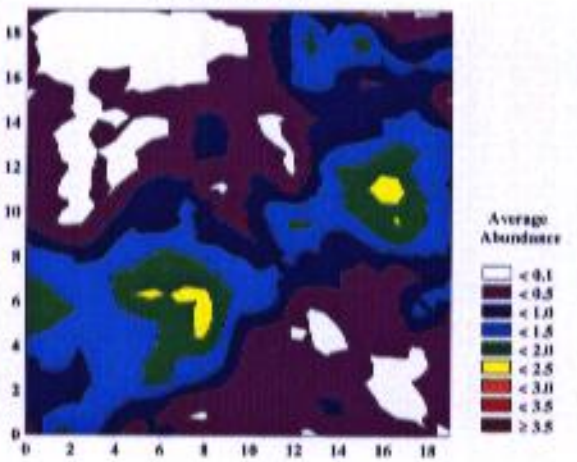
Lymnaeidae



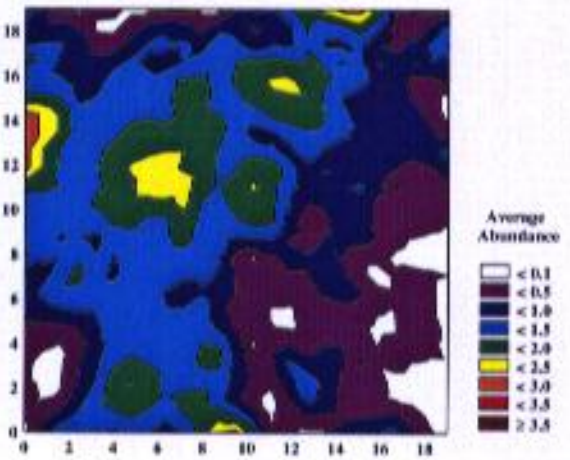
Physidae

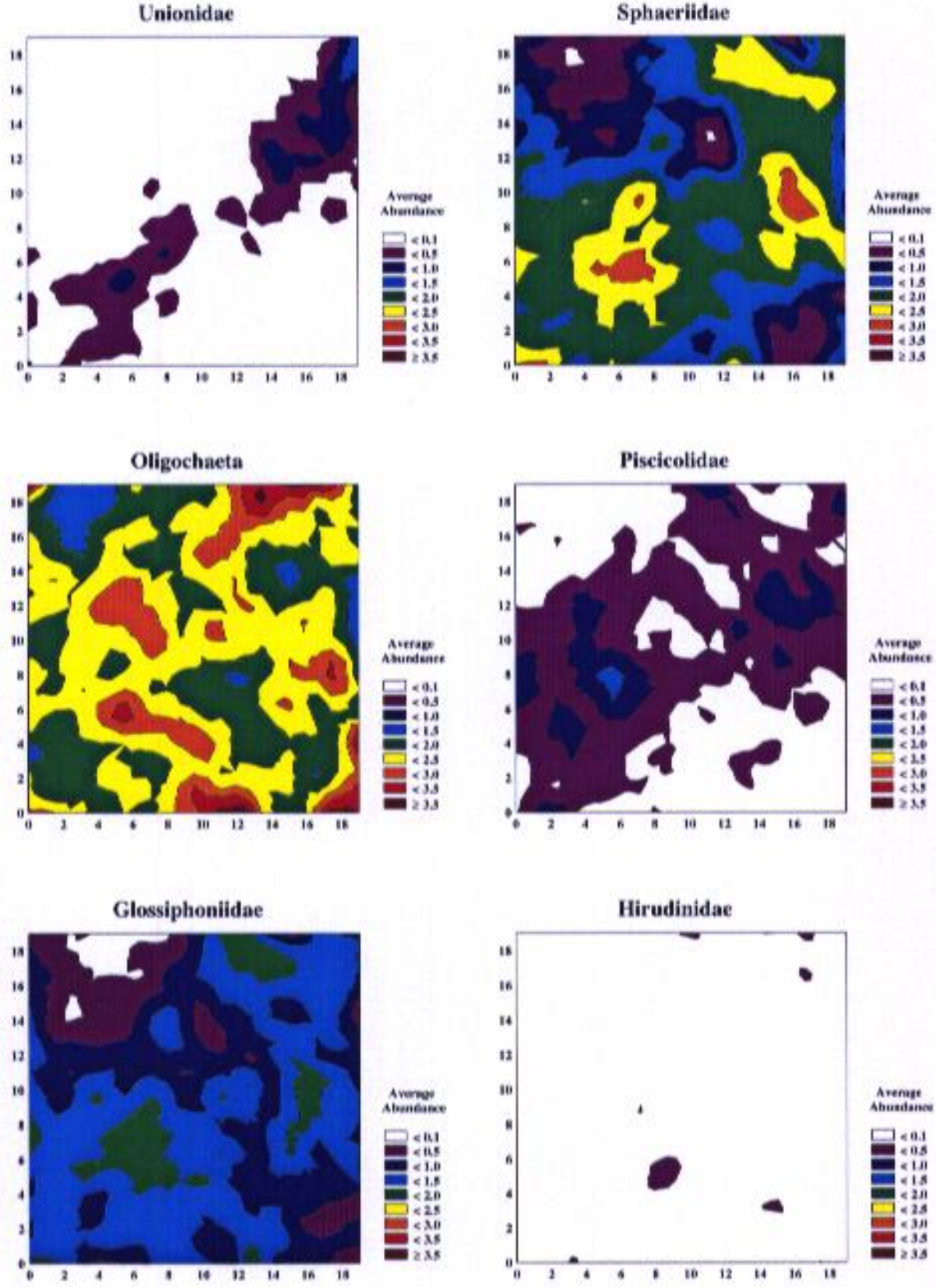


Planorbidae

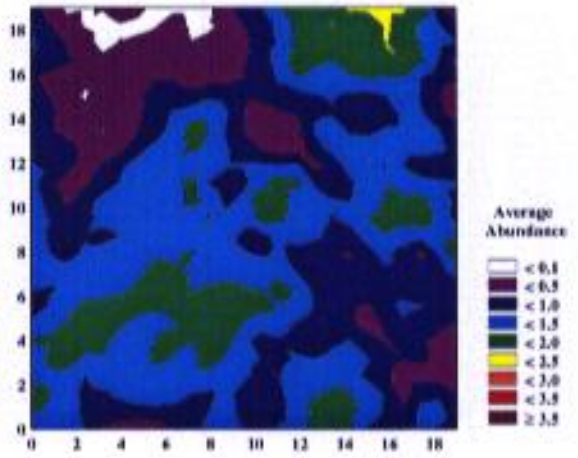


Ancylidae

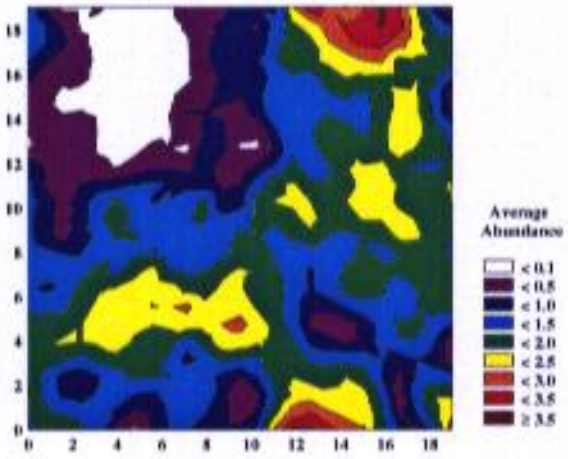



Unionidae
Sphaeriidae
Oligochaeta
Piscicolidae
Glossiphoniidae
Hirudinidae
Average Abundance
< 0.1
< 0.5
< 1.0
< 1.5
< 2.0
< 2.5
< 3.0
< 3.5
≥ 3.5

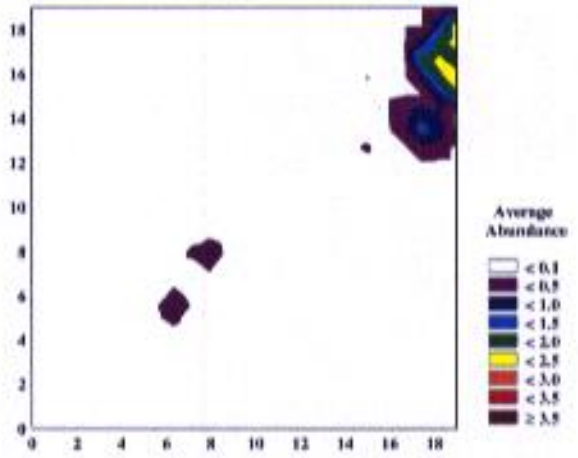
Erpobdellidae



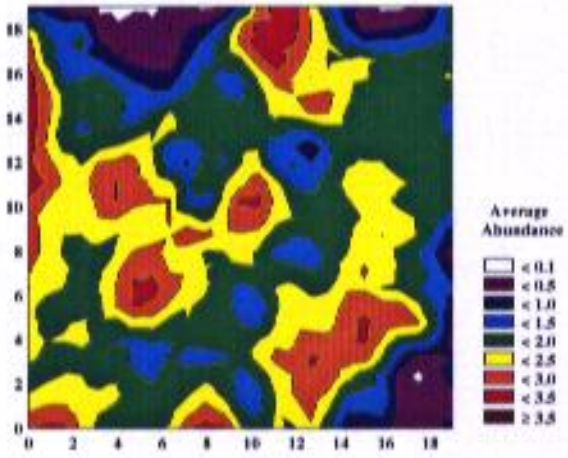
Asellidae



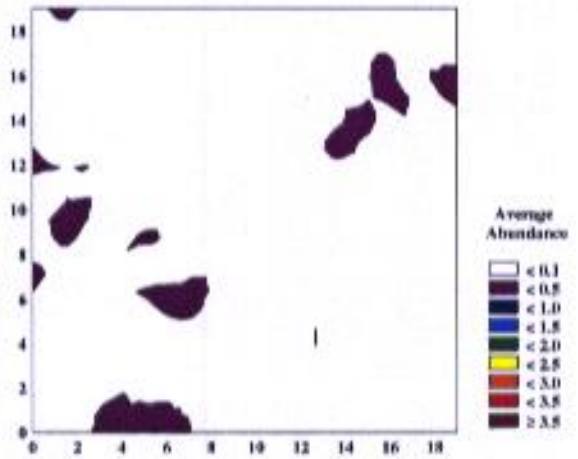
Corophiidae



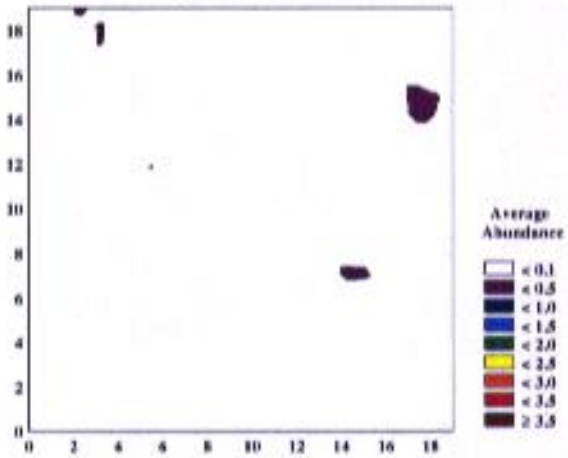
Gammaridae



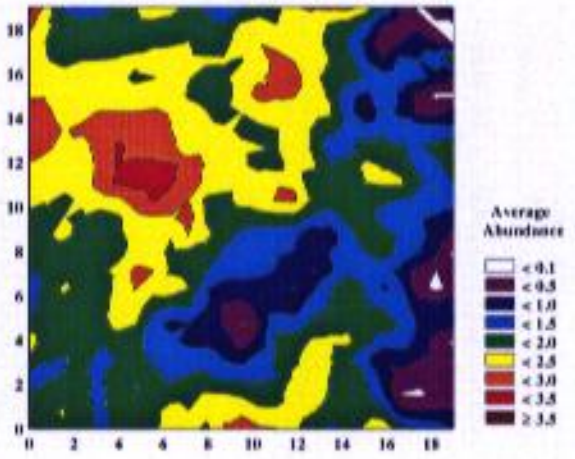
Astacidae



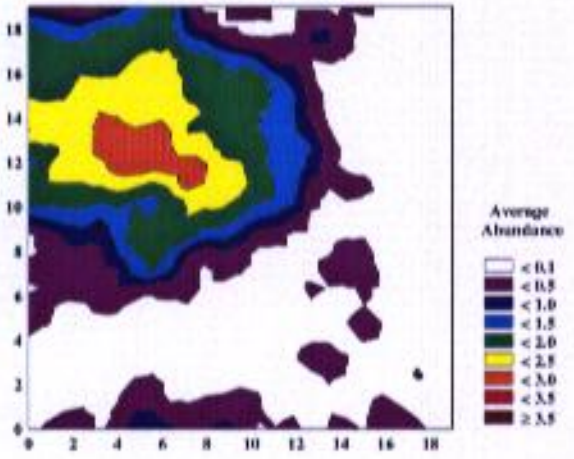
Siphlonuridae



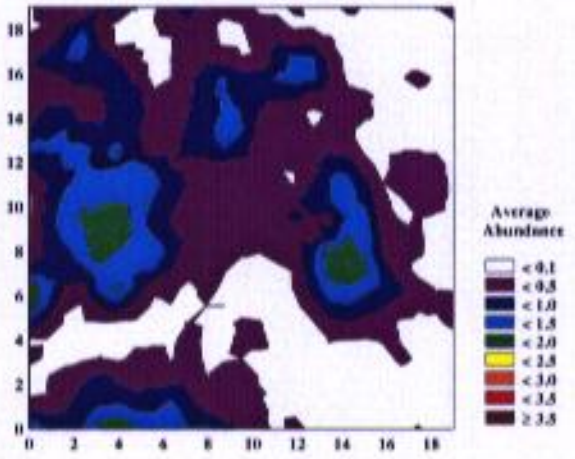
Baetidae



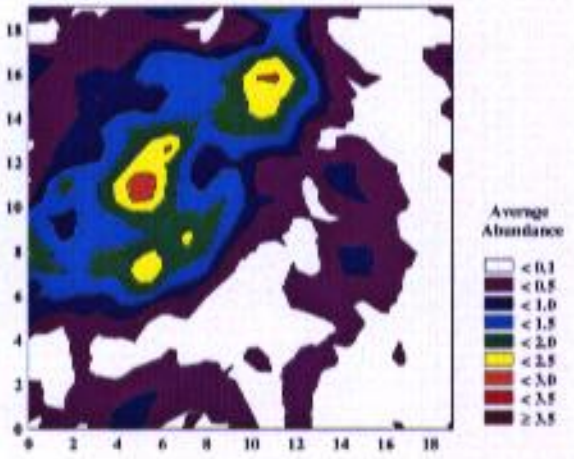
Heptageniidae



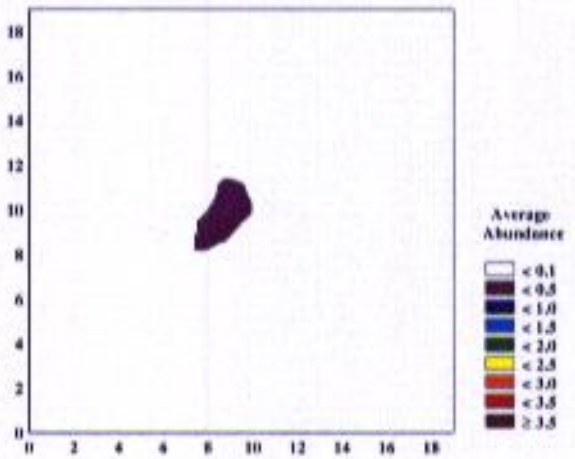
Leptophlebiidae



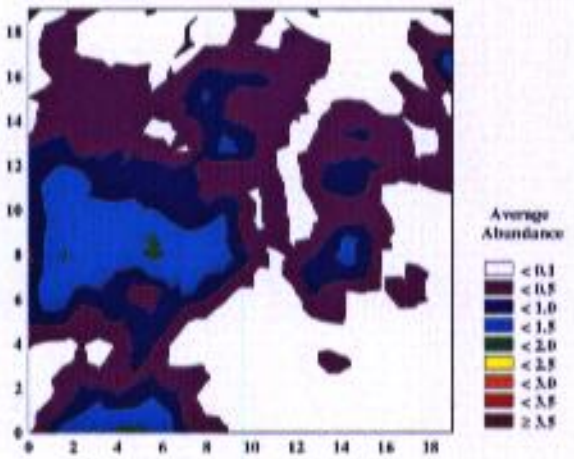
Ephemerellidae



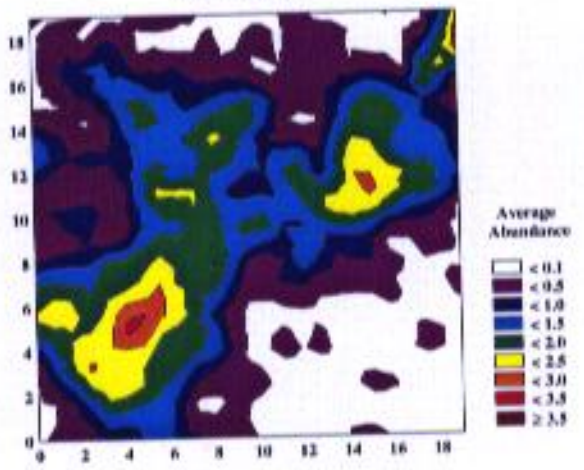
Potamanthidae



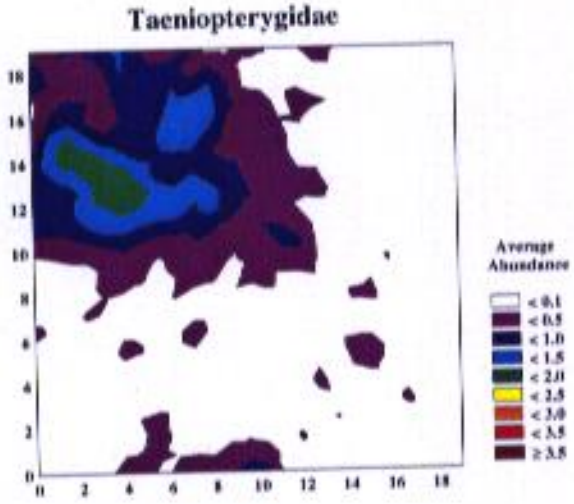
Ephemeridae



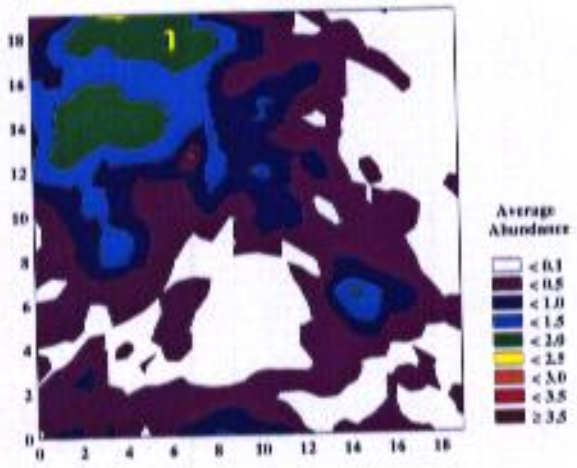
Caenidae



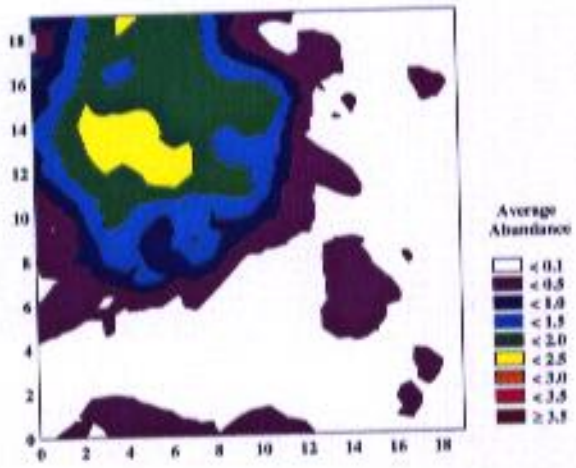
Taeniopterygidae



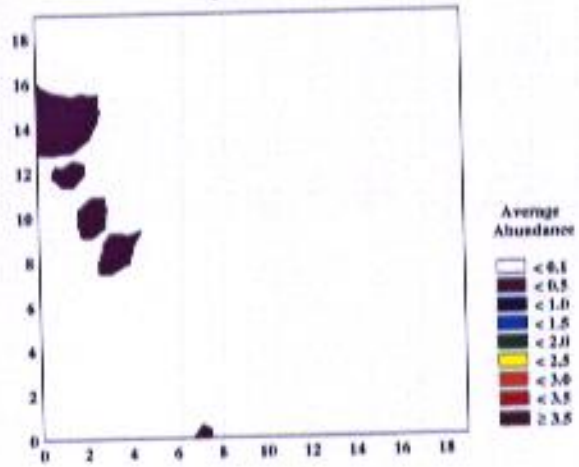
Nemouridae



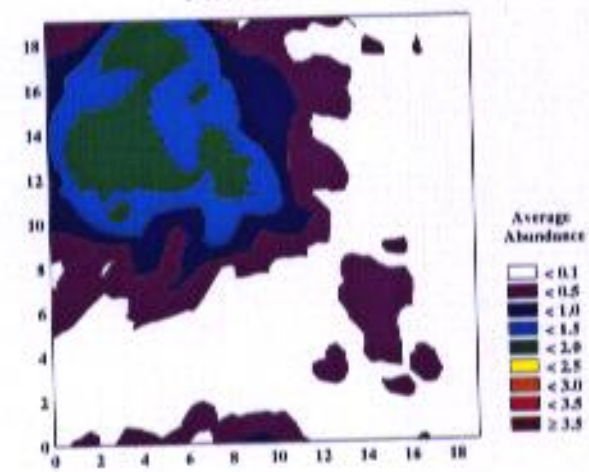
Leuctridae

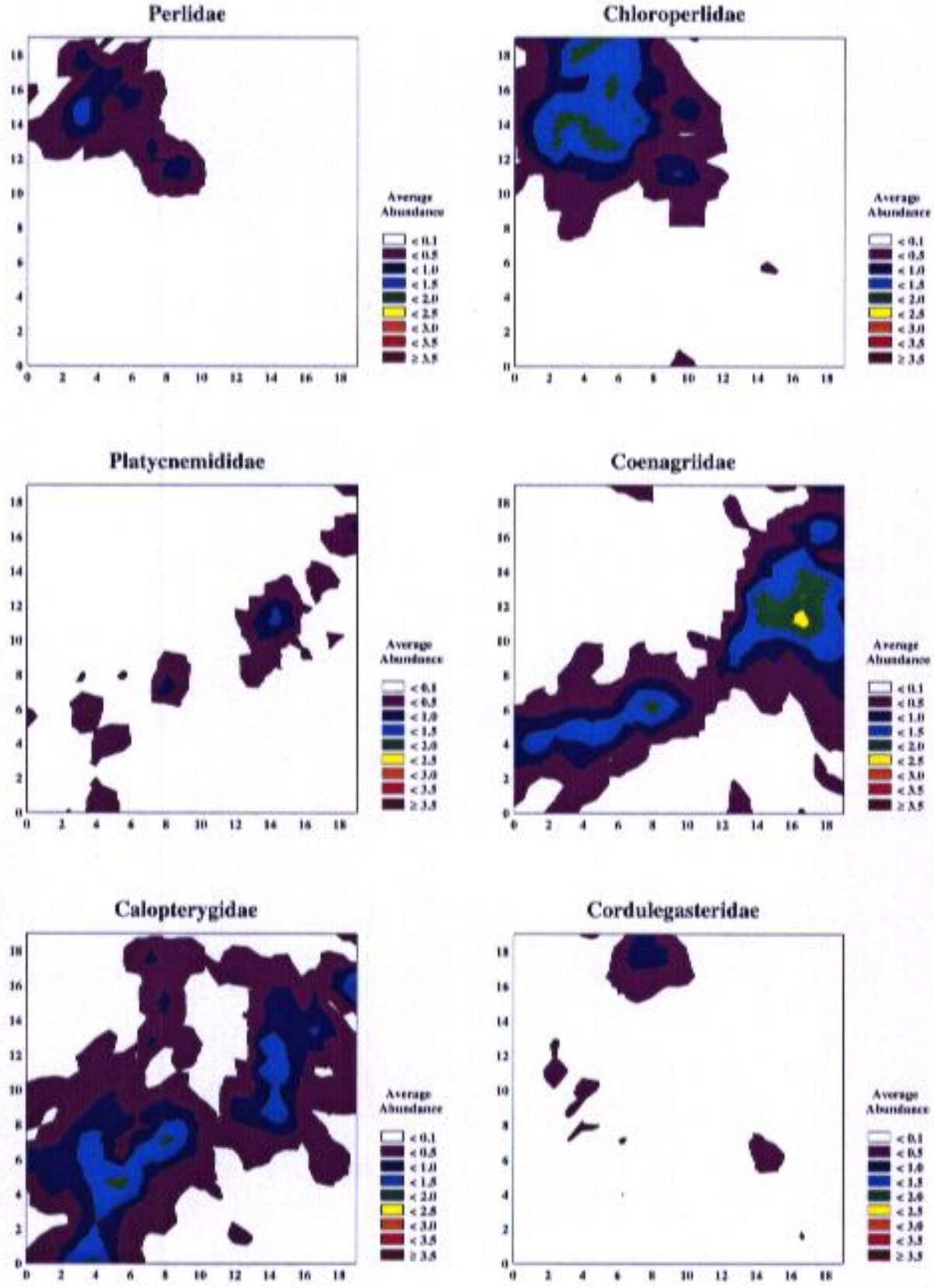


Capniidae

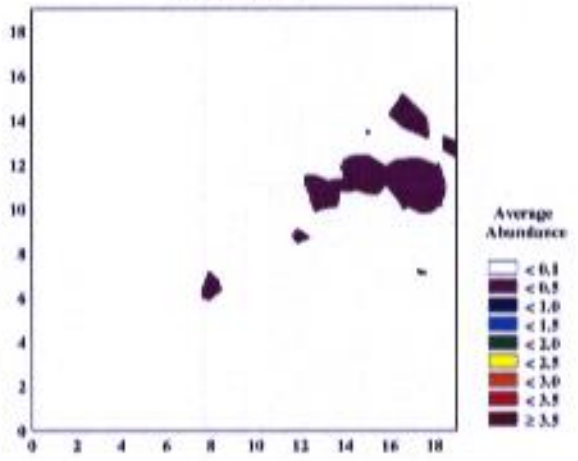


Perlodidae

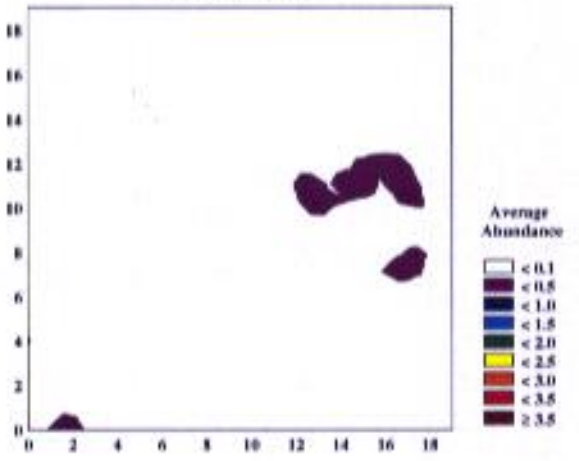



Perlidae
Chloroperlidae
Platycnemididae
Coenagriidae
Calopterygidae
Cordulegasteridae
Average Abundance
< 0.1
< 0.5
< 1.0
< 1.5
< 2.0
< 2.5
< 3.0
< 3.5
≥ 3.5

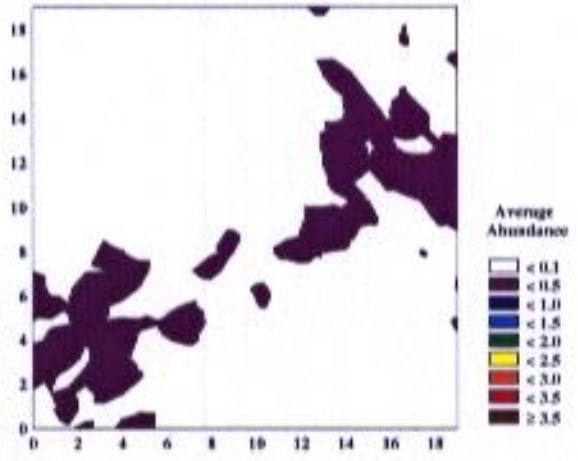
Aeshnidae



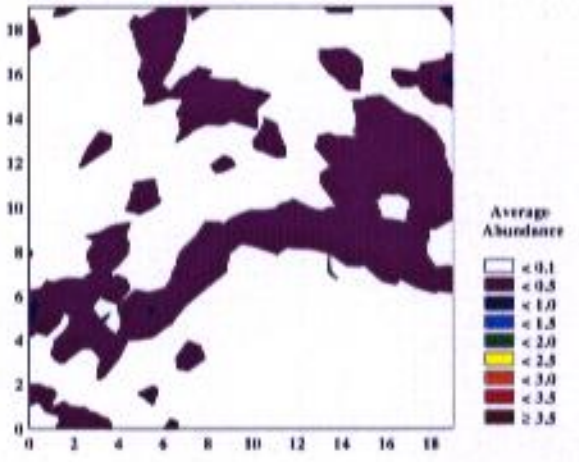
Libellulidae



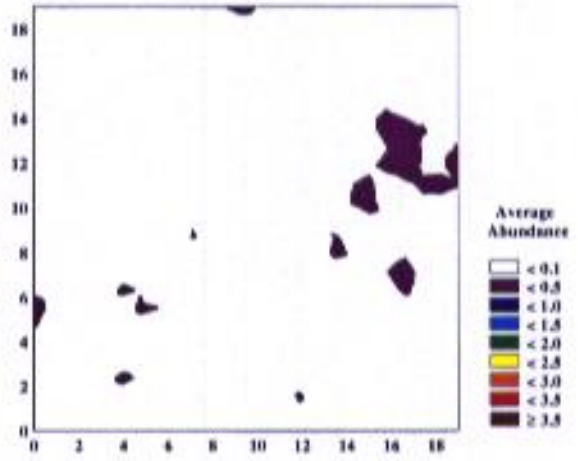
Hydrometridae



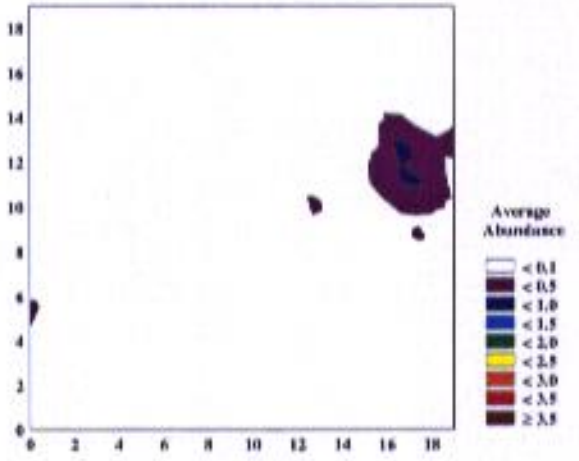
Gerridae



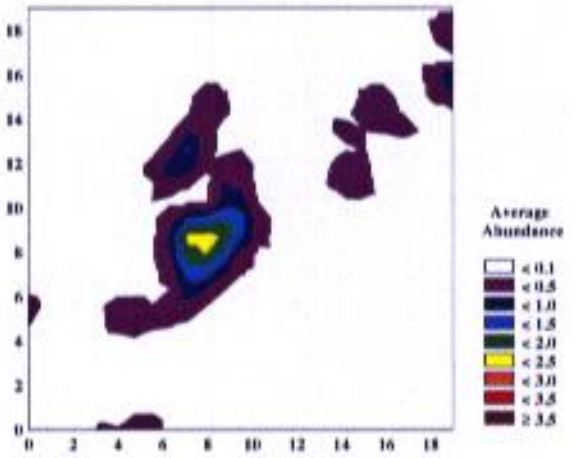
Nepidae



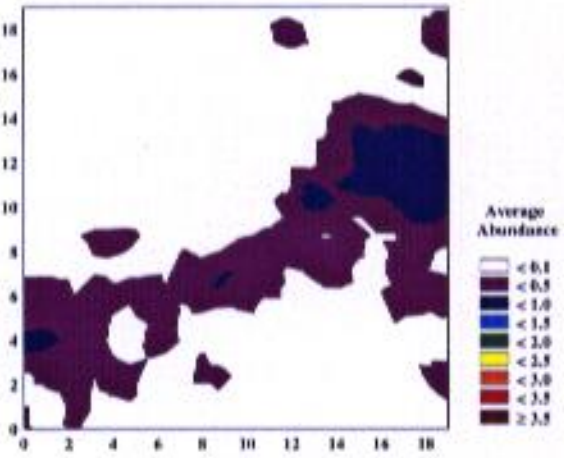
Naucoridae



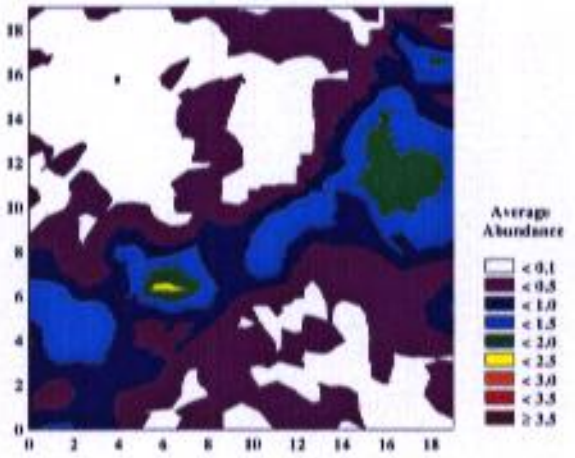
Aphelocheiridae



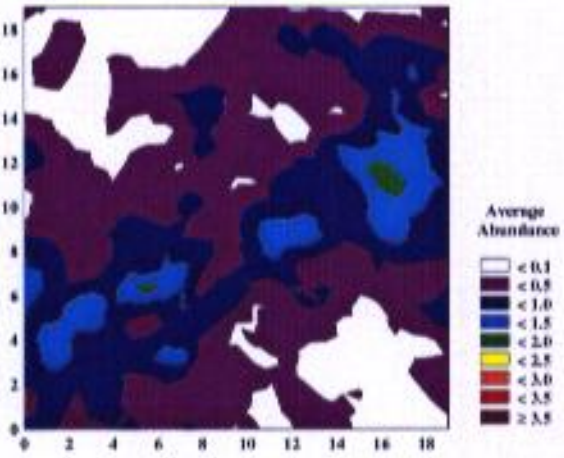
Notonectidae



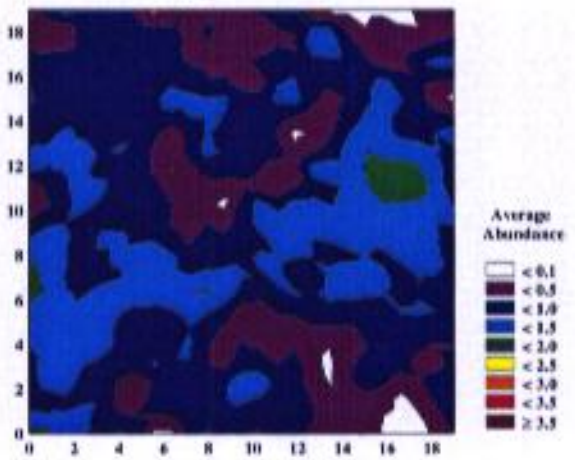
Corixidae



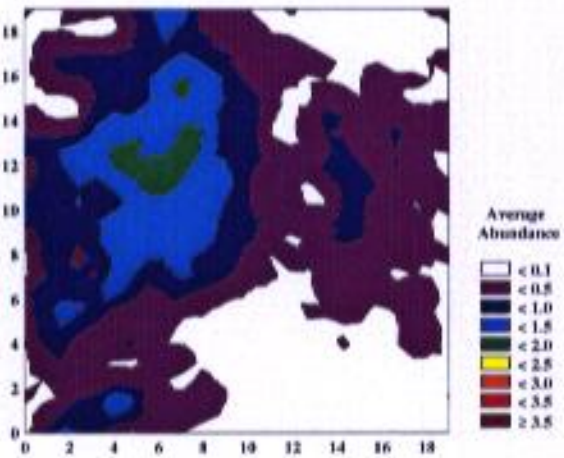
Haliplidae

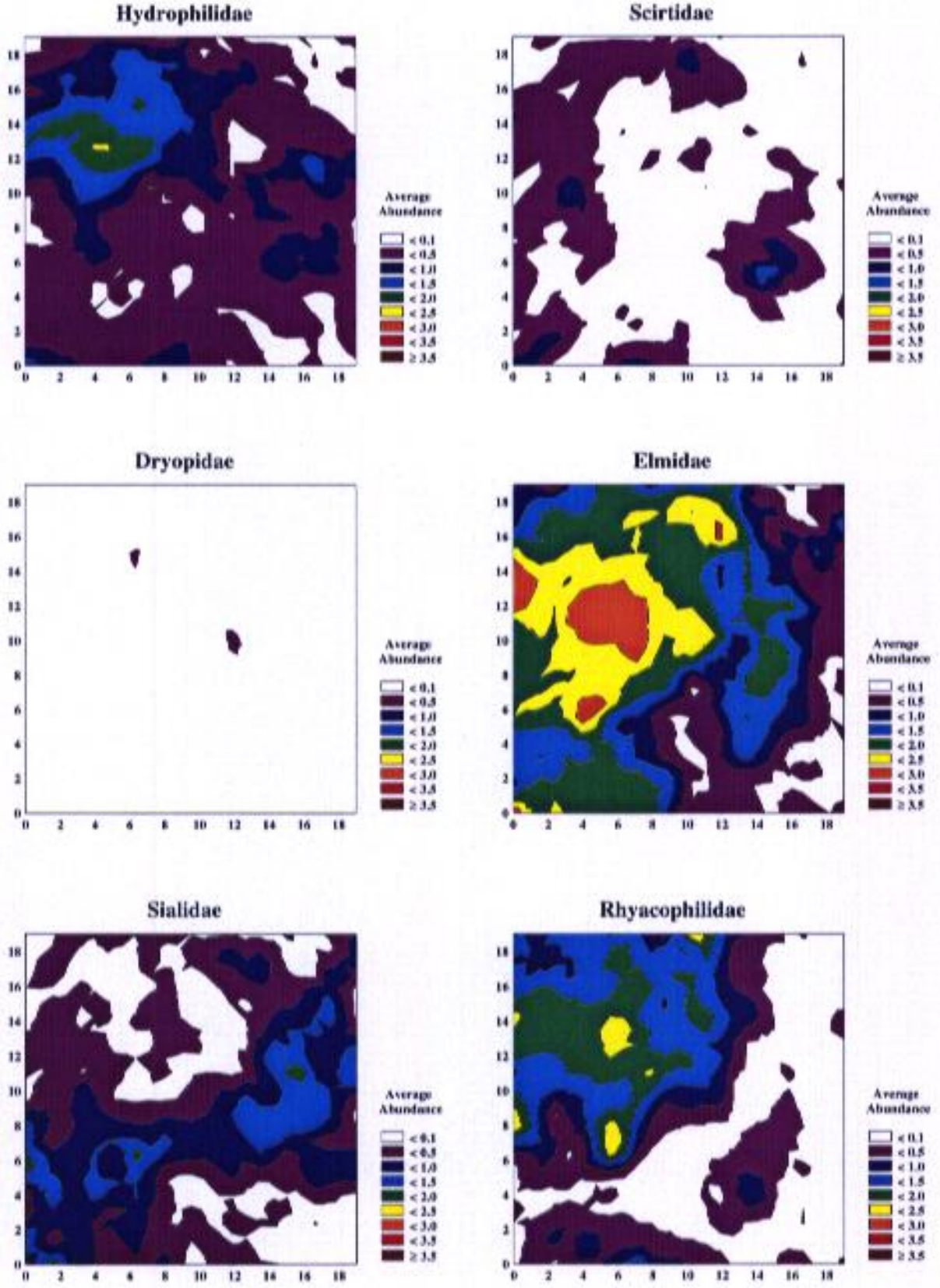


Dytiscidae

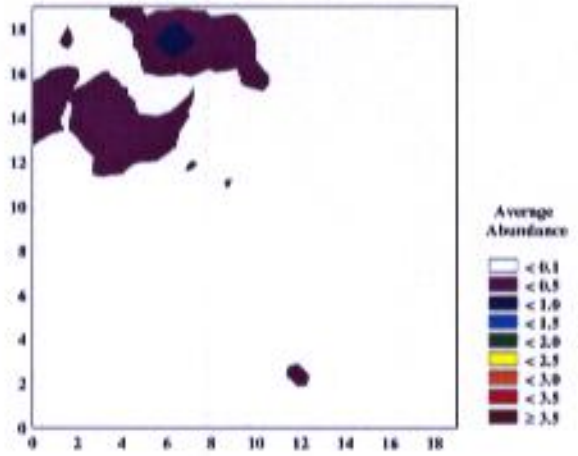


Gyrinidae

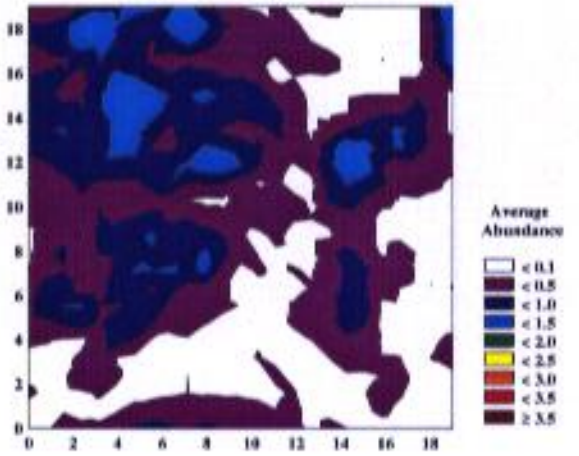



Hydrophilidae
Scirtidae
Dryopidae
Elmidae
Sialidae
Rhyacophilidae
Average Abundance
< 0.1
< 0.5
< 1.0
< 1.5
< 2.0
< 2.5
< 3.0
< 3.5
≥ 3.5

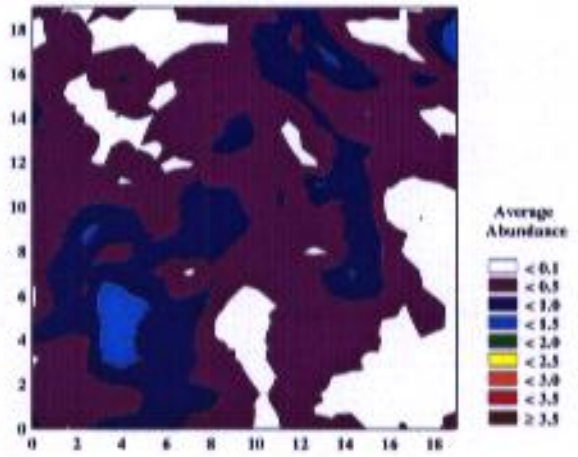
Philopotamidae



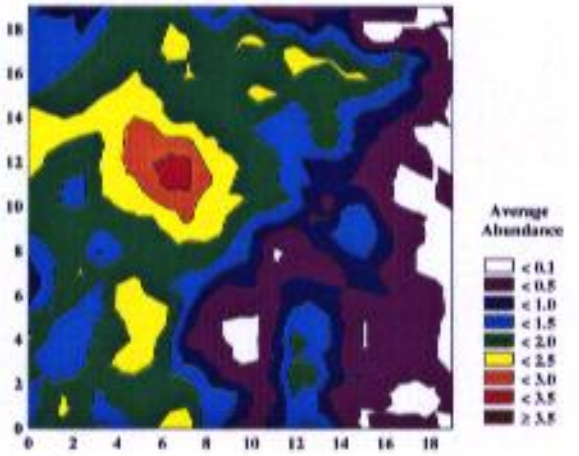
Polycentropodidae



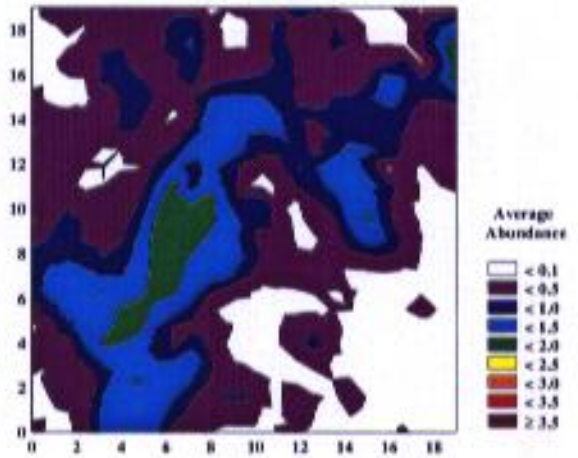
Psychomyiidae



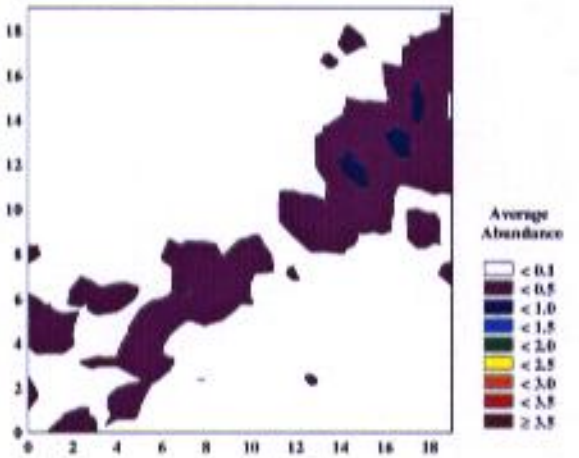
Hydropsychidae



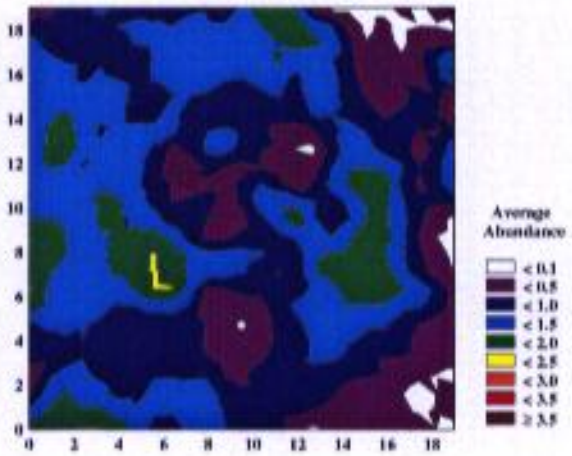
Hydroptilidae



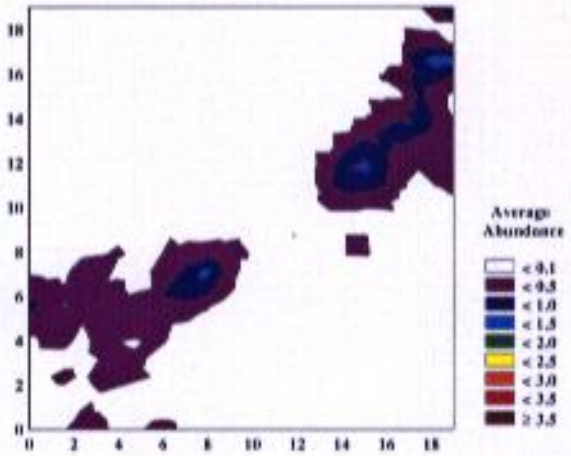
Phryganeidae



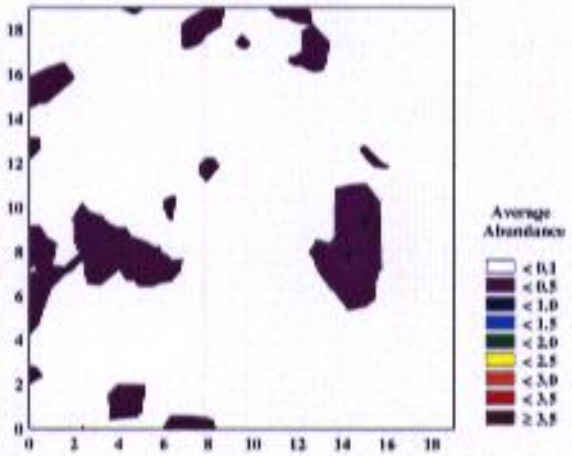
Limnephilidae



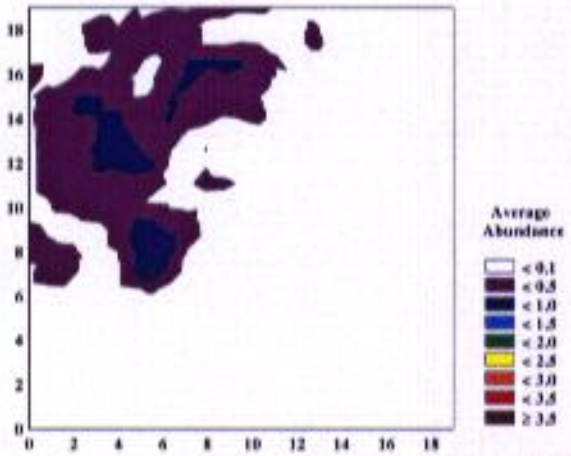
Molannidae



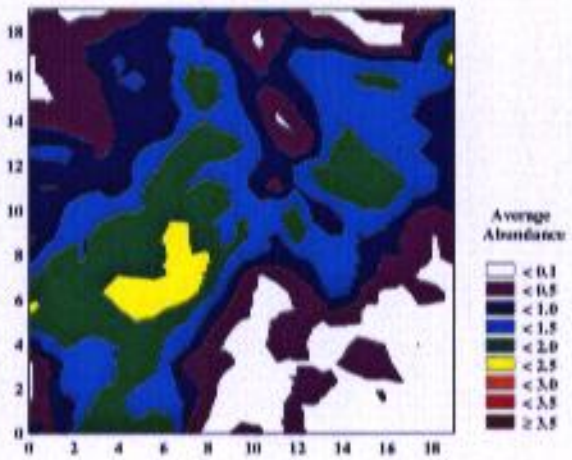
Beraeidae



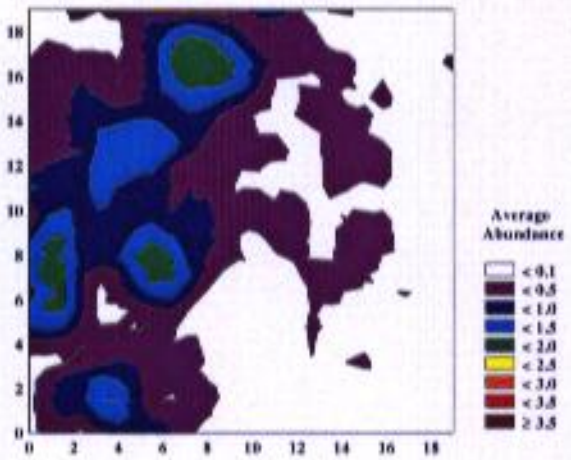
Odontoceridae



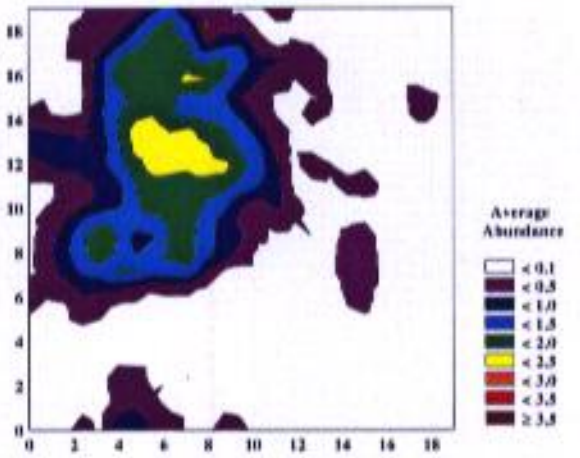
Leptoceridae



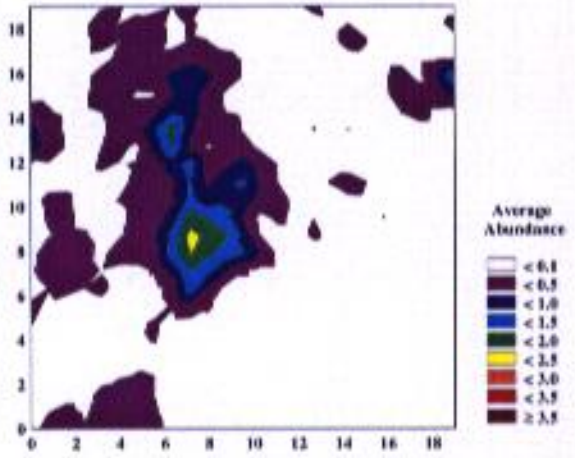
Goeridae



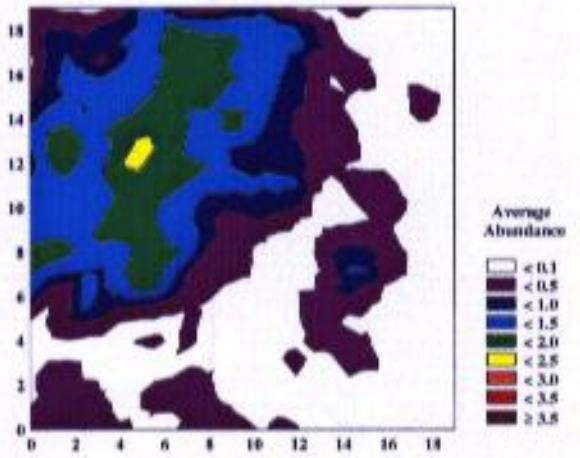
Lepidostomatidae



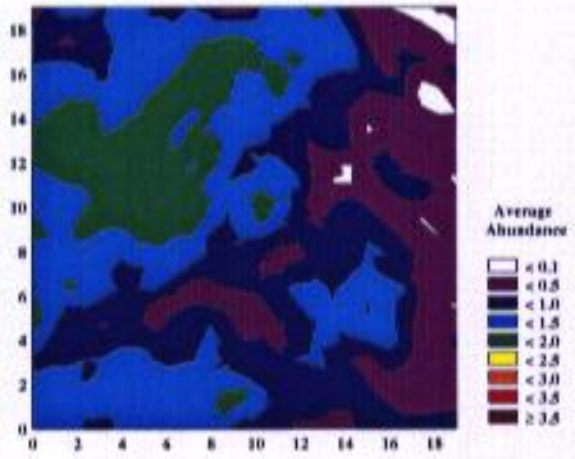
Brachycentridae



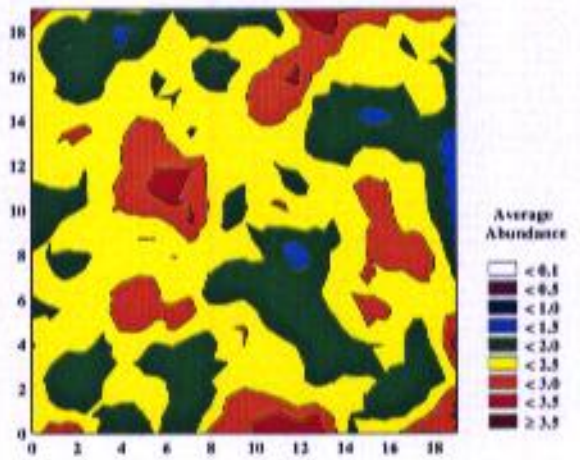
Sericostomatidae



Tipulidae



Chironomidae



Simuliidae