



Refinement of artificial intelligence-based systems for diagnosing and predicting river health

Report – SC030189

The Environment Agency is the leading public body protecting and improving the environment in England and Wales.

It's our job to make sure that air, land and water are looked after by everyone in today's society, so that tomorrow's generations inherit a cleaner, healthier world.

Our work includes tackling flooding and pollution incidents, reducing industry's impacts on the environment, cleaning up rivers, coastal waters and contaminated land, and improving wildlife habitats.

This report is the result of research commissioned and funded by the Environment Agency.

Published by:

Environment Agency, Horizon House, Deanery Road,
Bristol, BS1 5AH

www.environment-agency.gov.uk

ISBN: 978-1-84911-243-7

© Environment Agency – September 2011

All rights reserved. This document may be reproduced with prior permission of the Environment Agency.

The views and statements expressed in this report are those of the author alone. The views or statements expressed in this publication do not necessarily represent the views of the Environment Agency and the Environment Agency cannot accept any responsibility for such views or statements.

Further copies of this report are available from our publications catalogue: <http://publications.environment-agency.gov.uk> or our National Customer Contact Centre: T: 08708 506506
E: Henquiries@environment-agency.gov.uk.

Author(s):

Paisley M.F., Trigg, D.J., Martin, R., Walley, W.J., Andriaenssens, V*, Buxton, R., & O'Connor, M.

Dissemination Status:

Publicly available

Released to all regions

Keywords:

River invertebrates, biomonitoring, pollution, diagnosis, modelling, pressure, pattern recognition, Bayesian network, artificial intelligence

Research Contractor:

Centre for Intelligent Environmental Systems,
Faculty of Computing, Engineering & Technology,
Staffordshire University, Stafford ST18 0AD
* Environment Agency

Research Collaborator:

Scottish Environment Protection Agency

Environment Agency's Project Manager:

John Murray-Bligh, Operations Directorate

Project Number:

SC030189

Product Code:

SCHO0911BUBL-E-E

Evidence at the Environment Agency

Evidence underpins the work of the Environment Agency. It provides an up-to-date understanding of the world about us, helps us to develop tools and techniques to monitor and manage our environment as efficiently and effectively as possible. It also helps us to understand how the environment is changing and to identify what the future pressures may be.

The work of the Environment Agency's Evidence Directorate is a key ingredient in the partnership between research, guidance and operations that enables the Environment Agency to protect and restore our environment.

This report was produced by the Research, Monitoring and Innovation team within Evidence. The team focuses on four main areas of activity:

- **Setting the agenda**, by providing the evidence for decisions;
- **Maintaining scientific credibility**, by ensuring that our programmes and projects are fit for purpose and executed according to international standards;
- **Carrying out research**, either by contracting it out to research organisations and consultancies or by doing it ourselves;
- **Delivering information, advice, tools and techniques**, by making appropriate products available.

Miranda Kavanagh
Director of Evidence

Executive Summary

This report presents the results of a project funded by the Environment Agency which builds on the previous creation of two software systems to diagnose and predict river health from biological and environmental data, namely the River Pressure Diagnostic System (RPDS) and the River Pressure Bayesian Belief Network (RPBBN). RPDS is a pattern recognition system to diagnose likely pressures at a river site. RPBBN is a reasoning system that can diagnose chemical concentrations from a biological community, or predict likely changes in a biological community from changes in chemical concentrations.

An early aim of our project was to use the RPDS database to define chemical standards needed to protect ecological quality. This was achieved by developing Thresholder, a software application which searches the 1995 river survey database to determine the chemical concentrations needed to support the invertebrate fauna predicted by RIVAPCS (River Invertebrate Prediction and Classification System) at all general quality assessment sites. A second early objective was to use the same database to help determine potential reference sites to act as targets for temporal trajectories in RPDS.

The main aims of the project were to enhance the software systems RPDS and RPBBN. The specific objectives were as follows:

- Substantially extend the dataset on which the data models are based.
- Revise and test the data models on which the systems are based.
- Extend the functionality of the two systems and combine into one 'integrated system'.

Extending the dataset took much longer than originally anticipated and affected the progress of the remaining work. In particular, the integrated system could not be developed in the remaining time (the third objective), and the two software systems have been kept separate.

The dataset has been substantially extended temporally, geographically and in terms of the variables included. The new dataset covers the ten-year period 1995-2004 instead of the single year 1995. The dataset covers Scotland as well as England and Wales. Variables relating to flow (except for Scotland), geology, land cover and land risk have been added to the chemistry and stress as diagnostic variables. The resulting spring and autumn datasets contain over five times more biological samples than the original systems.

To reflect increasing interest in linking changes in the biological community with the physical flow of rivers, two measures of flow were included in the new diagnostic variables. The first was the percentage impact (at 95 per cent exceedence probability) from LowFlows2000. To complement this, a second measure was developed to estimate the flow condition at each site at the time the sample was taken. This was based on interpolation from thirty years of monthly flow records at gauged sites in England and Wales. The ecological significance of this measure was demonstrated by the fact that those taxa more likely to occur in wet conditions were more sensitive, and those more likely to occur in dry conditions were more tolerant, according to their revised biological monitoring working party (BMWP) scores.

Following preliminary tests, new MIR-max models were produced for the full dataset for both spring and autumn. In all cases, the number of bins (clusters of samples with

similar biological composition) was kept the same as in the original models, namely 250. The original spring and autumn models contained more than 6,000 samples each (with an average of 24 samples in each bin), whereas the new spring model contains 32,100 samples and the new autumn model 31,400 (with averages of 128 and 126 samples in each bin respectively). The sample data in the original spring and autumn models was from 6,000 sites in England and Wales, whereas the sample data in the new spring and autumn models cover 9,100 and 8,800 sites respectively in England, Wales and Scotland.

The new spring and autumn cluster models were ordered by MIR-max to produce hexagonal output maps with the same side-length (10, 15 and 20 bins) as the original models. Each map was rotated to align it as closely as possible with the originals to ease comparison. After adding the new diagnostic variables to the spring and autumn models, RPDS 2.0 was revised to RPDS 3.0 by streamlining the operations involving database queries and including Scotland on the geographical map panel. The output maps in RPDS 3.0 for particular variables are qualitatively similar to those of the same variables in RPDS 2.0, demonstrating that the clustering and ordering in the new models are similar to those of the original models despite the large increase in data. From this we conclude that the models are sound and represent reality rather than an artefact of the sample data, and that the models contain sufficient data: more data is unlikely to affect the overall models. The geographic locations of the samples in clusters occupying similar positions in the hexagonal output map are also similar in RPDS 2.0 and RPDS 3.0.

Preliminary evaluation of the new flow variables shows that the percentage impact at 95-percentile flow (flow exceeded 95 per cent of the time) is negatively correlated with distance from source, as might be expected. Preliminary evaluation of the flow condition variable, on the other hand, shows a relationship with the taxa as the clusters containing samples taken in wetter years tend to be those with higher average score per taxon (ASPT), while those taken in drier years tend to be those with lower ASPT.

As with the MIR-max models, the new Bayesian Belief Network (BBN) model was derived from a substantially larger quantity of data. The original BBN model was based on 3,600 spring and autumn matched samples, whereas the new spring and autumn models are based on 16,200 and 15,800 matched samples respectively. Several other changes were made. The structure of the model was modified. The larger dataset meant that chemical statistics could be based on percentile values over three years prior to sampling (the chemical statistics that are used to manage water quality) rather than mean values over the preceding three months, and that five states could be used for the taxonomic variables instead of four.

Dependent testing of the network including all of these changes against the original indicated major improvements in the predictions of total ammoniacal nitrogen and dissolved oxygen. The prediction of the flow condition variables was poor, but this was not unexpected given the far fewer connections to the taxonomic variables. Independent testing to assess the impact of each of the changes suggested that the use of percentile statistics, permitted by the increased dataset, was the factor contributing to the greatest improvements.

Following meetings with potential users, new versions of the RPDS and RPBBN software have been produced which incorporate several modifications designed to optimise them for operational use. Their usability has been improved and additional functionality has been incorporated to meet the requirements of the Water Framework Directive.

Acknowledgement

Thanks are expressed to the following Environment Agency staff for their help and support on on this project:

- John Murray-Bligh
- Amanda Veal
- Richard Hemsworth
- Paul Logan
- Ian Humpheryes
- Emma Pemberton
- Chris Moore
- Alice Hiley
- Emily Parker
- Tomar Chierici
- Jan Stipala
- Rachel Anning
- John Waddingham
- Graeme Storey
- Grant McMellin

We would also like to express our gratitude to the following:

- Robin Guthrie, Scottish Environmental Protection Agency
- Nathan Critchlow-Watton, Scottish Environmental Protection Agency
- Dominic Habron, Scottish Environmental Protection Agency
- Mark Caulfield, Blucel Ltd
- San Shunmuga, Staffordshire University.

Contents

1. Introduction	1
2. Determination of chemical thresholds for GQA sites	4
Introduction	4
Chemical thresholds	4
Rivers Intercalibration Project (R.I.P.)	4
R.I.P. software	5
Outcomes of the R.I.P.	6
‘Pollution Sensitivities of Taxa’ website	7
First phase of development	8
Second phase of development	11
Using the Thresholder software	12
Derivation of chemical threshold values	15
Future work	15
3. Determine potential reference sites	16
Introduction	16
Generate list of proposed reference sites from RPDS database	16
Summarise status of current reference sites according to RPDS	17
4. Construction of project database: biological data	28
Introduction	28
Standardising Environment Agency and SEPA data	28
Duplicate sample data	29
Project taxonomic group list for Environment Agency database	29
Validation by checking for errors	31
Validation using BMWP assessment data	32
Completed biological database	33
5. Construction of project database: environmental and chemical data	37
Introduction	37
Sample environmental data	37
Chemical data	38
Spatial matching	40
Matching biological and chemical sites	41
Derivation of chemical statistics	42
Acid neutralising capacity	43
Use of toxicity data	43
6. Construction of project database: stress and GIS data	45
Stress data	45
Data from geographical information systems	48
7. Construction of project database: flow data	51
Introduction	51
LowFLows 2000	51
Development of a method for estimating flow condition	52
8. Revision and testing of MIR-max models	63
Review of MIR-max	63
Review of variables in the input vector	65
Reproducing the original models based on 1995 data	65

Criteria for discretising the continuous variables	66
Modifying the bandings for the environmental variables	73
Rationalising environmental variables used in training	74
Final test model	79
Final models with full dataset based on average alkalinity	80
Final models with full dataset based on calcareous geology	81
9. Revision and evaluation of RPDS software	83
Introduction	83
Development of RPDS 3.0 software	83
Quantitative evaluation of RPDS 3.0	83
Extensions to functionality of RPDS	95
10. Modifications to RPDS software	96
Introduction	96
Data tasks	96
Changes to user interface	98
Batch mode	102
User feedback workshop	105
11. Revision and testing of BBN models	107
Review of original network	107
BBN creator	108
Revisions to causal network	109
Summary of dataset used	111
Derivation of conditional probability matrices	112
Testing and evaluation	113
Extensions to functionality of RPBBN	116
12. Modifications to RPBBN software	118
Introduction	118
Changes to RPBBN model	118
Changes to user interface	121
Batch mode	124
Items not undertaken	126
Additional work	127
Rewrite updating procedure	135
User feedback workshop	138
13. Summary and conclusions	139
Background to project	139
Summary of project outcomes	139
Extension of dataset	139
Estimation of flow condition and impact on taxa	140
Revision of MIR-max models	140
Revision of RPDS software	141
Revision of BBN model	141
Revision of RPBBN software	142
Overall conclusion	142
References	143

Glossary

145

Appendix A	Investigation into potential use of toxicity data	148
Appendix B	Impact of flow condition on occurrence of taxa	170
Appendix C	Proposed data specification	174
Appendix D	MIR-max user guide	187
Appendix E	PISCES codes for sector, activity and pressure	209
Appendix F	Stress categories and associated activity, source and pressure codes	218
Appendix G	BBN creator user guide	224

List of Figures

Figure 2.1	Plot options panel of R.I.P. software	5
Figure 2.2	Plot panel showing scatter plot of ASPT against alkalinity and two user-defined range boxes	6
Figure 2.3	Pollution Sensitivities of Taxa website	7
Figure 2.4	Screenshot of Thresholder software with abundance values for taxa loaded from external file containing part of 1995 River Survey data	10
Figure 2.5	Screenshot of Thresholder software showing user defining abundance values manually	10
Figure 2.6	Screenshot of Thresholder software showing report detailing results of a query	11
Figure 2.7	Screenshot of main Thresholder window, showing abundance values for invertebrate fauna of selected sample, and corresponding query report	13
Figure 2.8	Screenshot of Modify Sample/Manual Input dialogue box	14
Figure 2.9	Screenshot of Thresholder Options dialogue box, showing report and query parameters available to user	14
Figure 3.1	Three different membership value distributions that could give same weighted mean result	19
Figure 7.1	Distribution of 124 gauging stations with complete monthly flow records or less than five per cent missing for January 1976 - December 2005	53
Figure 7.2	Variation in ranking score against mean ranking score across 121 stations for February	54
Figure 7.3	Variation in ranking score against mean ranking score across 121 stations for August	55
Figure 7.4	Locations of gauging stations and indication of number of serious errors	59
Figure 8.1	Illustration of distribution of sample data for attribute X_j among n classes	63
Figure 8.2	Screenshot of division of sample distribution of average alkalinity into nine equally sized bands	74
Figure 9.1	(a) Output maps from RPDS 2.0 (spring 1995 model): (i) Number of families, (ii) BOD and (iii) Elmidae	86
Figure 9.1	(b) Output maps from RPDS 3.0 (spring 1995 model): (i) Number of families, (ii) BOD and (iii) Elmidae	86
Figure 9.2	(a) Output maps from RPDS 2.0 (autumn 1995 model): (i) ASPT, (ii) pH and (iii) Heptageniidae	87
Figure 9.2	(b) Output maps from RPDS 3.0 (autumn 1995 model): (i) ASPT, (ii) pH and (iii) Heptageniidae	87
Figure 9.3	(a) Geographic map panel on RPDS 2.0 showing spatial locations of samples in cluster containing least diverse sites	88
Figure 9.3	(b) Geographic map panel in RPDS 3.0 showing locations of samples in cluster corresponding closely to that illustrated in Figure 9.3 (a)	88
Figure 9.4	(a) Geographic map panel on RPDS 2.0 showing spatial locations of samples in cluster containing samples with high ASPT values	89
Figure 9.4	(b) Geographic map panel in RPDS 3.0 showing locations of samples in cluster corresponding closely to that illustrated in Figure 9.4(a)	89
Figure 9.5	New geological variables in RPDS	90
Figure 9.6	New land cover variable in RPDS3	91
Figure 9.7	(a) Output maps in spring model for (i) percentage impact at Q95 and (ii) distance from source; (b) Output maps in autumn model for (i) percentage impact at Q95 and (ii) distance from source.	92
Figure 9.8	Flow condition variable in RPDS 3.0 spring model	93
Figure 9.9	Flow condition variable in RPDS 3.0 spring model	94
Figure 10.1	Description box below hexagon for total ammoniacal nitrogen	99

Figure 10.2	Image and descriptive text for indicator Caenidae	100
Figure 10.3	Range of scale values for selected indicators on current cluster only	100
Figure 10.4	Range of scale values for selected indicators on current cluster and archive sample data	101
Figure 10.5	Range of scale values for selected indicators when comparing current cluster and input sample data	101
Figure 10.6	Activation of batch mode with dialogue box for selecting options	104
Figure 10.7	Report following the run in batch mode	105
Figure 11.1	Causal belief network of original model	107
Figure 11.2	Revised causal belief network	109
Figure 12.1	Screenshot of report display panel showing information on predictions and index values for individual variables and summary for whole network	123
Figure 12.2	Screenshot of charts display panel with current and stored states	124
Figure 12.3	Batch mode 'Input Variable Selection' dialogue box	125
Figure 12.4	Batch progress dialogue box	125
Figure 12.5	Screenshot of prototype 'Group Manager' application, designed to allow users to construct and modify templates	128
Figure 12.6	Chart panel and report bar	130
Figure 12.7	Screenshot of RPBBN with 'Manual Propagation' option selected, prior to updating of the model with new evidence	131
Figure 12.8	Domain manager dialogue box	133
Figure 12.9	Causal network of 'site type predictor' BBN	134
Figure 12.10	Example of mapping between two probability tables. Adapted from Huang and Darwiche (1996).	137

List of Tables

Table 3.1	Top 20 least stressed sites ordered by mean spring/autumn ranking. Rankings based on mean predicted stress, predicted using <i>best matching cluster</i> method.	20
Table 3.2	Top 20 least stressed sites ordered by mean spring/autumn ranking. Rankings based on mean predicted stresses, predicted using <i>weighted mean</i> method.	21
Table 3.3	Top 20 least stressed sites ordered by mean of <i>best matching cluster</i> and <i>weighted mean</i> rankings.	21
Table 3.4	Top 20 most stressed sites ordered by mean spring/autumn ranking. Rankings based on mean predicted stresses, predicted using <i>best matching cluster</i> method.	22
Table 3.5	Top 20 most stressed sites ordered by mean spring/autumn ranking. Rankings based on mean predicted stresses, predicted using <i>weighted mean</i> method.	23
Table 3.6	Top 20 most stressed sites ordered by mean of <i>best matching cluster</i> and <i>weighted mean</i> rankings.	23
Table 3.7	Top 20 most stressed sites in spring, with stress identified as predicted stress value greater than two standard deviations from mean value for all samples.	25
Table 3.8	Top 20 most stressed sites in autumn, with stress identified as predicted stress value greater than two standard deviations from mean value for all samples.	25
Table 3.9	Top 20 sites identified as having stresses in both seasons ranked by mean number of stresses	26
Table 4.1	The eighty-two BMWP taxa	35
Table 4.2	Taxa contributing to the eleven composite BMWP families	35
Table 4.3	Taxa contributing to BMWP taxa Chironomidae and Oligochaeta	36
Table 4.4	Summary of biological samples by year and agency for spring	36
Table 4.5	Summary of biological samples by year and agency for autumn	36
Table 4.6	Summary of sampling sites by agency and season	36
Table 5.1	List of 13 environmental variables used in previous project	37
Table 5.2	Definition of chemical variables	39
Table 5.3	Number of samples containing each chemical determinand for three-year sample period as minimum required sample population is increased	43
Table 6.1	Completeness of stresses database	46
Table 6.2	Abundance of most commonly perceived stresses for 2003	47
Table 6.3	Geology categories	49
Table 6.4	Land cover categories	49
Table 6.5	Land risk categories	50
Table 7.1	Extent of spatial variation in ranking score across 121 gauging stations	54

	for each month, averaged over 30 years	
Table 7.2	Actual and predicted ranking scores and conditions for August record for typical gauging station	56
Table 7.3	Summary for 121 stations with averages etc taken over 30 years	57
Table 7.4	Actual (vertical) and predicted (horizontal) flow condition by month	58
Table 7.5	Actual (vertical) and predicted (horizontal) flow condition averaged over all months	58
Table 7.6	Taxa for which probability of absence was generally greater in 'wet' conditions than 'dry', for riffle sites in the spring	60
Table 7.7	Taxa for which probability of absence was generally less in 'wet' conditions than 'dry', for riffle sites in the spring	61
Table 8.1	Ranking of variables in original spring RPDS model	68
Table 8.2	Ranking of variables in 'clone' spring RPDS model	69
Table 8.3	Distribution of samples between five equally sized bands for each environmental variable, ordered according to their ranking in Table 8.2	70
Table 8.4	Results for original RPDS data using equal percentile bandings	71
Table 8.5	Results for RPDS 'clone' using equal percentile bandings	72
Table 8.6	Three rankings (based on mean of pair-wise correlation, MI tests and MI values based on MIR-max model using only macroinvertebrate taxa) and their mean	75
Table 8.7	Results for environmental training variables reduction tests using alkalinity	76
Table 8.8	Results for environmental training variables reduction tests using calcareous geology	77
Table 8.9	Mean entropy and standard deviation for each cluster and each variable based on analysis of macroinvertebrate variables, plus MI for whole model	79
Table 8.10	Results for revised five environmental variables using alkalinity (left) and calcareous geology (right)	80
Table 8.11	Distribution of samples by region and year for spring MIR-max model based on average alkalinity	81
Table 8.12	Distribution of samples by region and year for autumn MIR-max model based on average alkalinity	81
Table 8.13	Distribution of sites for spring and autumn MIR-max models based on average alkalinity	81
Table 8.14	Distribution of samples by region and year for spring MIR-max model based on calcareous geology	82
Table 8.15	Distribution of samples by region and year for autumn MIR-max model based on calcareous geology	82
Table 8.16	Distribution of sites for spring and autumn MIR-max models based on calcareous geology	82
Table 9.1	List of requirements for additional functionality to RPDS	95
Table 11.1	Distribution of matched biological and chemical sample data by region and year available for revised BBN model, spring	112
Table 11.2	Distribution of matched biological and chemical sample data by region and year available for revised BBN model, autumn	112
Table 11.3	Distribution of sites for matched biological and chemical sample data for revised BBN	112
Table 11.4	Results of dependent tests on RPBBN 2.0 against RPBBN 1.0 expressed in terms of Spearman rank correlation coefficients between predicted and recorded values of chemical variables	114
Table 11.5	Performance characteristics of RPBBN 2.0	114
Table 11.6	States of variables in Tables 10.4 and 10.5 together with their prior probabilities	115
Table 11.7	Description of independent tests	115
Table 11.8	Results of independent tests on RPBBN 1.0 expressed in terms of Spearman rank correlation coefficients between predicted and recorded values of chemical variables	116
Table 11.9	List of requirements for additional functionality to RPBBN	117
Table 12.1	Suggested chemical bands and distribution of database samples	119
Table 12.2	Chemical bands used in revised 'Southern' RPBBN and distribution of samples	120
Table 12.3	Different types of groups in templates	129
Table 12.4	Comparison of times taken by HUGIN and lazy algorithms to initialize and perform update on RPBBN-P2, RPBBN-A and RPBBN-S models.	137

1 Introduction

This project's work packages relate to the following three original objectives:

1. To help the Environment Agency's work for the Water Framework Directive (WFD).
2. To develop diagnostics and modelling for river basin planning and programmes of measures.
3. To develop an ecological quality classification scheme (if the regulatory authorities were unable to develop RIVPACS (Moss *et al.*, 1987)).

The third objective provided an insurance against any problems that would have prevented the Environment Agency from developing RIVPACS (River Invertebrate Prediction and Classification System) to meet the needs of the WFD river invertebrate classification of ecological quality. Previous work (Walley *et al.*, 1998) showed that artificial intelligence (AI) technology was able to deliver a good classification based on average score per taxon (ASPT) and number of taxa that could be related back to predictions of reference, as required by the WFD. Had it not been possible to develop RIVPACS, work packages to develop an AI-based classification would have been implemented. This explains the original title of the project (*Development of an Integrated Classification System for Rivers and Lakes*). This option was not taken up. RIVPACS was used as a basis for classification largely because it had proved adequate and its statistical basis (including its shortcomings) was already well understood by ecologists in the regulatory agencies.

Two early activities in our project were designed to help with the first objective, as follows:

- Use the original project database to determine chemical thresholds for General Quality Assessment (GQA) monitoring sites at good ecological status to help establish chemical standards to protect invertebrates.
- Determine potential reference sites using River Pressure Diagnostic System (RPDS) software to help identify targets for temporal trajectories.

The work undertaken on these two aims is described in Sections 2 and 3 respectively.

The main core of the project addressed the second aim, in which the work undertaken in previous projects (Walley *et al.*, 2002 and Walley *et al.*, 1998 respectively) was extended. Two computer systems based on artificial intelligence techniques had been developed in the most recent of those projects: RPDS (River Pressure Diagnostic System, formerly River Pollution Diagnostic System, based on pattern recognition), and RPBBN (River Pressure Bayesian Belief Network, formerly River Pollution Bayesian Belief Network, based on plausible reasoning). Both systems were based on a dataset of biological (macroinvertebrate) and environmental samples taken in the spring and autumn of 1995 at over 6,000 sites in England and Wales, for which there was matched chemical data for 3,600 sites. The main objectives of this phase of the project were as follows:

- Substantially extend the dataset on which the systems were based.
- Revise and test the data models on which the systems were based.
- Extend the functionality of the two systems and combine them into an 'integrated' system.

The dataset was extended in the following ways:

- Biological, chemical and environmental data was included from the national river surveys in 1995-2004.
- Data for Scotland was included as well as for England and Wales.
- Additional biological, chemical and environmental data was included.

Data from other surveys was considered but data from the River Habitat Survey (RHS) was rejected because it was not standardised, and experience from the previous project suggested that only 15 per cent commonality would be achieved with the rest of the data. Data from the Countryside Survey 2000 (Centre for Ecology and Hydrology, Dorset) included matched RHS data (invertebrate samples and some limited chemistry), but this was rejected too. Although the data may have been useful (the Countryside Survey has a different distribution of sites, usually headwaters, which complement the GQA survey sites which are predominantly at the downstream end of streams and rivers), information about the location of sites would have to be withheld (a condition of using the data) thus requiring the data to be excluded from display on maps, for example. Although data for lakes was not expected to be included, the project was anticipated to act as a feasibility study for lakes.

Construction of the extended dataset from national survey data became a major task and delayed the rest of the project. The initial approach was to add to the biological and chemical datasets for 1995 and 2000 incrementally with annual summaries supplied by the Environment Agency. However, the extent of inconsistencies and incompatibility between the annual increments made this approach unviable. The inconsistencies probably arose from the use of different queries to the main Environment Agency database or from 'snapshots' taken as the database evolved over time. Consequently, new requests were made for both biological and chemical data for the entire period 1993-2004 to ensure a dataset of reliable integrity. Construction of the biological dataset thereafter was an extensive and time-consuming task, largely because tools within the Environment Agency's biological database to extract family data had not been developed; the process is described in Section 4.

Additional biological variables in the original project scope included diatom, macrophyte and phytoplankton data. Fish data was considered but rejected as being too different. The inclusion of other data was investigated but delays in the availability of diatom and macrophyte data, limitations in the coverage of phytoplankton data, and delays in collating the macroinvertebrate data led to a decision to restrict biological data to macroinvertebrates only.

Environmental parameters associated with the biological samples and chemical sample data are described in Section 5. Matching the biological and chemical samples required both sets of spatial coordinates to be validated (because the biological and chemical data are sampled at different sites). Procedures for this and subsequent matching are also described.

Section 6 describes the updated stress data from 2003 (the original RPDS dataset included stress data from 1995), as well as additional land cover and soil risk data derived from geographical information systems (GIS) that may be indicative of pressures.

Flow data was included as statistics based on long-term averages (from LowFlows 2000) and as the relative flow condition before the biological samples were collected (based on time series data). This data could only be included for sites in England and Wales (not Scotland). The procedure developed for estimating flow condition and subsequent evaluation of its ecological significance in terms of macroinvertebrates are described in Section 7.

The databases constructed here are among the largest of their kind, and have already been used for a variety of purposes outside this project. However, because of delays caused by some of the tasks required to construct the databases, the goal of producing an 'integrated system' had to be abandoned in favour of redeveloping the two separate systems.

Section 8 documents the revised MIR-max models produced from the new datasets, while preliminary evaluation of the new models and modifications to RPDS software are covered in Sections 9 and 10 respectively. Section 11 deals with the revised BBN model and some preliminary testing, while modifications to RPBBN software are presented in Section 12. The project is summarised in Section 13.

2 Determination of chemical thresholds for GQA sites

Introduction

The work described in this section was undertaken early in the project and was designed to inform the development of regulatory standards for protecting macroinvertebrates in rivers. The initial stages of this work began as part of a previous project and culminated in the production of software called the Rivers Intercalibration Project or R.I.P. The purpose of the R.I.P was to establish the relationship between biological quality measures and chemical parameters, and to translate biological quality standards into chemical standards.

This work fed directly into the design of a work package in which the biological community rather than a biological quality score was used as the basis for obtaining chemical standards.

In this section, R.I.P software is described and the outcomes of the work discussed to provide some background. Following this, the work undertaken in this project is described, namely the development of the Thresholder software.

Chemical thresholds

For WFD, it is necessary to define chemical quality standards for rivers, that is, chemical concentrations that support the ecological quality status. This would permit rivers to be monitored for conformance to chemical quality. The problem with defining quality is that it is a subjective concept and the concentrations that correspond to 'good' or 'poor' quality are somewhat meaningless unless the impact of the chemicals on some other aspect of the river, such as the biology, is used as a yardstick.

Biological quality is a key aspect of river management and the WFD, and quality standards already exist based on ecological quality indices (EQIs). Therefore, using biological quality standards in the definition of chemical thresholds would appear to be a sensible approach, providing the necessary yardstick and automatically calibrating the two sets of standards.

Rivers Intercalibration Project (R.I.P.)

The purpose of R.I.P. was to study the relationship between biological and chemical parameters and use this information to define a set of chemical quality standards. This process in turn would provide crisper and more easily measurable targets for the analysis of rivers based on chemical concentrations, rather than the more subjective criteria of 'quality'. At the outset of the project, the intention was to derive a set of suggested chemical standards from analysis of the project database, which contained the combined biological and chemical sample data from the 1995 River Survey of England and Wales. However, it quickly became apparent from the initial analysis that the relationships between biology and chemistry were insufficiently clear-cut to allow a set of standards to be easily identified. This meant that some degree of expert interpretation would be required when selecting values to be used as standards. As a result, the outputs of the project changed from producing a set of suggested standards

to producing software that would help experts to define a set of standards, by analysing the 1995 River Survey database.

The software produced was known as R.I.P. At its core, R.I.P. is simply a scatter-plotting program that allows the relationships between variables to be shown quickly. R.I.P. also incorporates tools to define potential standard boundary values. The definition of the ranges of values that define the standard simply involves drawing a box on the scatter plot with the sides of the box on the x-axis and y-axis defining the upper and lower bounds. For example, when drawing the quality bounds for ecological quality index for ASPT, 'good' quality lies in the range of 0.9 to 1.0, therefore the sides of the box on this axis occur at these values. The report information provided by R.I.P. includes bounds drawn on the x- and y-axes, correlation coefficient for the whole scatter plot and sets of points that fall within the bounding boxes.

R.I.P. software

R.I.P. software allows the user to load delimited text files, which are files in which all the values are stored as text and each variable/field is separated by a special delimiting character. Any pair of continuous variables within the file can be plotted against each other.

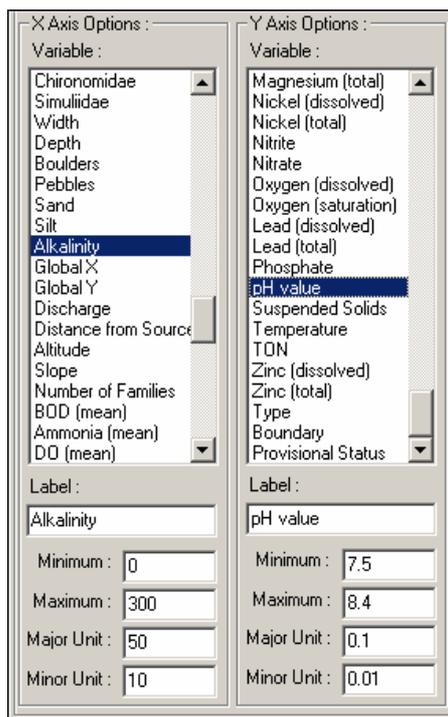


Figure 2.1 Plot options panel of R.I.P. software.

Figure 2.1 shows the plot options panel of R.I.P., which lists the variables in the currently loaded file and provides a facility to modify the parameters of the plotted axis. Once two variables have been selected they are plotted on the plot panel, along with the number of samples and the r correlation coefficient, as shown in Figure 2.2.

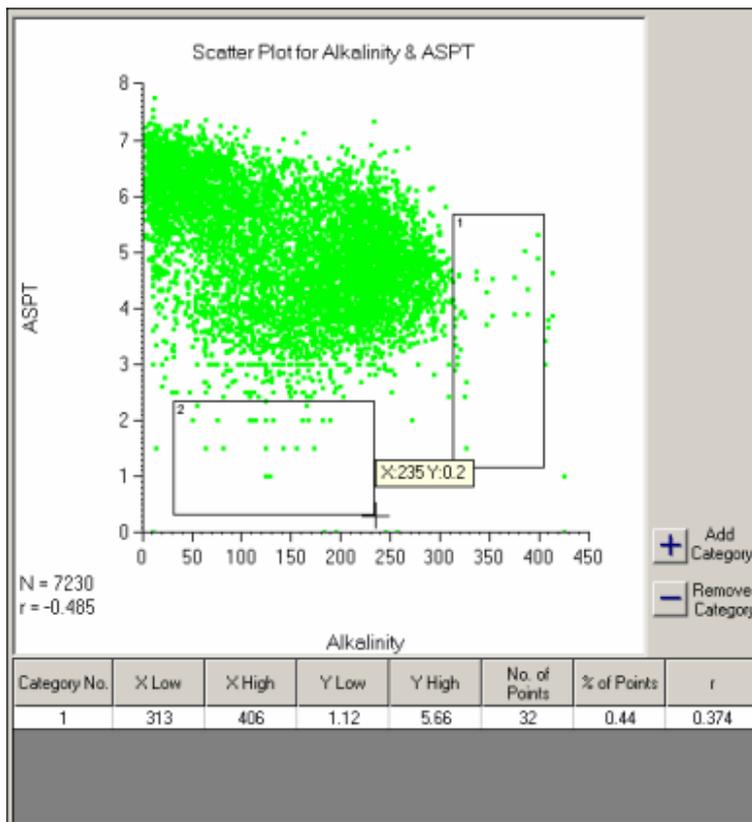


Figure 2.2 Plot panel showing scatter plot of ASPT against alkalinity and two user-defined range boxes.

This plot can then be used to define boundary values for the different quality classes. This is achieved by the user choosing to 'add category', then drawing a box on the plot. As the user draws the category box, information on the high and low values it encompasses on both axes, the number and percentage of points and the r correlation coefficient for these points are automatically updated in the report beneath the plot (see Figure 2.2). R.I.P. also allows the plotted points to be coloured by a 'quality' variable, the variable must be categorical and have seven states or less. The purpose of this feature is to allow the spread of samples to be easily identified in terms of quality.

Once a set of categories/boundaries have been defined, R.I.P. has the option to print the plot and the accompanying category information to provide a hard copy of the work.

Outcomes of R.I.P.

The main outcome of R.I.P. was that it showed that the relationships between biological quality measure and chemical pressures were not as strong as anticipated. When plots of quality measures against chemicals were produced, instead of the desired diagonal line of points indicating a strong, clear and identifiable relationship, the majority were simply clouds of points, like those shown in Figure 2.2. This meant that defining chemical standards by this method would be difficult because it would inevitably involve some element of expert opinion so any decisions could be contentious, leaving standards open to criticism and potentially difficult to enforce. Ultimately, R.I.P. showed that using natural breaks in the distribution of ecological quality and chemical variables was not a practical solution and a more systematic and objective approach was required.

This approach was based principally on data analysis. The key differences from the R.I.P approach were firstly, the biological component of the analysis would be multivariate, with a community rather than a single quality value, and secondly,

selection of the threshold values would be based on an automated scan of the available data rather than manual selection. The source of the data was the pollution sensitivities of taxa database, which was used as the source of the 'Pollution Sensitivities of Taxa' website.

'Pollution Sensitivities of Taxa' website

The data produced in the 1995 River Survey of England and Wales contained a wealth of information on the macroinvertebrate fauna and chemical and environmental parameters at more than 6,000 sampling sites. As part of a previous project (Walley *et al.*, 2002), some preliminary analysis of the pollution sensitivity of the taxa was undertaken and the results published in 1999 on the 'Pollution Sensitivities of Taxa' website, shown in Figure 2.3, at <http://www.soc.staffs.ac.uk/research/groups/cies2/>. The focus was primarily on the ranges of chemical and environmental parameters at which the taxa were found with different levels of abundance. The website displayed the results for each of the chemicals for every taxon for which there was data. However, this work was a spin-off from the development of the original RPBBN, and no further activity was undertaken once the project was complete.

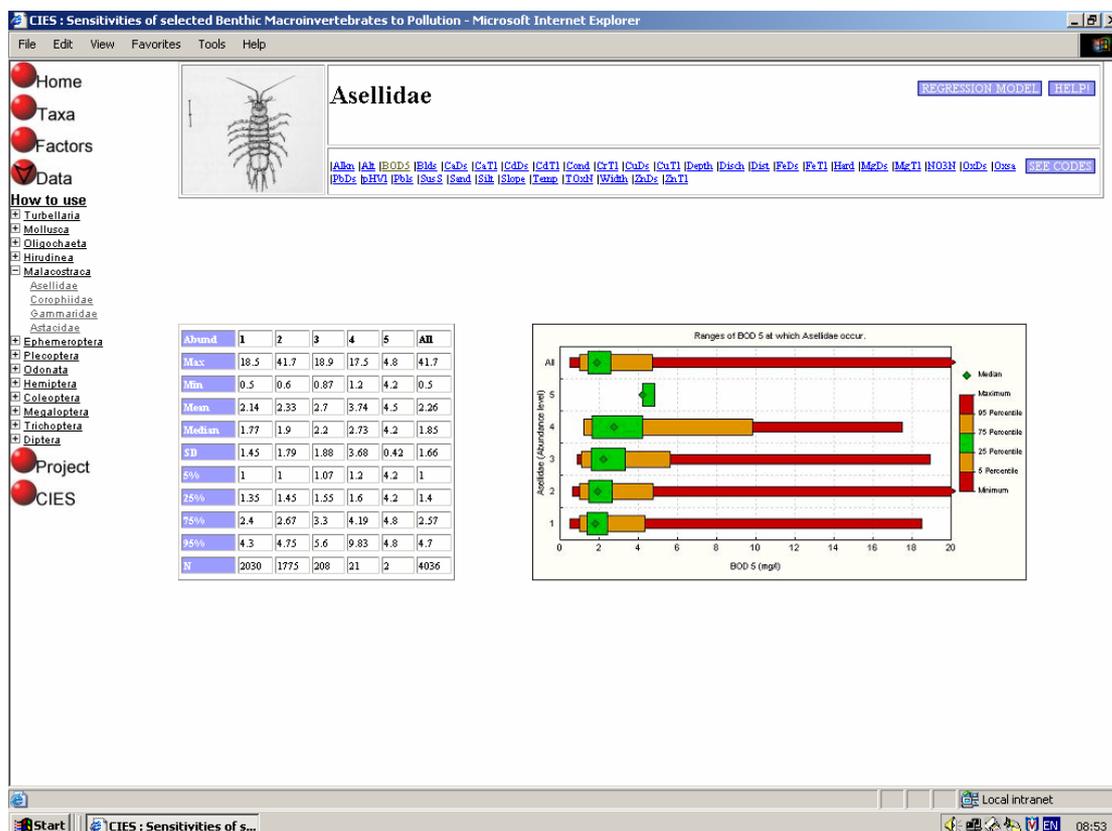


Figure 2.3. Pollution Sensitivities of Taxa website. Horizontal bars show distribution of taxon (Asellidae in this example) against concentration of chemical (BOD in this example). Green, orange and red indicate range of quartiles, fifth and 95th percentiles and range of all data. Top bar represents presence regardless of abundance and four bars below represent distributions for different log₁₀ abundance categories of the taxon.

Subsequent feedback from the Environment Agency and others suggested the need for further development. The main requirement was for a more flexible and powerful method of manipulating and presenting the underlying data, in particular the ability to

derive results for combinations of taxa and/or chemicals. Such functionality would enable predictions to be obtained of the abundance levels of taxa or threshold concentrations for chemical parameters at a sample site. This would help us to determine chemical standards to protect ecological quality by defining the chemical concentrations that support the invertebrate fauna predicted by RIVPACS at all GQA sites. The specific objectives were as follows:

- Determine the target invertebrate fauna at each site using RIVPACS predictions of the probabilities of occurrence and abundances of families.
- Scan the 'Pollution Sensitivities of Taxa' website to determine the concentrations (range and percentile statistics) of each chemical that supports the predicted taxa at their predicted abundances.
- Pool the concentration ranges (and statistics) for each chemical at each site. These are the thresholds for the site. Include an indication of the reliability of the thresholds, such as the number or proportion of predicted taxa on which they are based.

The original intention was to achieve these aims by redeveloping the 'Pollution Sensitivities of Taxa' website. However, difficulties with this approach led to the development of stand-alone software call 'Thresholder', the steps to which are described below.

First phase of development

Data

The 'MChm95' database contained individual tables comprising the environmental, chemical and macroinvertebrate data for each sample site in England and Wales. A composite data table 'CreatureStats' was created from the original data to record the maximum, minimum, mean, medium, standard deviation, 5th, 25th, 75th and 95th percentiles of each environmental or chemical parameter for every taxon at each of the sampled abundance levels¹. For example, the database contained each of the statistical values of chloride for Gammaridae at the abundance levels at which it was recorded. The table contained 11,500 records for 44 environmental and chemical parameters and 76 taxa at up to six different abundance levels. The data in the 'CreatureStats' table provided the basis for the 'Pollution Sensitivities of Taxa' website, with individual pages representing a record in the table.

Design of the original website

When the 'Pollution Sensitivities of Taxa' website was originally constructed, consideration was given to providing a means to manipulate elementary information in the 'CreatureStats' table and so enable the user to extract data for different combinations of taxa or chemical and environmental parameters. However, this was abandoned because of practical problems. The key difficulty was representation of all the possible combinations of variables on the website. If the website was static, this would involve storing hundreds of thousands of pages to cover all possible combinations. On the other hand, if the site was dynamic the pages would not be stored but created in response to a query using some form of Common Gateway Interface (CGI) programming². At the time, the effort required to develop either option was considered to be unjustified.

¹ Abundance levels were log₁₀ categories 0= 0, 1-9 = 1, 10-99 = 2, 100-999 = 3, 1,000-9999 = 4 and 10,000+ = 5.

² A loose definition of CGI is a program whose output is a web page. This means that the web page can include dynamic information extracted from a database or produced by some algorithm.

Design of the new website

In designing the new 'Pollution Sensitivities of Taxa' website, the first decision was whether it would be static or dynamic. Many pages would need to be stored for a static website, and it would be difficult and time-consuming to modify. A dynamic website, on the other hand, would offer a readily expandable way to present the threshold data. However, development of a dynamic site would have its own problems. Enabling CGI programs to run on the web server would increase security risk and crashes of the CGI programs could potentially disable the entire web server³. In addition, CGI programs would be more difficult to test and debug than normal applications: firstly, because the inputs and outputs would be transmitted between web browser and server across a network; and secondly, because the code would be run remotely on the server rather than on the computer on which the system would be developed. The potential for problems when creating the final web-based system (such as serious delays in development or major difficulties associated with maintenance of the server) would be increased by the need for several cycles of re-development and modification.

Creating a prototype system as a Windows application

An obvious way to reduce the likelihood of a serious problem was to undertake the initial development of the final system as a Windows[®] application, and subsequently rewrite it as a web application. An additional benefit of this strategy was less development time for the prototype because the skills to develop the Windows[®] application as a Visual Basic[®] project were readily available. This approach was adopted and the resulting 'Thresholder' software was developed as a prototype. Thresholder received taxon abundance values as inputs and produced an output report of thresholds (maximum and minimum values) for chemical and environmental variables based on samples whose taxonomic parameters match those of the input values.

The system was designed to identify the ranges of chemical concentrations or values at which a predefined set of taxa at specified abundance levels were found to exist in the MChem95 database. Options for the input taxa data were designed to be biological sample data, RIVPACS community predictions or values entered manually for single taxa or assemblages. The screenshot in Figure 2.4 shows abundance values loaded from a file containing part of the 1995 River Survey data, while Figure 2.5 shows the user manually inputting abundance values. A report of the output is shown in Figure 2.6.

Once the abundance values for the taxa were set, the database was queried to extract the range for each chemical at which all taxa at the specified abundance level had been recorded. The ranges produced in the report (see Figure 2.6) represented the union of values produced for the individual taxa, rather than values based on the combination representing the entire assemblage. For example the maximum-minimum range for alkalinity based on Asellidae at abundance level 3 and Gammaridae at abundance level 1, was based on the maximum-minimum range for Asellidae (all records where Asellidae occur at abundance level 3) joined with the range for Gammaridae (all records at level 1), rather than the maximum-minimum range based on the combination of Asellidae at level 3 and Gammaridae at level 1. This approach was adopted to maximise the range of queries for which the system could produce a result.

³ This was a particular problem because Staffordshire University hosts the website and its web-based resources are extremely important to its function. Therefore, security breaches and/or crashes on the web server would represent a threat to the day-to-day workings of the university and its business.

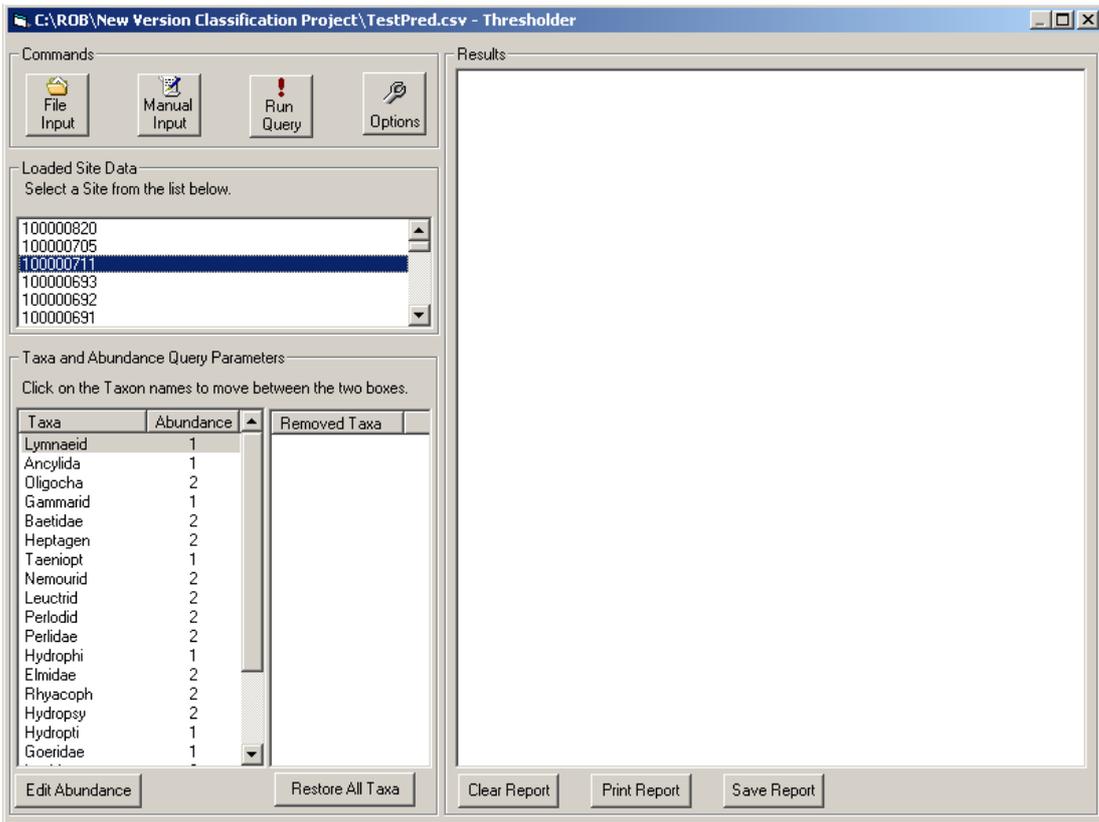


Figure 2.4 Screenshot of Thresholder software with abundance values for taxa loaded from external file containing part of 1995 River Survey data.

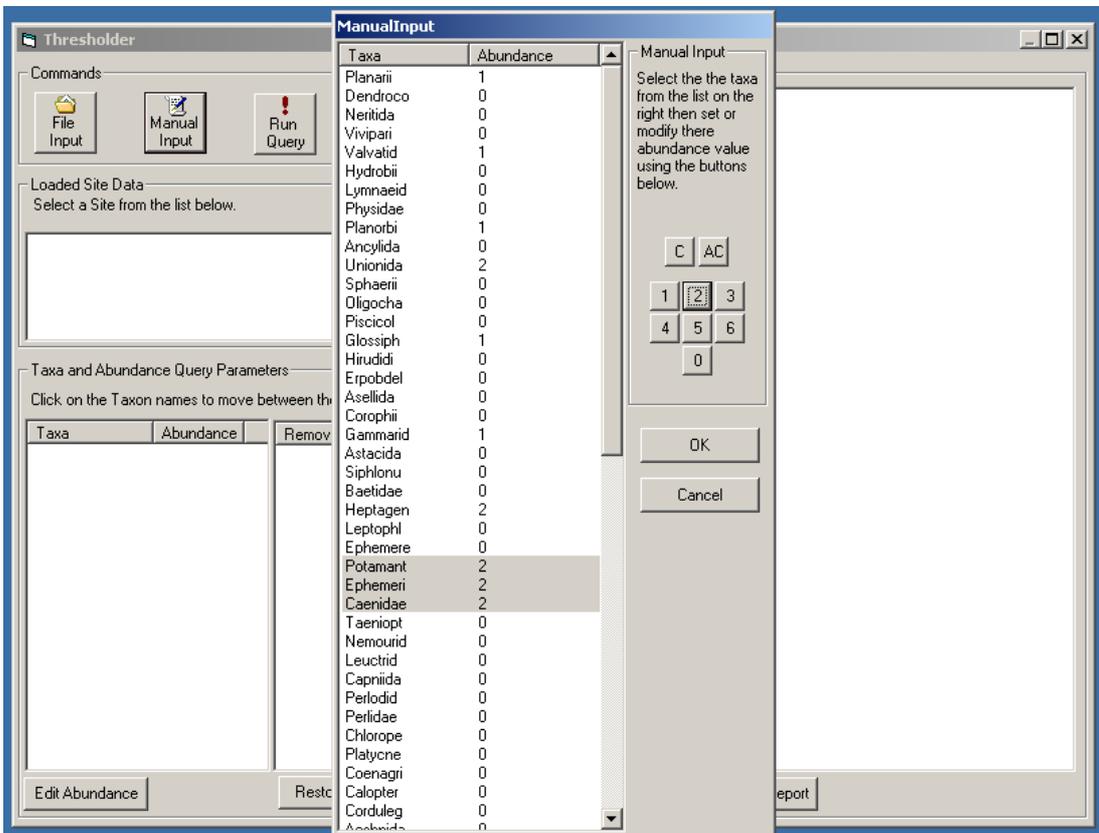


Figure 2.5 Screenshot of Thresholder software showing user defining abundance values manually.

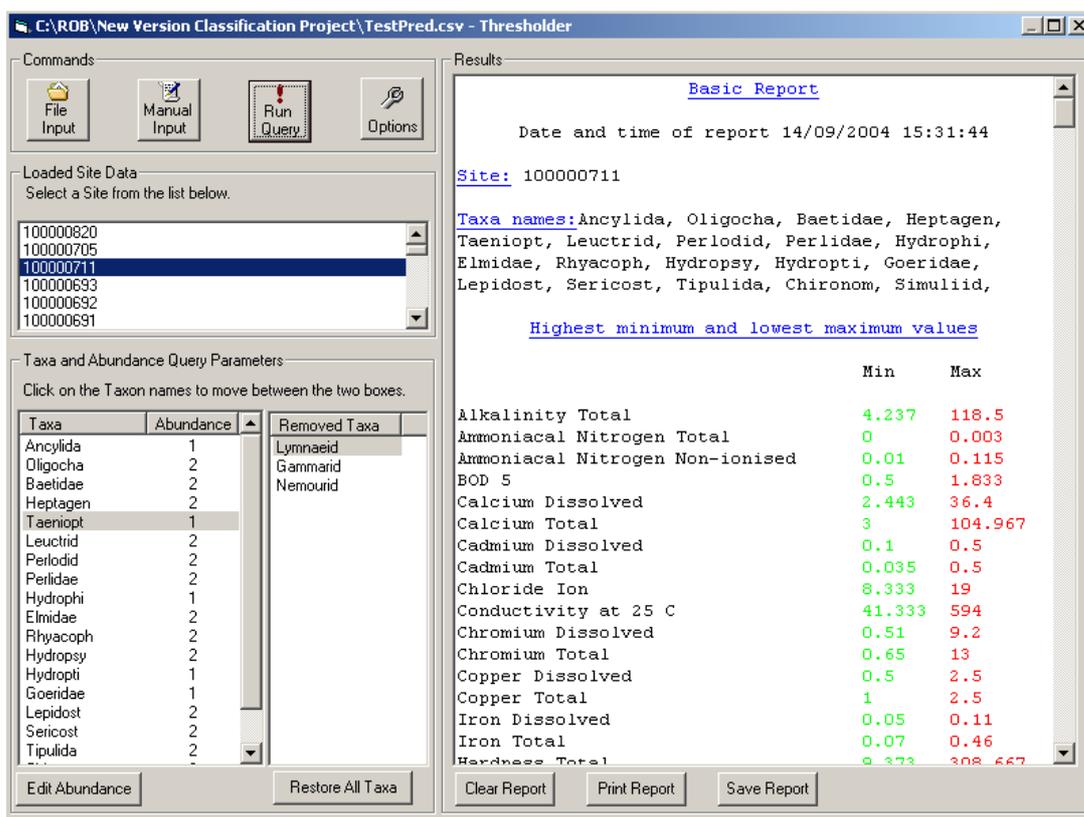


Figure 2.6 Screenshot of Thresholder software showing report detailing results of a query.

Second phase of development

Following delivery of the prototype to the Environment Agency for evaluation, the software was refined in response to users' suggestions. The main improvements requested were:

- A facility to import RIVPACS results and predictions to derive the ranges of expected chemical concentrations.
- Inclusion of seasonal as well as annual concentration values (a feature missing in the original 'Pollution Sensitivities of Taxa' database) and inclusion of information on the 90th and 10th percentiles (to permit compatibility with regulatory thresholds and environmental standards).

These required the creation of a new database, as well as modification to the prototype.

Importing RIVPACS data

Production of the RIVPACS target invertebrate fauna predictions⁴ involved firstly converting the 1995 GQA sample data into the RIVPACS input file format. Two of the RIVPACS prediction functions were used for this: *Option 1. BMWP families and BMWP indices*, which produced predictions of the probability of capture of each of the BMWP (biological monitoring working party) families and; *Option 2. Abundance for all families*

⁴ Only BWMP-family taxa can be used as the basis for chemical threshold predictions because these are the only taxa for which there is data in the 'pollution sensitivities of taxa' database.

and abundance index, which produced predictions of the abundance of the BMWP families.

A method of combining probabilistic information on the likelihood of capture with predicted abundance values needed to be defined. Following consultation with the Environment Agency, it was proposed that prediction of target fauna would only include those taxa for which the probability of capture exceeded a predefined level. A probability of 50 per cent was proposed as the suggested cut-off level, given that it had already been used by the Environment Agency to predict abundance-related LIFE index for CAMS (catchment abstraction management system) environmental weighting.

Modification to the software

The 'Pollution Sensitivities of Taxa' database was modified to incorporate seasonal values and additional percentile values. In order to scan the database, the Thresholder software was modified to incorporate:

- A facility for the user to change aspects of the scanning process by simply altering parameters in the software.
- A prototype Windows® application that could be redeveloped as a web application for inclusion in the revised 'Pollution Sensitivities of Taxa' website.

By providing the user with the ability to change parameters of the scanning process such as composition of biological community and abundances of taxa (see next section), it was hoped that further modifications to the 'Pollution Sensitivities of Taxa' database or adjustments to the threshold capture probability would only require adjustments to the options provided in the user interface, and not changes to the underlying application code. In the longer term, developing the system to accommodate different types of scans at this stage would reduce the amount of development required later for publication on the internet.

To calculate chemical thresholds for the input target invertebrate fauna, the Thresholder software performed two processes. Firstly, it scanned the 'Pollution Sensitivities of Taxa' database and extracted the records corresponding to each taxon at the abundance level specified in the input vector. Secondly, it compared the maximum/minimum and percentile range data of each chemical in all the extracted records. The aim of this process was to identify the range of concentrations/values of chemicals common to each taxon. This involved finding the lowest maximum/upper percentile value and the highest minimum/lower percentile value. Once complete, these values were then output to provide an idea of the range of chemical concentrations within which all taxa in the input vector would be likely to be found.

Using the Thresholder software

As the software initialises, it loads the 'Pollution Sensitivities of Taxa' database which contains the statistics for each taxon at different levels of abundance. The main user interface window is displayed in Figure 2.7 and has four separate regions:

- *Commands* – containing buttons that start the main functions of the application. These functions are *File Input* and *Manual Input* for inputting invertebrate fauna data, *Run Query* and *Run Batch Query*⁵ for initiating a

⁵ The terms 'query' and 'scan' are used synonymously in the discussion of the Thresholder software.

scan of the database and *Options* for modifying the parameters for the scan and report.

- *Loaded Sample Data* – displays all the sample data loaded from an input file.
- *Sample Query Parameters* – displays abundance values of the invertebrate fauna of the current sample.
- *Report* – displays the results of the query.

The user is then given the option of manually entering abundance values of invertebrate fauna or loading sample data from a file. The file formats accepted at the time of writing are:

- Comma-delimited (.csv)⁶.
- Report files produced by the *Abundance for all families and abundance index* prediction option of RIVPACS. (If a RIVPACS abundance prediction file is loaded the user is also presented with the option to load a file that contains corresponding probabilities of capture, produced by the *BMWP families and BMWP indices* prediction option of RIVPACS.)

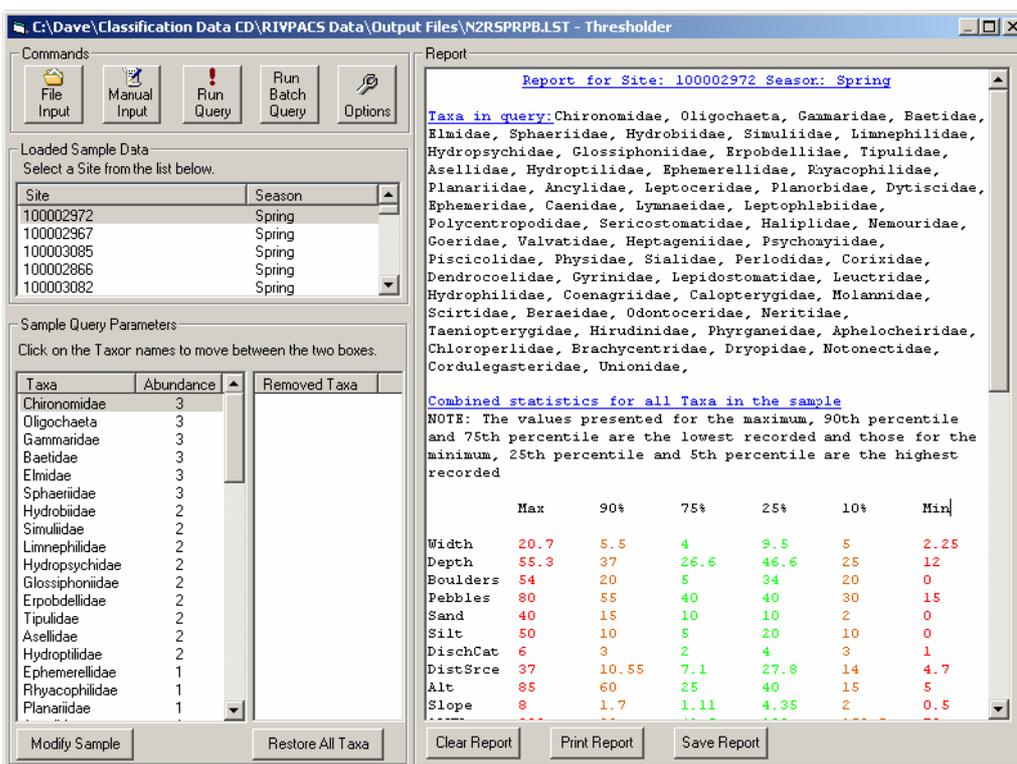


Figure 2.7 Screenshot of main Thresholder window, showing abundance values for invertebrate fauna of selected sample, and corresponding query report.

If sample data is loaded from a file, all samples in the file are displayed in the *Loaded Sample Data* section of the main screen. By simply clicking on the appropriate sample identifier, the user can select its invertebrate fauna data as the basis of a query. At this point, the *Sample Query Parameters* list box is populated with invertebrate data from the relevant sample.

⁶ Comma delimited files need to meet file format requirements before they can be loaded by the system.

The *Sample Query Parameters* section of the main window provides the user with the option of removing individual taxa from the query or modifying their values (Figure 2.8). Once satisfied with the abundance values and the invertebrate taxa to be used in the query, the *Run Query* button can be clicked to provide the user with a report on the expected maximum, minimum and percentile statistics for a range of chemical parameters. Alternatively, once a sample input file has been loaded, the *Run Batch Query* can be selected to produce a comma-delimited (.csv) file containing the predicted chemical statistics for each of the samples in the input file.

The user is also provided with several options that change the operation of the Thresholder software. These options are split into two sections: the first modifies the format of the on-screen report, while the second modifies aspects of the query (number of samples on which statistics for individual taxa need to be based for consideration in the final results, and the cut-off level for the probability of capture, if applicable) (Figure 2.9).

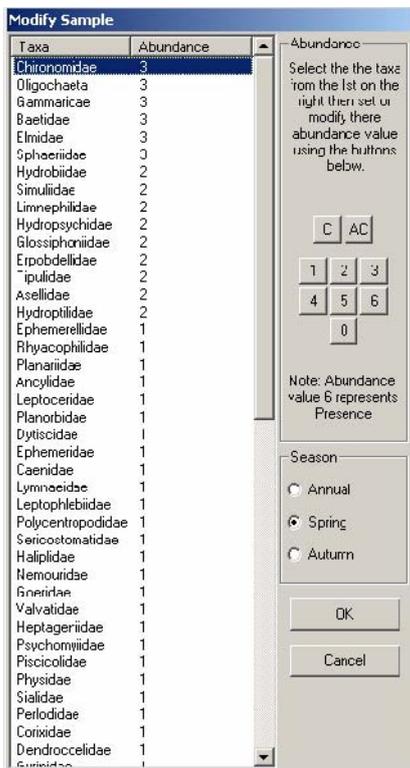


Figure 2.8 Screenshot of Modify Sample/Manual Input dialogue box.

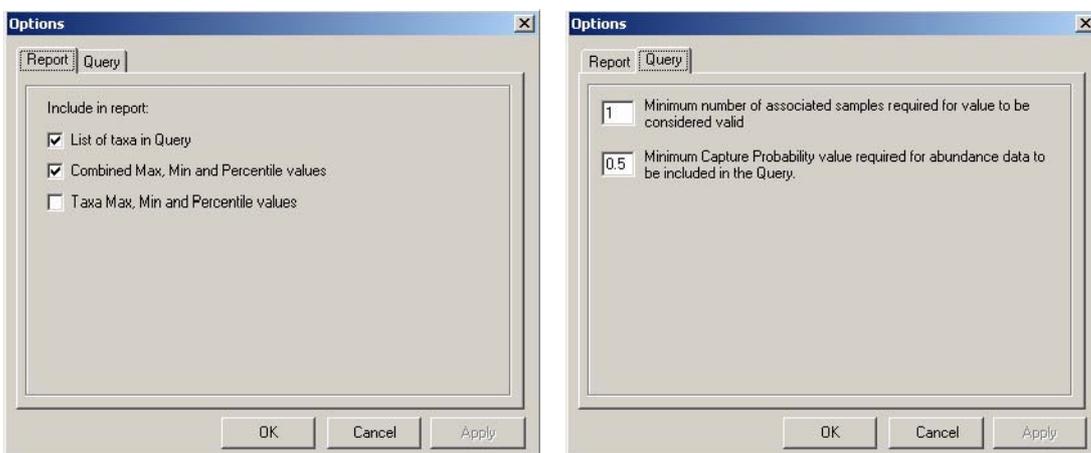


Figure 2.9 Screenshot of Thresholder Options dialogue box, showing report and query parameters available to user.

Derivation of chemical threshold values

The chemical threshold results based on the target invertebrate fauna predicted by RIVPACS were delivered to the Environment Agency in February 2005. The Thresholder software produced two sets of results: the first based on the statistics of all taxa regardless of the size of the sample from which the statistics were derived; and the second based solely on taxa whose statistics were based on 36 records or more⁷, the larger sample size ensuring greater reliability.

The probabilities of capture and predicted abundances of the target invertebrate fauna were loaded directly from the RIVPACS output files. The *Run Batch Query* command was then used to produce the final results, which include maximum-minimum, 90th-10th and 75th-25th percentile ranges for each the target invertebrate communities.

The validity of using either set of results as the sole basis to derive chemical threshold values is questionable. It might be argued that they are based on poorly substantiated statistics, or on only a subset of the total amount of information potentially available. The results are probably best used as a guide for experts with the ability to interpret the data to arrive at better-informed decisions on the chemical thresholds finally imposed.

Future work

Because of the demands of the other phases of the project, it was not possible to implement Thresholder as a web application, and this must be left for future work.

⁷ The figure of 36 records was set in consultation with the Environment Agency, as this was roughly the minimum number of records required to obtain a reasonable estimate of the 90th percentile value.

3 Determine potential reference sites

Introduction

Objectives

Following the derivation of chemical thresholds to support invertebrate fauna, two further aims of the project related to the task of identifying unstressed river sites which could be used to define reference conditions. Specific objectives were as follows:

- Search the RPDS database for sites which could be used for this purpose. Such sites could then be used to identify clusters in the data model to act as target clusters when improvements in quality are tracked over time.
- Assist in the screening and confirmation of the current set of reference sites used by RIVPACS.

The work undertaken to meet these aims is described below.

Generate list of proposed reference sites from RPDS database

Introduction

The River Pressure Diagnostic System (RPDS) is based on a cluster model created using the Mutual Information and Regression maximisation (MIR-max) algorithm (a review of the algorithm can be found in Section 8 of this report). This algorithm is an unsupervised clustering method, meaning that the criteria used to perform the clustering and assign samples to clusters are not imposed beforehand. Although this method has the advantages of being objective and being able to operate without any further information, the final model does not include any indication of the set of criteria used to construct the clusters or the types of pattern that they represent. Hence the clusters in the model need to be examined, their 'type' identified and a name or label assigned to them. This examination process is known as 'interpreting' the model.

One of the main outstanding tasks to enable full use to be made of RPDS was to interpret the model in terms of river quality. Such an analysis would enable an estimate of the quality of a new input sample to be made by simply identifying the quality band of the cluster that it was classified to. In an operational setting, this would provide a quick and sophisticated multivariate method of assigning a quality rating to a new sample. This method would also provide a means of evaluating the quality ratings assigned to samples using other methods. This would allow RPDS to be used as a screening tool to provide further assurance for quality assessments.

Interpreting the RPDS model

The initial objective was to interpret the model to identify clusters that were of reference quality. This involved the following two-stage process.

- An initial screening - in which existing information and techniques for assessing quality were used to identify candidate reference samples and clusters.
- A detailed analysis of candidate samples/clusters - in which samples and clusters received closer inspection by expert biologists to assess the appropriateness of the initial classification.

To perform the initial screening, each of the samples in the original RPDS training data set (the N2R database) was assessed using four different criteria where possible:

1. A biological GQA grading of A or B.
2. A chemical GQA grading of A.
3. No perceived stresses at the site.
4. All recorded chemical concentrations within 90th percentile range.

The biological and stress data needed for this was available for all 12,078 samples in the data set. However, because of the paucity of matched chemical and biological sites, the necessary chemical data was available for only 7,230 of the samples, so only about 60 per cent of the samples could be assessed using all four criteria - the remainder were screened using the biological GQA and stress data.

The initial screening identified 1,234 candidate reference samples. Of these, 526 had chemical data and so met all four criteria. The remaining 708 samples met the biological and stresses criteria only, in the absence of chemical data.

The next stage of the process was a more detailed analysis of the candidate reference sites by expert biologists. The candidate samples were extracted from the database and were sent to John Murray-Bligh (Environment Agency) and John Davy-Bowker (CEH) for this.

Summarise status of current reference sites according to RPDS

Introduction

Implementation of the Water Framework Directive (WFD) requires information on a range of sites that are of high quality and are representative of the different types of rivers found in the UK. The purpose of collecting this information is to develop a 'reference' model capable of predicting, for any site in the UK, the composition of the macroinvertebrate community to be expected if environmental conditions were in WFD reference condition (subject to minimal human pressure or pristine, see REFCOND, 2003). The quality of a particular site can then be expressed as the extent to which the community observed at the site deviates from that expected in reference state.

The collection of a large amount of 'reference' sample data had already been undertaken as part of the development of RIVPACS. However, not all reference samples used for RIVPACS were of WFD reference quality: they were the best available and for some types of stream there were no examples where human influences were minimal or absent. Also, whereas RIVPACS sites were chosen according to their chemical water quality, a much broader range of pressures were encompassed by WFD including flow, morphology and alien species. As the developers of RIVPACS, the Centre for Ecology and Hydrology (CEH, Dorset) screened the sample data (Davy-Bowker *et al.*, 2007). However to complement this process it was proposed that RIVPACS reference data be classified by the River

Pollution and Diagnostic Systems (RPDS) to indicate any potential stresses on these reference sites.

Diagnosis using the River Pressure Diagnostic System (RPDS)

The principal purpose of the River Pollution and Diagnostic System (RPDS) is to diagnose stresses that may be affecting a sample taken from a site, by classifying the sample solely on the basis of its macroinvertebrate community and a few environmental parameters. The diagnosis is based on the degree of similarity between the new sample and the cluster(s) that it is assigned to. The RPDS database includes a wide variety of data and so RPDS is able to provide predictions for a range of environmental parameters, including chemical concentrations and perceived stresses.

Making predictions of environmental parameters using RPDS classifications

Like any model that classifies rivers into types, RPDS essentially partitions a continuous distribution, with clusters representing areas of the distribution. The mean values calculated from all the samples assigned to the cluster are taken to represent the 'centre point' of the cluster. This centre point represents the set of values that is most indicative of the cluster as a whole. Membership of the cluster is defined by similarity to the centre point. New samples can be classified according to their position in relation to the centre points of different clusters. These classification membership values reveal a great deal about the sample, its position in relation to the clusters and within the model itself. However, to generate a predicted value for a variable from a set of classification results requires all information in the results to be condensed into a single value, which creates problems.

The following two methods were used to make predictions of environmental parameters using the results of the classification of RIVPACS sites using RPDS.

1. Predictions based on the values of environmental parameters within the cluster that is the best match. This is the simplest method of deriving predictions for a sample, because the values are taken directly from the *best matching cluster*.
2. The *weighted mean* method, which uses the membership values and the mean values of environmental variables for the cluster to calculate a predicted value. The method is defined by the general formula:

$$\bar{X}_w = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$$

where:

N = number of clusters

w_i = weightings (in this case similarity values = membership values: in RPDS's case the change in mutual information).

x_i = cluster values (in this case the mean of the chemical concentration for the samples in that cluster).

Only clusters with positive values of the membership function were used, ensuring that the weighted mean was based on clusters that were genuinely similar to the sample. The weightings are negative if adding the sample to a cluster changes the mutual information (MI) value for the model for the worse.

The key problem with the *best matching cluster* method is that it takes no account of the other cluster membership values. Even if a sample has several cluster membership values that are almost identical, only the highest value will be considered and the information held within the similarity that it shares with the others clusters is lost. The *weighted mean* method, on the other hand, suffers from ambiguity introduced when values are combined and the difficulty of interpreting fractional predicted values for discrete parameters (see Figure 3.1). On the other hand, the weighted mean is relatively easy to implement as an initial form of analysis and provides valuable information on the result of classification.

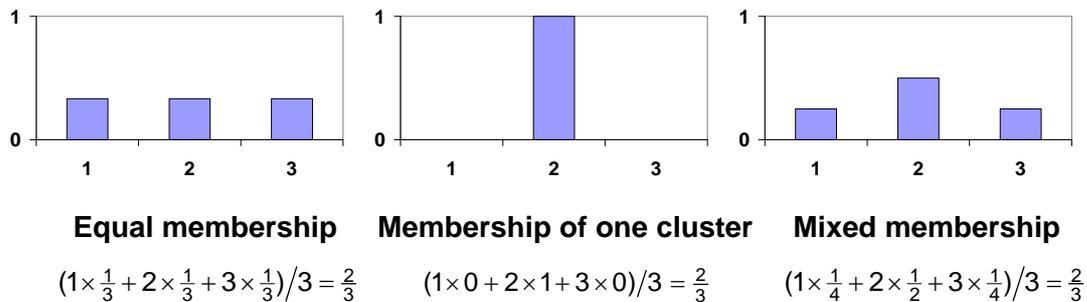


Figure 3.1 Three different membership value distributions that could give the same weighted mean result.

Ranking of sites using seasonal and mean total stress values

Three different types of analysis were performed on the *best matching cluster* and *weighted mean* results. The sites were ranked using seasonal and mean total stress values, analysis of the correlation of ranking results and by the identification of predicted values that may indicate the impact of a stress.

The total stress value is the total of all the predicted stress values. Predictions for individual stresses will not be whole numbers because they are derived from the stress values for the cluster, which are the mean of the sample values in that cluster. The total of all the stress values will therefore also tend to be fractional.

Values predicted by the *best matching cluster* and *weighted mean* methods were 'noisy' because of the use of mean cluster values. This was the case even for *best matching cluster* predictions, where even the high quality potential 'reference' clusters had low values for some stresses. This limited the information that could be derived from the results without undertaking more detailed analyses. Because of this, the initial results simply took the form of a ranking of sites in terms of total stresses. These results assumed that the distribution of noise was reasonably consistent and the more 'significant' predicted stress values would influence the rankings. Tables 3.1-3.3 and 3.4-3.6 show the 20 least and most stressed sites respectively, according to the criteria: *best matching cluster*, *weighted mean* predictions and the mean of the two.

Table 3.1 Top 20 least stressed sites ordered by mean spring/autumn ranking. Rankings based on mean predicted stress, predicted using *best matching cluster* method.

SiteID	Site	River	Spring Rank	Autumn Rank	Mean Rank
5619	Bredwardine	Wye	3	6.5	4.75
CL04	Mainholm Ford	Ayr	31.5	6.5	19
3409	Gainford	Tees	31.5	16	23.75
SEPA_N45	Syre	Naver/Mudale/Meadie	31.5	16	23.75
20301601	Glarryford Bridge	Main/Clogh River	4.5	69.5	37
4203	Moffat	Annan	48.5	29.5	39
3407	Barnard Castle	Tees	31.5	48	39.75
3509	Bardon Mill	South Tyne	31.5	48	39.75
3703	Laighlands	Teith	31.5	48	39.75
4211	Brydekir	Annan	31.5	48	39.75
20100701	Killymore Bridge	Owenkillew River/Broughderg Water	31.5	48	39.75
SW07	Mether-uny-Mill Bridge	Gweek River	80.5	22	51.25
HI10	Moy Bridge	Conon/Bran	123.5	6.5	65
SEPA_E07	Haugh of Kercock	Tay/Dochart/Fillan/Cononish	123.5	6.5	65
3101	Langdale End	Derwent	4.5	127.5	66
SEPA_N54	Rhisalach	Kirkaig/Ledbeg	58	86.5	72.25
SEPA_N55	d/s Loch Borrulan	Ledmore/Loin Duibh	58	86.5	72.25
6105	Nuns Bridge, Thetford	Thet	19.5	127.5	73.5
4905	Kingledores	Tweed	72.5	86.5	79.5
3603	Middleton	Wansbeck	65	101	83

Table 3.2 Top 20 least stressed sites ordered by mean spring/autumn ranking. Rankings based on mean predicted stresses, predicted using *weighted mean* method.

SiteID	Site	River	Spring Rank	Autumn Rank	Mean Rank
1207	Fawler	Evenlode	9	4	6.5
23601401	Tullyreagh Cross	Colebrooke River	20	12	16
313	Flowerpot	Exe	2	37	19.5
8521	Middle Bere	Bere Stream	26	18	22
1807	Ketford	Leadon	57	1	29
3704	Blackdub	Teith	27	31	29
SEPA_N01	Brouster	Shetland: Upper Loch of Brouster	50	39	44.5
23601501	Wattle Bridge	Finn River	67	26	46.5
1901	Perry Farm	Perry	78	19	48.5
601	Patney	Avon	60	47	53.5
2103	Colston Bassett	Smite	77	34	55.5
6801	Grange Wood	Middlemarsh Stream	21	90	55.5
6005	Temple Balsall	Blythe	103	10	56.5
1081	Carter's Lodge	Hammer's Pond Tributary	35	81	58
1301	Wotton	Tilling Bourne	79	38	58.5
NI_3	B84 Road Bridge	Owenreagh River	73	55	64
8309	Whitehouse Farm Ford	Bure	113	17	65
2005	Field	Blithe	72	65	68.5
3144	Newgate Foot	Long Gill	53	86	69.5
2905	Ouse Bridge	Derwent	133	8	70.5

Table 3.3 Top 20 least stressed sites ordered by mean of *best matching cluster* and *weighted mean* rankings.

SiteID	Site	River	Total Mean Rank
203	Oathill Farm	Axe	101.125
23601501	Wattle Bridge	Finn River	105.125
5619	Bredwardine	Wye	131.625
8521	Middle Bere	Bere Stream	134.375
5383	Bratley	Bratley Water	134.875
20100701	Killymore Bridge	Owenkillew River/Broughderg Water	137.125
5713	Crickhowell	Usk	143.5
7122	King's Farm	Moors/Crane	151.25
20101101	Ballynahatty	Drumragh River/Ballynahatty Water/O	157.875
3101	Langdale End	Derwent	165
ST02	Isle Of Bicton	Severn	167.625
3509	Bardon Mill	South Tyne	174.625
5864	Mordiford	Lugg	178.25
20301601	Glarryford Bridge	Main/Clogh River	179.75
6801	Grange Wood	Middlemarsh Stream	182
SEPA_N60	Poolewe	Ewe/Kinlochewe River/Abhainn Bruach	184.25
3407	Barnard Castle	Tees	186.875
6840	Gasper	Unnamed	188
605	Bulford	Avon	193.25
3704	Blackdub	Teith	194

Table 3.4 Top 20 most stressed sites ordered by mean spring/autumn ranking. Rankings based on mean predicted stresses using *best matching cluster* method.

SiteID	Site	River	Spring Rank	Autumn Rank	Mean Rank
2621	Earlham	Yare/Blackwater	834.5	835	834.75
6921	Runnymede	Thames/Isis	819	801	810
3317	Acaster Malbis	Ouse/Ure	809.5	801	805.25
1203	Evenlode	Evenlode	787.5	821.5	804.5
20301801	Rock Bridge	Kells Water	797	811.5	804.25
20200501	Corick Bridge	Roe	785	816.5	800.75
3111	Thorganby	Derwent	744.5	832.5	788.5
SEPA_N07	Stackhoull	Unst: Burn of Mailand/Caldback	760.5	811.5	786
1605	Pont Gogoyan	Teifi	797	773	785
1611	Llechryd	Teifi	797	773	785
4983	Chesterfield Ford	Whiteadder Water	797	773	785
20100001	Donnelly's Bridge	Foyle/Mourne/Strule/Camowen	797	773	785
8421	Lower Brook	Test	739	829.5	784.25
1409	Meadgate	Lee	744.5	820	782.25
2619	North of Barford	Yare/Blackwater	822	735.5	778.75
2303	d/s Hedingham	Colne	832	710.5	771.25
1209	Cassington	Evenlode	830	710.5	770.25
23600301	Killynoogan	Termon River	812.5	727	769.75
1405	Panshanger	Mimram	734.5	797	765.75
1307	Tilford	Wey	779.5	749.5	764.5

Table 3.5 Top 20 most stressed sites ordered by mean spring/autumn ranking. Rankings based on mean predicted stresses, using *weighted mean* method.

SiteID	Site	River	Spring Rank	Autumn Rank	Mean Rank
SEPA_W02	d/s Cattadale	Islay:Laggan/Barr	806	830	818
20100801	Cloughery Bridge	Glenelly River	830	791	810.5
HI08	Strathan	Arkaig/Dessarry	818	798	808
SEPA_N27	Leachd Thuilm	Skye:Brittle	803	812	807.5
SO03	Nr. Southwick House	Southwick Burn/Boreland Burn	808	789	798.5
4901	Fingland	Tweed	816	776	796
SEPA_N21	Sourin	Rousay:Suso Burn	777	809	793
NI_33	Kilnasaggart Bridge	Kilnasaggart	732	835	783.5
4203	Moffat	Annan	799	767	783
HI07	Shiel Bridge	Shiel	748	817	782.5
NH07	u/s Balderhead Reservoir	Balder	774	787	780.5
NI_2	Broughderg Bridge	Owenkillew River/Broughderg Water	793	761	777
SEPA_W07	Monyquil	Arran: Machrie Water	781	773	777
ST01	Llandinam	Severn	747	806	776.5
7305	Ariundle Oakwood NNR	Strontian	710	831	770.5
SEPA_E01	u/s Auchinner Bridge	Water of Ruchill	826	707	766.5
2901	Grange-in-Borrowdale	Derwent	706	825	765.5
SEPA_W34	A8003 Bridge	Ruel	744	783	763.5
SEPA_N36	Meavaig	North Harris: Meavaig River	770	755	762.5
20500101	Glynn	Glynn River/Glenoe Water	784	732	758

Table 3.6 Top 20 most stressed sites ordered by mean of *best matching cluster* and *weighted mean* rankings.

SiteID	Site	River	Total Mean Rank
20200501	Corick Bridge	Roe	765.125
3111	Thorganby	Derwent	680.5
SEPA_N43	Strathmore	Hope	673.25
6693	u/s Dowles Manor	Dowles Brook	661.75
4305	Fionn-Abhainn	Fionn Abhainn	658
3007	Braystones	Ehen/Liza	654
NH07	u/s Balderhead Reservoir	Balder	654
FO03	Pitcruvie Castle	Boghall Burn/Keil Burn	653
2513	Marston Trussel	Welland	645.875
23601201	Drumkeenagh	Black River	643.125
NI_28	Gortin Bridge	Gortin Water	641.125
6847	Farrington	Unnamed	635.875
4105	d/s Barr	Stinchar	632.875
3301	Keld	Swale	625.5
SEPA_W12	u/s Gaodhail	Mull:Forsa	625.5
107	Brocton	Camel	624.375
3003	u/s Keekle	Ehen/Liza	623.125
23600501	Ederny	Kesh River/Glendurragh River	623.125
20601701	Forkhill Lower Bridge	Kilcurry River/Forkhill River	621.25
3005	d/s Keekle	Ehen/Liza	618.625

The sets of results for the *best matching cluster* and *weighted mean* were markedly different, for no site appeared in the lists of 20 least stressed or 20 most stressed sites based on both criteria. This result was disappointing and indicated that the inclusion of the additional clusters used by the *weighted mean* method had swamped the values of the *best matching cluster*.

Another notable feature of the results was that although the sites identified by the *weighted mean* method were all from different rivers, some rivers re-occurred in the *best matching cluster* results. For example, the Tees re-occurs in the 20 least stressed sites and the Yare/Blackwater, Evenlode and Teifi all re-occur in the 20 most stressed sites.

The combined mean results generally contain sites that occur in either of the other sets of results. In particular the list of 20 least stressed sites contains seven sites (Bewardine, Langdale End, Bardon Mill, Glarryford Bridge, Barnard Castle, Grange Wood and Blackdub) which appear in both the other two lists. The division of the sites between the lists is uneven, with the five that appear first in the list coming from the *best matching cluster* results and the remaining two coming from the *weighted mean* results. In the list of 20 most stressed sites, Corick Bridge and Thorganby appear from the *best matching cluster* results; and u/s Balderhead Reservoir from the *weighted mean* results. These results are reassuring because they indicate that there was some commonality between the sets.

Analysis of correlation of ranking results

Additional analysis was undertaken using a rank correlation test. Rankings were compared between spring and autumn seasons for the same method and between methods in the same season. Seasonal correlations for each method were low, with r values of 0.27 and 0.33 for the spring and autumn rankings for the *best matching cluster* and *weighted mean* methods respectively. However, the correlations between the rankings for the same season were much higher, with r values of 0.65 and 0.66 for spring and autumn respectively. This indicated that the main source of disparity between the individual sets of rankings was the season rather than the method used.

Identification of predicted values that may indicate the impact of a stress

As mentioned earlier, background 'noise' in the *weighted mean* results made it difficult to identify predicted stress values that might have indicated the impact of a stress. This was a key aspect of the study, because although the ranking of sites was useful, it did not indicate the factors that made one site worse than another.

One possible method for identifying the stresses impacting on a site was to identify the predicted stress values that deviated sufficiently from the baseline noise to be considered significantly different. This was done by calculating the mean value for each of the stresses based on all the samples and identifying values that were different by more than two standard deviations from the mean. All values nearer than two standard deviations from the mean were considered to be primarily the product of noise and were assigned a 'zero' and all those above were considered to be a 'true' stress and assigned a 'one'. As before, the number of stresses for a site was totalled to provide a ranking, but in addition, the stresses responsible for that ranking could also be identified.

The 20 most stressed sites in both spring and autumn are shown in Tables 3.7 and 3.8 respectively. Like the best matching cluster and weighted mean predictions, the seasonal results show few similarities: in fact only one site appears in both lists (Sordale on the Thurso River). As before, there were discernable differences between

the spring and autumn results. A further list of sites considered to be stressed in both seasons was produced by matching sites with stresses in both spring and autumn. Table 3.9 lists the 20 most stressed sites according to this list.

Table 3.7 Top 20 most stressed sites in spring, with stress identified as predicted stress value greater than two standard deviations from mean value for all samples.

SiteID	Site	River	Total Stresses
20303101	Airport Bridge	Crumlin River	9
4111	Ballantrae	Stinchar	6
5305	Millyford Bridge	Highland Water	6
2211	Monk's Bridge	Dove	5
5607	Marlbrook	Lugg	5
20303401	Caledon Bridge	Blackwater	5
207	Whitford Bridge	Axe	4
229	Gammons Hill	Yarty	4
411	Great Torrington Town Mills	Torridge	4
1603	Tregaron Bog	Teifi	4
2513	Marston Trussel	Welland	4
3381	Hubberholme	Wharfe	4
3704	Blackdub	Teith	4
4807	Sordale	Thurso	4
5695	Folly Farm	Arrow	4
6801	Grange Wood	Middlemarsh Stream	4
9703	Glassoch Bridge	Bladnoch	4
20100801	Cloughery Bridge	Glenelly River	4
20301601	Glarryford Bridge	Main/Clogh River	4
23600301	Killynoogan	Termon River	4

Table 3.8 Top 20 most stressed sites in autumn, with stress identified as predicted stress value greater than two standard deviations from mean value for all samples.

SiteID	Site	River	Total Stresses
NI_33	Kilnasaggart Bridge	Kilnasaggart	8
2507	Banthorpe Lodge	Glen	6
20400201	Iderown Bridge	Dervock River/Stracam River/Dougher	6
2607	Worthing	Wensum	5
2721	Ribchester Bridge	Ribble/Gayle Beck	5
2911	Workington	Derwent	5
3007	Braystones	Ehen/Liza	5
NI_24	Carrols Bridge	Crew Burn	5
SEPA_N09	Bouster	Yell: Easter Burn of Bouster	5
SEPA_W02	d/s Cattadale	Islay: Laggan/Barr	5
SEPA_W06	Drochaid Bheag	Islay: Duich/Torra	5
2505	Little Bytham	Glen	4
4807	Sordale	Thurso	4
5003	Bidwell Farm	Otter	4
5005	Monkton	Otter	4
5381	Vereley	Ober Water	4
5856	Leominster	Main Ditch	4
6201	u/s Brackley	Unnamed	4
6242	Nine Wells	Nine Wells Spring	4
6413	Tootle Bridge	Brue	4

Table 3.9 Top 20 sites identified as having stresses in both seasons ranked by mean number of stresses.

SiteID	Site	River	Autumn Stresses	Spring Stresses	Mean Stresses
20303101	Airport Bridge	Crumlin River	1	9	5
NI_33	Kilnasaggart Bridge	Kilnasaggart	8	1	4.5
4111	Ballantrae	Stinchar	2	6	4
2211	Monk's Bridge	Dove	3	5	4
4807	Sordale	Thurso	4	4	4
20400201	Iderown Bridge	Dervock River/Stracam River/Dougher	6	2	4
20303401	Caledon Bridge	Blackwater	2	5	3.5
2513	Marston Trussel	Welland	3	4	3.5
2721	Ribchester Bridge	Ribble/Gayle Beck	5	2	3.5
3007	Braystones	Ehen/Liza	5	2	3.5
SEPA_W02	d/s Cattadale	Islay: Laggan/Barr	5	2	3.5
23600301	Killynoogan	Termon River	2	4	3
TA04	u/s Tay Confluence	Braan	2	4	3
5401	Hadman's Place	Beult	3	3	3
9711	Spittal	Bladnoch	3	3	3
20302101	Dundermot Bridge	Killagan Water	3	3	3
SEPA_W04	u/s Duich confluence	Islay: Laggan/Barr	3	3	3
5381	Vereley	Ober Water	4	2	3
6413	Tootle Bridge	Brue	4	2	3
7311	Anaheilt	Strontian	4	2	3

Feedback

It was pointed out (by John Davy-Bowker, CEH) that many of the most stressed sites (Tables 3.4-3.6, 3.7-3.9) were from Scotland or Northern Ireland. The likely reason for this was that RPDS was based on data from England and Wales, which generally has a less harsh climate and hence richer biological communities. Although the Scottish Environmental Protection Agency (SEPA) and EANI reference sites are of good quality for their environment, they appear to have been identified as of a lower quality compared to the reference sites for England and Wales, which are naturally richer. The inclusion of SEPA data in the current project should alleviate this problem if the exercise is repeated.

Conclusions

Preliminary lists of RIVPACS reference sites that may be stressed have been produced. Of the three lists, the one based on values that differ from the mean by more than two standard deviations is probably the most useful. This analysis includes an attempt to remove the impact of noise on the results and it identifies particular stresses that may affect the sample sites. The classification values on which this is based are the same as those used to produce the *best matching cluster* and *weighted mean* results, which provide alternative perspectives and additional information on classifications made by RPDS.

The main differences in the sets of predictions are caused by season rather than method of analysis. This provides reassurance that the classifications made by the seasonal RPDS models are consistent. However, the difference between the

predictions made by the seasonal models is of interest and requires further analysis to determine its implications.

Finally and most importantly, the results should only be considered as part of an interim analysis. Screening of the clusters in the RPDS model is vital to obtaining a more complete and consistent set of predictions and analyses. Attention should be focussed on the new RPDS model, which includes sites from Scotland as well as England and Wales, the data validation for which is described in the next sections of this report.

4 Construction of project database: biological data

Introduction

The biological database was fundamental to the project and ensuring the quality of its data was of paramount importance. As mentioned in Section 1, construction of the dataset was a major task which severely impacted the time available for the rest of the project. However, the dataset that was produced is one of the largest of its kind ever to be assembled. The stages required in its development are described below.

Standardising Environment Agency and SEPA data

The Environment Agency data came principally from the Environment Agency's BIOSYS database, which contains an extensive amount of macroinvertebrate data recorded at various taxonomic levels using the National Biodiversity Network (NBN) coding system and structured according to the NBN Data Model (see www.nbn.org.uk for details). Whilst the sophistication of this data model makes the database able to deal with different types of data recorded using different species dictionaries, it can make extraction and modification of the data awkward. The SEPA database, on the other hand, is at a much earlier stage of development. Unlike the Environment Agency, however, chemical and biological data is kept in the same database. The majority of macroinvertebrate data in SEPA's database is recorded using the latest versions of the modified Furse-Maitland Code and identified only to family level.

In order to standardise the biological data provided by the Environment Agency and SEPA for inclusion in a unified database, it was first necessary to obtain a match between the codes or names used to identify taxa in each dataset. The Environment Agency used the NBN (National Biodiversity Network) system to encode samples, and this has the advantage of enabling cross-referencing with other taxonomic checklists and codes through a 'Taxon Dictionary'. As a result, the sample data supplied by the Environment Agency had a 'SORT_CODE' field that contained the revised Furse-Maitland Code (Maitland, 1977)⁸ version 1.1 for each taxon. The SEPA sample data also contained the Maitland Code for each taxon; however the codes were from version 3.1 of the checklist published on 2 July 2003. An attempt to produce a match based on the checklists produced 777 unmatched codes and 332 mismatches where the codes were the same but either the taxon or spelling of the taxon's name differed.

To obtain a better match, the process focused initially on only the taxa that appeared in the SEPA sample data and names of the taxa were used as the match criteria instead of codes. The reason for matching by names was that they appeared to be modified less frequently than the codes and the problem of mismatches caused by the swapping of codes was removed. Modifications to the matching process led to all but 16 taxa being matched, and after consultation with the Environment Agency the problems with these were resolved.

⁸ Although referred to as the Maitland Code for historical reasons, taxonomic codes used by the Environment Agency and SEPA are revised versions of the original Maitland Code published in 1977. The revised code was originally published in 1989 and was developed by Mike Furse (Centre for Ecology and Hydrology, CEH), Ian McDonald (Thames Water Authority) and Bob Abel (Department of the Environment) and has been maintained by Mike Furse. A copy of the most recent version of the code is available from the CEH website (<http://science.ceh.ac.uk/subsites/eic/ddc/furselist/index.htm>).

To produce a more robust solution to matching Environment Agency and SEPA data, a Maitland Code look-up database was constructed. Following the principles of the NBN Taxon Dictionary, this permitted cross-referencing between different versions of the Maitland Code.

Duplicate sample data

Analysis of datasets from the Environment Agency and SEPA revealed they both contained duplicated sample data. In the majority of cases, duplicate data was simply a copy; that is, the site, sample, and abundance value for the particular taxon were identical. However, there were cases in which the site and sample information was the same but recorded abundance differed. Whilst the problems of copied records could be rectified by simply deleting all but one of the copies, the problem of the same samples having different abundance recorded could not because there was no way to identify which was the 'correct' value.

The reason for the duplicate data in both cases was simplification of the structure of BIOSYS data to supply all the data in one table. In the BIOSYS database, each sample could have several records in other tables associated with it. When the multiple table structure was collapsed into one, each record from the related table was assigned a copy of the sample data that it was associated with.

The copied records were caused by collapsing the relationship in the BIOSYS database between the B4W_REASONS and B4W_SAMPLES tables, so for each sample with several 'reasons for sampling' there was a corresponding number of copies of the sample data. The solution to this problem was simply to group all the records on all fields other than 'Reason', which effectively removed all the copied records.

The same samples having different abundance values was caused by collapsing the relationship between the B4W_SAMPLES and B4W_ANALYSIS_TAXA tables. Again, the information associated with the original sample was the same but the recorded abundance differed when more than one laboratory analysis was performed on a sample and the results of those analyses differed. In this case, there was no easy way to remove the duplicates because the dataset lacked the extra information associated with the individual analyses that would enable them to be distinguished. The only resolution to this problem was to have the biological data re-extracted from BIOSYS with the additional 'analysis' fields included. Once this data had been acquired, it was necessary to reduce the analysis result for each sample to just one. The 'initial primary laboratory analysis' results were used because this was the main analysis performed on the majority of samples (other analyses checked the quality of this main analysis for audit or analytical quality control).

Project taxonomic groupings list for Environment Agency database

The processes involved in the construction of the Environment Agency and SEPA databases were very different. The main difference was that the abundance values recorded for higher taxonomic groupings were inclusive in the SEPA database, that is, the abundance recorded for a family included the abundances of all the species that were recorded; in the Environment Agency's BIOSYS database the abundance values were exclusive, that is, the abundance recorded for a family did not include the abundances recorded for its constituent species. To extract the SEPA biological data, all that was needed was a list of required family identifiers. However, for the

Environment Agency, a list of all the required taxonomic groupings and all the taxonomic levels below that grouping against which data had been recorded had to be produced, validated and maintained. This task was the most time-consuming part of constructing the biological database and the difficulties encountered are discussed here.

To construct family level biological sample records, a method for combining species level data at a higher taxonomic level was required. As the database software used at CIES lacked the ability to perform recursive queries and thereby directly exploit parent links in the NBN Code System database, it was necessary to write bespoke software to do the job. The software was called NBNTree.

NBNTree software derived results based on predefined taxonomies, with the flexibility to amend or redefine taxonomical relationships. This meant that the project database could be updated without requiring further data requests for the Environment Agency when taxonomic errors were identified. One of the incidental benefits of this software was that, during validation of results produced using the Environment Agency's implementation of the NBN system, it was able to identify some missing relationships. This information was fed back to the Environment Agency and resulted in changes to the BIOSYS database. Although BIOSYS now has the capacity to do this, NBNTree provides much greater flexibility and could be useful to those without access to BIOSYS and those who wish to use different taxonomies.

NBNTree software loaded all the information about the taxa and their relationships in the form of a network that could be traversed to identify ancestor and descendent links. The software was used to construct a list of recorded taxa and their associated taxonomic grouping(s), which was used in the process of combining the data to 'family' records. The construction of this taxa list involved the following four stages.

1. Identification of all taxa recorded in the BIOSYS data (this initial list included macroinvertebrates, plants, algae and diatoms).
2. Traversal of the 'parent' relationships to identify all ancestors and thus all taxonomic groupings involved, including those that may not have been recorded explicitly, for example when a species had been recorded but there were no records containing that species' family.
3. Revision of the family list, to remove all unwanted families.
4. Traversal of the 'child' relationships to identify and label all taxa associated with a selected family/group.

Ensuring that the final list was accurate and well maintained was vital to the accurate reconstruction of sample data. Unfortunately, taxonomy is dynamic and not necessarily consistent, making this task difficult and time-consuming, not least because implementation of the NBN Code System by the Environment Agency included errors and inconsistencies. Producing the biological data for the 'intercalibration' exercise and for the revision of BMWP scores (Paisley *et al.*, 2007) required the process to be repeated several times.

Missing family links

Vital to the success of the process was the presence of all relevant parent links in the NBN Code System, otherwise it might not have been possible to identify the appropriate grouping for a taxon. During an inspection of one of the draft ancestor lists in stage two, it became apparent that a small number of taxa lacked parental links to a family grouping. There were two reasons for this: either the parent relationships stopped before family level or they skipped it. Identification of all taxa affected required a modification to NBNTree so that it checked for the occurrence of a particular taxonomic level amongst the ancestors of a taxon. This not only required additional

functionality but also for the 'taxon ranking' details in the data to be imported. Once the affected taxa were identified, they were sent to the Environment Agency to validate the broken links, to amend the NBN Code where necessary and to supply the missing taxonomic groupings for the project's list. The link between non-native *Pacificastacus* and native Astacidae was broken intentionally, to ensure that only native crayfish were included in the calculation of BMWP indices.

'Sub-family' groupings

Further problems with the identification of family groupings related to the NBN Code System for ranking of taxa and inconsistencies in the definition of parent links. In the BIOSYS database, some families shared the same 'taxon ranking' as tribes. As mentioned previously, the taxon ranking value was used to check whether a taxon had a 'family level parent' and this led to some taxa that only had links to a tribe being validated as having a family level parent. This was not too great a problem, but it highlighted the inconsistency in the recording of parental links, with some taxa in the same family being linked to tribe and not to the family and *vice versa*. The majority of these problems were associated with Chironomidae, and to achieve consistency it was decided to record all chironomids at tribe level, and allow them to be considered separately or combined to family level. It was therefore necessary to send the list of chironomids whose parent was Chironomidae to the Environment Agency, so that they could be assigned to tribes. This revised list was then incorporated into the master taxonomic groupings list.

Removal of unwanted sites

The data supplied from BIOSYS included data from a range of different categories of water bodies, some of which were artificial (dykes, ditches and canals) and were excluded from our study. The BIOSYS site data contained a field for 'water body type', which enabled the majority of these unwanted samples to be removed from the project database. The remaining samples were then analysed to ensure that this process had successfully identified all such samples. This was done by scrutinising the description field of the site. First, a list of all the words used in the description field along with their frequency of occurrence was created. This list, containing 7,357 words, was then checked for words that might indicate a category of water body not included in this study and these were recorded in a secondary list. All the site records that contained one or more of the words on this secondary list were then extracted. This dataset, containing 935 records, was checked to ascertain whether the sites were truly located in unwanted water body categories. This check was necessary to ensure that the words referred to the actual water body category at the site and were not simply part of the description, for example 'River X near Dyke Y' or 'River X at Canal Street.' This process revealed 548 sites which appeared to be located on one of the unwanted categories of water body. This list of 'problem' sites was sent to the Environment Agency for validation.

Validation by checking for errors

Errors in the data are of three basic types: value type errors (that is, values in the wrong format or of the wrong type, such as text instead of a number); 'implausible' erroneous values, where a value has been input incorrectly and is implausibly large or small; and 'plausible' erroneous values, where a value has been input incorrectly but the erroneous value is still plausible. Of these three types of errors the first two are the easiest to identify. Value type errors simply require a format/type check to be performed, which can usually be done by the database software itself; and 'implausible' errors which more often than not can be revealed using basic statistical techniques designed to identify outlying values.

Plausible errors are much more difficult to identify because they will not be identified as outlying. The only way they can be identified is by checking whether the potentially erroneous value is consistent with the pattern of values for other variables in the sample. This requires a more sophisticated method of analysis (using neural networks, for example – see Walley *et al.* 1998) able to generate expected patterns of values against which the sample data can be ,

Since the late 1990s when data was compiled for the preceding project (Walley *et al.* 2002), the centralisation and investment in the storage and handling of data in the Environment Agency has greatly improved the quality of data held. Simple analysis revealed no value type errors and only a small proportion could statistically be considered outliers with a reasonable degree of confidence. However, none of the outlying values identified were extreme and although high, were still plausible. This suggested that the Environment Agency had already performed some validation on the data.

Validation by using BMWP assessment data

Once the taxonomic groupings list had been constructed and checked, it was used to extract the raw sample data and combine the records to produce family level data. A list of families common to both the Environment Agency and SEPA was used to extract the equivalent data from the SEPA database and produce records in a common format in one unified database.

Validation of the biological data was discussed with the Environment Agency project management team, when it was suggested that expert opinion could be used to identify values that could be erroneous. It was agreed that the 50 highest abundance values for each taxon would be sent to the Environment Agency for evaluation. Values identified as potentially erroneous were then checked against lab records to confirm they were correct. Only a few such values were identified and queries about these were sent to the laboratories.

Because individual abundance values recorded for each taxon had been combined into a single record for each family, it was possible to confirm that records in the project database were a true representation of what was originally sampled. Both the SEPA and Environment Agency samples had BMWP-indices associated with them, and by calculating these indices from our constructed records, it was possible to perform some degree of validation. The assurances provided by this validation process differed for the SEPA and Environment Agency dataset and were dealt with as follows:

Validation of SEPA dataset

The majority of SEPA data lacked any sample BMWP values. Fortunately, a process of using existing and calculated BMWP assessment data to validate the sample data in the SEPA database had already been undertaken. It was therefore possible to compare our calculations against the SEPA assessment to provide some degree of validation. However, because both sets of results would be based on values from the database, the validation process would not be able to identify erroneous sample data, only discrepancies in the way that the data was handled.

An initial comparison of BMWP values produced over 30 differences. The variation was caused by a difference in the method used to check for the presence of a BMWP taxon. In this project abundance was important and so the criterion for 'presence' was a check for a valid abundance value - if an abundance value was invalid or missing the taxon was ignored. The criterion used by SEPA, on the other hand, was simply the

occurrence of a taxon identifier. Therefore, a difference indicated records that contained invalid abundance values.

Records with differences were analysed further to identify erroneous values and their causes. By far the biggest problem was that the abundance value was simply missing, although there were cases where errors were simply syntactical and the correct value could be recovered. Because this project required abundance values for every taxon in a sample, the 22 records with missing or unrecoverable values were removed from the dataset.

Validation of Environment Agency dataset

The Environment Agency supplied a complete list of BMWP values for all samples. However, these values were the result of calculations made by the BIOSYS database and so were based on the recorded sample data. Therefore, as with the SEPA validation process, the results only indicated how closely the BMWP sample constructed from the project database replicated that held on BIOSYS. Again, as with SEPA, the difference between sets of BMWP result would indicate only discrepancies in the way in which the original sample data was manipulated to derive the BMWP scores, and not errors in the recorded sample data.

Initial validation tests produced thousands of differences between the project and the Environment Agency's BMWP values, in one case up to 16,000. However, the source of most of these errors was traced to project values and omissions in the sets of taxa that were combined to form the BMWP 'families'. Once the missing taxa had been identified and included in the relevant composition lists, the number of differences dropped to just 353. The cause of the remaining differences was not so easy to detect. A small subset of samples was selected for closer analysis, which involved extracting the raw sample data, checking it and then calculating the BMWP values manually. When the results of the manual calculations agreed with those of the project results, the full set of 353 samples were sent to the Environment Agency to establish if BIOSYS was responsible for the differences.

The Environment Agency analysis identified the following reasons for the remaining discrepancies between BIOSYS and project results.

1. Changes in the classification of species considered to belong to BMWP families.
2. BIOSYS counting scores that were associated with genera.
3. Errors in data entry, where there had been a failure to record an abundance value.

Reasons 1 and 2 relate to the way in which the BMWP-score had been calculated, therefore any records that differed for these reasons were still valid and could remain in the dataset. However, the records that contained data entry errors had to be removed.

Completed biological database

The final stage in the construction of the biological database was to combine the validated Environment Agency and SEPA datasets. This was done by using a look-up table in which the names and codes used to identify taxa in the two datasets had been matched together.

The 82 BMWP taxa used are given in Table 4.1. Advances in taxonomy since Maitland (1977) have caused some taxa to be removed from several of the families listed in Table 4.1 and placed in new families. Table 4.2 lists the contributing taxa to eleven of

the BMWP taxa. Contributing taxa to two further BMWP taxa, Chironomidae and Oligochaeta, are listed in Table 4.3. Data assigned to Chironomidae that had not been allocated to one of the sub-families or tribes listed here contributed to the BMWP taxon Chironomidae.

The Environment Agency and SEPA biological sample data is summarised by region for spring and autumn for the years 1995 to 2004 in Tables 4.4 and 4.5 respectively. Table 4.6 summarises the sites from which the samples were taken for each season. Not all of this data could be used in the models. MIR-max models required environmental data in the input vector in addition to biological data, which reduced the number of samples that could be used. BBN models required matched chemical data as well, resulting in a further reduction. The steps required to produce the final datasets are described in the next three sections.

Although the work needed to produce the datasets was considerable, they are amongst the largest of their type ever to have been compiled, and have formed the basis for several other projects in addition to the development of the AI systems described in this report. They have been used for the WFD 'Intercalibration' exercise, revision of BMWP scores (Paisley *et al.*, 2007), research into ecological impacts (De Zwart *et al.*, 2008, Kapo *et al.*, 2008), chemical investigations (Comber and Georges, 2007) and for the MEM (Macro-Ecological Model) Project (Holzkämper *et al.*, 2008, Kumar *et al.*, 2008).

Table 4.1 The eighty-two BMWP taxa.

TRICLADA Planariidae Dendrocoelidae	PLECOPTERA Perlidae Chloroperlidae Taeniopterygidae Perlodidae Capniidae Leuctridae Nemouridae	COLEOPTERA Gyrinidae Scirtidae Dryopidae Elmidae Hydrophilidae Dytiscidae Haliplidae Hygrobiidae
MOLLUSCA Neritidae Viviparidae Ancyliidae Unionidae Hydrobiidae Sphaeriidae_Pea_mussels Lymnaeidae Planorbidae Valvatidae Physidae	ODONATA (Damsel flies) Calopterygidae Lestidae Platycnemididae Coenagriidae	MEGALOPTERA Sialidae
OLIGOCHAETA Oligochaeta	ODONATA (Dragon flies) Cordulegasteridae Aeshnidae Libellulidae Corduliidae Gomphidae	TRICHOPTERA (Caseless) Philopotamidae Polycentropodidae Rhyacophilidae Psychomyiidae Hydropsychidae
HIRUDINA Piscicolidae Glossiphoniidae Erpobdellidae Hirudinidae	HEMIPTERA Aphelocheiridae Hydrometridae Gerridae Mesoveliidae Nepidae Naucoridae Pleidae Notonectidae Corixidae	TRICHOPTERA (Cased) Odontoceridae Lepidostomatidae Goeridae Brachycentridae Sericostomatidae Beraeidae Molannidae Leptoceridae Phryganeidae Limnephilidae Hydroptilidae
CRUSTACEA Astacidae Corophiidae Gammaridae Asellidae		DIPTERA Simuliidae Tipulidae Chironomidae
EPHEMEROPTERA Siphonuridae Heptageniidae Ephemeridae Leptophlebiidae Ephemerellidae Potamanthidae Caenidae Baetidae		

Table 4.2 Taxa contributing to the eleven composite BMWP families.

BMWP Taxon Name	Contributing Taxa
Ancylidae	Ancylidae, Acroloxidae
Dytiscidae	Dytiscidae, Noteridae
Gammaridae	Gammaridae, Crangonyctidae, Niphargidae
Hydrobiidae	Hydrobiidae, Bithyniidae
Hydrophilidae	Hydrophilidae, Hydraenidae
Limnephilidae	Limnephilidae, Apatania
Planariidae	Planariidae, Dugesiidae
Psychomyiidae	Psychomyiidae, Ecnomidae
Rhyacophilidae	Rhyacophilidae, Glossosomatidae
Siphonuridae	Siphonuridae, Ameletus
Tipulidae	Tipulidae, Cylindrotomidae, Pediciidae, Limoniidae

Table 4.3 Taxa contributing to BMWP taxa Chironomidae and Oligochaeta.

BMWP Taxon Name	Contributing Taxa
Chironomidae	Chironomini, Diamesinae, Orthoclaadiinae, Podonominae, Prodiamesinae, Tanytarsini
Oligochaeta	Enchytraeidae, Glossoscolecidae, Haplotaxidae, Lumbricidae, Lumbriculidae, Naididae, Tubificidae

Table 4.4 Summary of biological samples by year and agency for spring.

Year	EA									SEPA				Grand Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	Total	E	N	W	Total	
1995	724	727	864	1266	476	1123	563	812	6555	168	251	16	435	6990
1996	806	319	303	1057	268	320	113	83	3269	76	123	16	215	3484
1997	729	344	303	800	284	427	229	22	3138	144	250	16	410	3548
1998	629	409	485	1090	325	254	297	64	3553	169	162	20	351	3904
1999	640	519	511	675	450	255	367	20	3437	226	238	20	484	3921
2000	528	868	867	1160	521	1193	559	830	6526	215	234	21	470	6996
2001	23	67	26	117	146	27	122	12	540	242	196	11	449	989
2002	456	444	441	768	456	455	382	314	3716	268	279	26	573	4289
2003	463	444	477	588	357	421	350	365	3465	726	498	22	1246	4711
2004	422	420	504	457	276	384	319	309	3091	689	459	504	1652	4743
Total	5420	4561	4781	7978	3559	4859	3301	2831	37290	2923	2690	672	6285	43575

Table 4.5 Summary of biological samples by year and agency for autumn.

Year	EA									SEPA				Grand Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	Total	E	N	W	Total	
1995	736	811	852	1493	470	1119	539	811	6831	131	170	4	305	7136
1996	728	380	155	808	259	213	195	36	2774	78	133	0	211	2985
1997	739	330	413	1161	304	196	102	16	3261	121	140	0	261	3522
1998	693	404	319	1017	353	263	294	12	3355	183	98	19	300	3655
1999	508	499	562	623	351	245	378	29	3195	223	78	20	321	3516
2000	564	619	589	851	377	1016	556	725	5297	197	156	16	369	5666
2001	285	324	202	716	363	220	181	58	2349	428	159	18	605	2954
2002	451	419	458	532	396	400	338	336	3330	257	155	25	437	3767
2003	447	432	448	573	346	389	341	332	3308	698	421	20	1139	4447
2004	400	381	496	456	289	341	310	322	2995	620	369	419	1408	4403
Total	5551	4599	4494	8230	3508	4402	3234	2677	36695	2936	1879	541	5356	42051

Table 4.6 Summary of sampling sites by agency and season.

	EA									SEPA				Grand Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	Total	E	N	W	Total	
Spring	1025	1354	1330	2205	1419	1901	1084	1063	11381	900	778	503	2181	13562
Autumn	1007	1357	1275	2156	1351	1780	1013	1110	11049	893	727	436	2056	13105

5 Construction of project database: environmental and chemical data

Introduction

The biological dataset described in Section 4 was combined with other datasets to produce the databases on which the revised MIR-max and BBN models were based. The input vector for the MIR-max models required the addition of environmental variables, whereas the BBN model required the addition of both environmental and chemical variables. The procedures used to achieve this are described in this Section.

Sample environmental data

The 13 environmental variables used in the previous project were linked with the biological data and are reproduced in Table 5.1. These were recorded for every invertebrate site by SEPA and Environment Agency and were also used by RIVPACS as predictor variables.

Table 5.1 List of 13 environmental variables used in previous project.

Variable	Description	Variable	Description
X	Global northing of NGR	DISCH	Discharge Category
Y	Global easting of NGR	BLDS	Boulders (% of substrate)
ALT	Altitude (m)	PBLS	Pebbles (% of substrate)
LDIST	Log ₁₀ distance from source	SAND	Sand (% of substrate)
LSLOPE	Log ₁₀ 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 (m)		

All of these, with the exception of X and Y, were used in the input vector for the original pattern recognition system. It was anticipated that a similar set of variables would be used in the extended models, with the possible exception of alkalinity. Alkalinity is no longer measured regularly by the Environment Agency, but given its importance in determining the composition of the biological community, alternative options were sought.

Because the value of alkalinity changes little with time, a straightforward option was to calculate the mean value for a site based on the recorded values contained in samples taken at the site. The mean value could then be used for all samples taken at that site, whether a value had been recorded or not. While the advantage of this was simplicity, the disadvantage was that no value would be available for samples taken at sites for which no values had ever been recorded. The same would apply to samples taken at new sites, for which historical records would be unavailable. The second option was to use the proportion of calcareous geology in the upstream catchment as a surrogate for alkalinity. The advantage would be that, in principle, a value would be available for all sites, with the disadvantage that, being the output of a GIS task, it might be difficult for a biologist to obtain a value for a new site.

The corresponding dataset for the BBN required matched chemical data, in addition to the biological and environmental data discussed so far. This required a number of further steps, starting with validation of the chemical dataset. It was then necessary to validate the spatial coordinates of the biological and chemical sites before finally producing the set of matched biological and chemical samples. These steps are described in the rest of this section.

Chemical data

As mentioned in Section 1, inconsistencies and incompatibilities in the data obtained from the Environment Agency in year-by-year increments resulted in the entire chemical (as well as biological) data for the period 1993-2004 being obtained again in a single retrieval from the Environment Agency's databases. The construction of the project's chemical database was completed more quickly than its biological counterpart, given the absence of time-consuming problems associated with taxonomy, although it was not without its own difficulties.

Data preparation and validation

The Environment Agency dataset contained over 24 million records and 5,500 determinands while the SEPA dataset contained over 180,000 records and 82 determinands. Given the large size of the Environment Agency database and the length of time required to run even simple queries, the Environment Agency database was rationalised by removing all redundant sites, determinands and samples. Both databases were then compared to define a common set of determinands with the same measurement units to produce a unified database.

To derive a list of chemical variables for the project, tables of sampling frequencies were produced for the Environment Agency and SEPA datasets. Based on the overall frequency of sampling, 42 determinands were then selected, as defined in Table 5.2. The frequencies of occurrence varied from roughly 90 per cent of samples down to five per cent.

Unwanted sites were eliminated by removing those that did not have the 'sample material' code recorded as 'river/running surface water', and as a result the number of sites in the list dropped from 6,662 to 6,062.

Potentially erroneous samples values were identified by producing the 'top 50' values for each variable that were then scrutinised by the Environment Agency project team (as had been done for the biological data). The consensus in the feedback was that although some values were anomalous, of greater concern was whether the sites were actually river sites. Additional data was supplied from the Environment Agency on the sample sites, which included a 'site type' field that enabled genuine river sites to be identified. Because some river sites may have been omitted from the original data retrieval, the entire chemical dataset was retrieved again, including this information, and the steps described above were repeated.

Table 5.2 Definition of chemical variables.

CIES Code	CIES Description	EA Code	EA Description	SEPA Code	SEPA Description	Unit
Alkn	AlkalinityTotal	162	ALKALINITY PH 4.5 - as CaCO3	200200	Alkalinity	mg/L
AmNI	Ammoniacal Nitrogen Non-ionised	119	AMMONIA UN-IONISED (CALCULATED)	250220	NonionNH3	mg/L
AmTN	Ammoniacal Nitrogen Total	111	AMMONIA - AS N	250200	Ammonia	mg/L
AsTI	Arsenic Total	6046	ARSENIC - AS AS	300250	As	µg/L
BOD5	BOD 5	85	BOD ATU as O2	220200	BOD (ATU)	mg/L
CaDs	Calcium Dissolved	239	CALCIUM DISSOLVED - AS CA	300125	Ca < 0.45?m	mg/L
CaTI	Calcium Total	241	CALCIUM - AS CA	300120	Ca	mg/L
CdDs	Cadmium Dissolved	106	CADMIUM DISSOLVED - AS CD	300195	Cd < 0.45?m	µg/L
CdTI	Cadmium Total	108	CADMIUM - AS CD	300190	Cd	µg/L
Chlo	Chloride Ion	172	CHLORIDE ION - AS CL	250400	Chloride	mg/L
Cond	Conductivity at 25 C	77	CONDUCTIVITY @25C	200160	ElecCond-25	µS/cm
CrDs	Chromium Dissolved	3409	CHROMIUM DISSOLVED - AS CR	300205	Cr < 0.45?m	µg/L
CrTI	Chromium Total	3164	CHROMIUM - AS CR	300200	Cr	µg/L
CuDs	Copper Dissolved	6450	COPPER DISSOLVED - AS CU	300215	Cu < 0.45?m	µg/L
CuTI	Copper Total	6452	COPPER - AS CU	300210	Cu	µg/L
FeDs	Iron Dissolved	6460	IRON DISSOLVED - AS FE	300165	Fe < 0.45?m	µg/L
FeTI	Iron Total	6051	IRON - AS FE	300160	Fe	µg/L
Hard	Hardness Total	158	HARDNESS TOTAL - as CaCO3	201100	Hardness	mg/L
HgTI	Mercury Total	105	MERCURY - AS HG	310400	Hg	µg/L
KTI	Potassium Total	211	POTASSIUM - AS K	300110	K	mg/L
MgDs	Magnesium Dissolved	235	MAGNESIUM DISSOLVED - AS MG	300135	Mg < 0.45?m	mg/L
MgTI	Magnesium Total	237	MAGNESIUM - AS MG	300130	Mg	mg/L
MnTI	Manganese Total	6050	MANGANESE - AS MN	300180	Mn	µg/L
NaTI	Sodium Total	207	SODIUM - AS NA	300100	Na	mg/L
NiDs	Nickel Dissolved	3410	NICKEL DISSOLVED - AS NI	300225	Ni < 0.45?m	µg/L
NI TI	Nickel Total	6462	NICKEL - AS NI	300220	Ni	µg/L
NO2N	Nitrite	118	NITRITE - as N	250240	Nitrite	mg/L
NO3N	Nitrate	117	NITRATE - as N	250250	Nitrate	mg/L
OPhos	Orthophosphate	180	ORTHOPHOSPHATE - as P	250300	o-Phosphate	mg/L
OxDs	Oxygen Dissolved	9924	OXYGEN DISSOLVED (INSTRUMENTAL - IN SITU) - AS O	210100	O2 - DO	mg/L
Oxsa	Oxygen % Saturation	9901	OXYGEN DISSOLVED (INSTRUMENTAL) - AS % SATN	210200	O2 -%sat	%
PbDs	Lead Dissolved	52	LEAD DISSOLVED - AS PB	300235	Pb < 0.45?m	µg/L
PbTI	Lead Total	50	LEAD - AS PB	300230	Pb	µg/L
Phos	Phosphate	192	PHOSPHATE	250320	P	mg/L
pHVI	pH	61	PH - AS PH UNITS	200100	pH	units
SiO2	Silicate	182	SILICATE REACTIVE DISSOLVED - AS SIO2	250430	Silicate	mg/L
SO4	Sulphate	183	SULPHATE - AS SO4	250410	Sulphate	mg/L
SusS	Suspended Solids	135	SOLIDS SUSPENDED @105C	140100	SuspSolids	mg/L
Temp	Temperature	76	TEMPERATURE WATER	110104	SampleTemp	°C
TOxN	Total Oxidised Nitrogen	116	NITROGEN TOTAL OXIDISED - AS N	250230	TON	mg/L
ZnDs	Zinc Dissolved	3408	ZINC DISSOLVED - AS ZN	300245	Zn < 0.45?m	µg/L
ZnTI	Zinc Total	6455	ZINC - AS ZN	300240	Zn	µg/L

Spatial matching

Methodology

Validation of the map coordinates supplied with the sample sites was extremely important. Once the biological and chemical databases were finalised, the biological and chemical sample sites had to be paired or matched with geographically close counterparts on the same stretch of water.

Spatial validation of grid references

The grid references of each site were plotted on the 1:50,000 digital river network base map supplied and information on location provided with the sample, such as an ID number or river name, was matched to a feature on the base map. The more specific the location of the feature, the more accurate the validation. For example, in most cases matching the river name of the plotted point to a river on the map would give a greater degree of confidence than simply matching by region or county.

SEPA sites

The 3,400 sites in the SEPA biological and chemical database contained a river ID. ArcGIS was used to match these with the river ID of the nearest section of the base map, provided the nearest section was within 100 m. Of the 3,400 SEPA sites processed, just 17 produced mismatches. Of these, the location of one site was out of position, ten corresponded to a meeting of two rivers where the site was closer to the 'incorrect river', and six corresponded to sites which were closer to a nearby 'incorrect' river than to the 'correct' river.

Environment Agency sites

Both biological and chemical sites had information on the water body name, region and a site description. Although the description offered valuable information, the range of entities used to define the location was too great and inconsistent between samples (names of towns, farms, streets and roads were commonly used). Obtaining this type of base map data and finding out which was relevant for each sample would have been expensive and impractical. Environment Agency regions were too broad spatially to offer any real confidence in the accuracy of the validation. This left only the information on the water body name as a viable means of validating the location of the sites. Even this had limitations, because matching a site to a river would not necessarily mean that the position along that river was correct.

However, a further method of spatial validation was provided by the 1:50,000 river network base map supplied by the Environment Agency. Each stretch of river on the base map included information on the chemical and biological site used to assess it. The sites were plotted on the map, matched to the nearest stretch of river (a process called 'snapping') and the ID of the site checked against the corresponding assessment ID for the stretch. This form of validation provided a higher degree of confidence in the accuracy of the map coordinates of the site because the stretches were much smaller geographic features than rivers or regions. The drawback was that the river network only had information on a subset of sites, which meant that matching by river name was the only option for the remaining sites.

ArcGIS software was used for the spatial matching of sample sites to river stretches. The result of this process was the joining of each of the sample site records with data from the closest stretch on the river network base map, along with a field giving the

distance between the two in metres. This dataset was used as the basis for the validation process.

As well as validating sites by matching an ID or name, a maximum limit of 100 metres was set for the distance between the site and stretch for the join to be valid. Of the original biological sites, about a quarter matched to the stretch for which they were named as the assessment site and were within the 100 metres limit. The remainder were matched by name.

The main difficulty when matching a site to a stretch by a water body name was that the names had to be identical. The names of water bodies were usually in the form of a 'distinctive' name followed by a word describing the type of water body, such as 'Lambwath Stream' or 'Hooton Brook'. Differences in spelling tended to be variations in the 'distinctive' name or the abbreviations of the category of water body, for example 'Brook', 'Brk' or 'Bk'. There was no way to deal with variations in the spelling of the 'distinctive' names other than checking them manually. It was possible to try and eliminate the problems associated with the water body category by removing these words from the description. The match was then made on the remaining 'distinctive' name words, resulting in a percentage match value.

All names of records that were within the 100 metres limit were checked, but the amount of effort put into this varied depending on the percentage match value and the distance. Those with 100 per cent match and a small distance between them were scanned mainly to check if the removed water body words matched. As the percentage dropped or the distance increased, the amount of effort put into the checking process increased. The greatest amount of effort was put into those with zero match because, though the water body names differed, they were still geographically close, so these records were those most likely to have variations in the spelling of the 'distinctive' name. The whole process was laborious and time-consuming but resulted in approximately 12,000 validated biological sites.

The same process was conducted on the chemical sites, resulting in approximately 6,000 validated chemical sites.

Matching biological and chemical sites

SEPA sites

Around 330 of the 3,400 sites were common chemical and biological sites. A search was made through the remainder to identify biological and chemical sites with matching river IDs that were within a threshold distance of 400 m, resulting in a total of 830 matched sites.

Environment Agency sites

The validated biological and chemical sites were matched according to stretch information, and using a threshold distance of 400 m. The resulting matches were checked with the locations of sewage treatment works and any pairs discarded when one was located upstream and the other downstream. This resulted in 5,300 pairs of matched sites.

Derivation of chemical statistics

Following the spatial matching of the chemical and biological samples, the chemical statistics required for each variable were derived. The diagnostic information for the pattern recognition system could accommodate a variety of statistics for each of the 42 chemical variables, with the impact on performance the only constraint - the more data that was included, the slower the application was likely to be. The key statistic for the majority of variables was the mean value over the three years prior to the sample date, with exceptions being the fifth percentile for PHVL, 10th for OXDS and OXSA, 90th for AMNI, AMTN and BOD5, 95th for SUSS and 98th for TEMP (matching the statistics used for chemical environmental standards). Other statistics could be included in the pattern recognition system if desired, following the results of performance tests. The Bayesian Belief Network, on the other hand, used a relatively small subset of the 42 chemical variables, and the model would be based on key statistics from the specification given above. Hence, for each of the 42 chemical variables, a range of statistics (mean, median, standard deviation, 5th, 10th, 90th, 95th, 98th percentiles) was generated for five different time periods prior to the sample date (three months, six months, one year, two years and three years).

The percentile values were initially estimated from the mean and standard deviation of the recorded values and the appropriate point of the normal distribution curve. The reliability of the values obtained was clearly dependent on the number of recorded values, especially for the more extreme percentile values, and a minimum threshold of N samples for a (100/N)th percentile statistic was adopted (that is, a minimum of 20 samples for a fifth percentile, for example). However, the values obtained were prone to distortion by the presence of outliers. To avoid this, an alternative was adopted based on simply ranking the values from smallest to greatest. The coverage of key statistics derived over a three-year period prior to the sample date is given for each chemical variable in Table 5.3

Table 5.3 Number of samples containing each chemical determinand for three-year sample period as minimum required sample population is increased. Determinands with asterisks are required in BBN model.

MEAS CODE	EA DETERMINAND DESCRIPTION	SAMPLE COUNT						
		>0	>4	>9	>14	>19	>24	>29
162	ALKALINITY PH 4.5 - as CaCO ₃ *	40470	37774	34258	30401	29169	27390	20805
111	AMMONIA - AS N*	44218	44128	43842	43397	42645	41544	34871
119	AMMONIA UN-IONISED (CALCULATED)	43707	42259	41613	40554	39020	36926	29998
6046	ARSENIC - AS AS	4041	3508	2935	2188	1995	1752	1443
85	BOD ATU as O ₂ *	44213	44123	43827	43371	42597	41230	34164
108	CADMIUM - AS CD	11592	9686	8234	7202	6600	5693	4562
106	CADMIUM DISSOLVED - AS CD	6751	5120	4041	3289	2818	2211	1708
241	CALCIUM - AS CA	26377	24942	22794	20181	18383	15916	11144
239	CALCIUM DISSOLVED - AS CA	5566	4322	4214	3638	3011	2691	2089
172	CHLORIDE ION - AS CL	36831	36567	35917	34991	33909	31982	26208
3164	CHROMIUM - AS CR	10490	8613	7499	6636	6053	5303	4193
3409	CHROMIUM DISSOLVED - AS CR	8408	6798	5669	4785	4148	3515	2715
77	CONDUCTIVITY @25C	15044	14922	14821	14331	13750	13047	11385
6452	COPPER - AS CU	12641	10756	9754	8827	8103	7181	5961
6450	COPPER DISSOLVED - AS CU	31708	30440	28563	26327	24389	21903	16755
158	HARDNESS TOTAL - as CaCO ₃	32332	31291	29760	27447	25345	23009	17029
6051	IRON - AS FE	7227	6378	5625	4777	4317	3731	3052
6460	IRON DISSOLVED - AS FE	6225	5376	4675	3850	3394	2867	2249
50	LEAD - AS PB	10557	8668	7570	6709	6135	5269	4166
52	LEAD DISSOLVED - AS PB	8514	6879	5756	4890	4276	3558	2783
237	MAGNESIUM - AS MG	26355	24933	22781	20173	18375	15903	11139
235	MAGNESIUM DISSOLVED - AS MG	5584	4360	4244	3667	3034	2710	2103
6050	MANGANESE - AS MN	3880	3272	2673	2149	1919	1399	1122
105	MERCURY - AS HG	4149	3893	3658	3128	2878	2640	2234
6462	NICKEL - AS NI	10664	8749	7661	6839	6272	5592	4284
3410	NICKEL DISSOLVED - AS NI	8756	7080	5932	5032	4397	3735	2751
117	NITRATE - as N	29570	27980	25196	23560	22349	21105	18126
118	NITRITE - as N	31851	30379	27879	26335	24984	23743	20578
116	NITROGEN TOTAL OXIDISED - AS N*	44197	43857	42999	41726	40111	38609	32629
180	ORTHOPHOSPHATE - as P*	42887	42012	40901	40046	39081	37708	31463
9924	OXYGEN DISSOLVED (INSTRUMENTAL - IN SITU) - AS O	33071	31233	30085	28814	27227	25415	20088
9901	OXYGEN DISSOLVED (INSTRUMENTAL) - AS % SATN*	39440	37575	36215	34700	33056	31101	25401
61	PH - AS PH UNITS*	44214	44128	43840	43405	42643	41489	34738
192	PHOSPHATE	5412	4753	4247	3846	3515	3054	2582
211	POTASSIUM - AS K	7319	6239	3801	3167	2702	2044	1457
182	SILICATE REACTIVE DISSOLVED - AS SiO ₂	6134	5588	4993	4548	4124	3210	2632
207	SODIUM - AS NA	2711	2411	2199	1883	1754	1457	1243
135	SOLIDS SUSPENDED @105C	32292	30060	27156	25032	22234	20279	16811
183	SULPHATE - AS SO ₄	4188	3757	3430	3102	2856	2154	1794
76	TEMPERATURE WATER	44216	44123	43821	43351	42601	41371	34758
6455	ZINC - AS ZN	34527	33321	31702	29623	27459	25280	19946
3408	ZINC DISSOLVED - AS ZN	8039	6114	4957	4182	3602	2907	2300

Acid-neutralising capacity

ANC (acid-neutralising capacity) was not included because of insufficient coverage of data.

Use of toxicity data

The feasibility of using toxicological data in the pattern recognition and plausible reasoning systems was investigated as an alternative to using survey data for chemicals with scant data. The study was undertaken by Veronique Adriaenssens of the Environment Agency. A short summary of the work is provided here and the full report is given in Appendix A.

- A literature review was presented on the links between macro-invertebrates and pesticides which might be relevant to the development of an impact response model such as a Bayesian Belief Network. Summaries of the results of both field-based and knowledge-based studies were given, with the conclusion that although there is plenty of information, clear causal relationships are difficult to define.

- A review of work on the impact of pesticides on macro-invertebrates undertaken by the Environment Agency and others was presented with the conclusion that current datasets are inadequate for detailed analysis and modelling, and that further work is required to develop methods for detecting stress caused by pesticides.
- Approach A. A general model with several areas of complexity was discussed. Relevant data sources were identified for estimating pesticide concentration (affected by land use, pesticide usage, soil type and leaching capacity) and toxic effect (affected by flow conditions, concentration of suspended solids, pH and dissolved organic carbon). Bayesian Belief Networks are suggested as a suitable approach because they can incorporate the causal links and model the interactions using data or other kinds of knowledge.
- Approach B. The use of species sensitive distributions (SSD) is proposed for determining the toxic effect on taxa. The SSD can be used to calculate the concentration at which a specified proportion of a species will be affected, and a community response can be estimated by extrapolation. Insufficient data is available to create the SSD, however.
- Approach C. The use of ms-PAF (Multiple Substance – Potentially Affected Fraction of species) is proposed to overcome the limitations of using SSD.

The feasibility study concluded that ms-PAF analysis would be a viable approach for incorporating pesticide data in the AI tools. Although it was outside the scope of the current project, further investigation is planned.

6 Construction of project database: stress and GIS data

Stress data

Stress data was used in the previous project as part of diagnostic information in the pattern recognition system. The data was based on the 1995 survey of rivers in England and Wales, and was summarised in Martin and Walley (2000), henceforth known as the “1995 Report”. Updated stress data were collected but problems in entering and returning the data led to delays. Analysis of the partial data that had been returned by November 2003 was reported in Martin *et al.* (2005), henceforth known as the “Interim Report”.

The 1995 Report detailed the results of the first systematic collation of types of environmental stresses that Environment Agency biologists believed were affecting the rivers of England and Wales in 1995. This was followed by the Interim Report, based on a partial return of similar data collected during autumn 2003. All the stress data which was collected and returned by Environment Agency biologists during 2003-05, including the data used in producing the Interim Report (referred to as the 2003 data), were analysed in the current project. The data were based on perceived stresses in 2003, with notes indicating any differences in stresses between 2000 and 2003. The data is summarised Martin and Paisley (2005)

The 1995 Report recommended a number of actions for improving the quality and reliability of data collection. After incorporating several modifications and additions, a revised Stress Recording System (SRS) was supplied to the Environment Agency in October 2001. However, for various reasons, the first set of data was not returned to CIES until September 2003. After the remaining stress data had been returned (the last in March 2005) a programme of validation and analysis was undertaken.

The number of GQA sites for which stress data was requested and from which data were eventually received, and the completeness of the returned data is shown in Table 6.1. Three regions (Southern, Thames and North East) returned stress data for all the requested GQA sites, and five other areas in four regions also returned stress data for all sites. The remaining areas returned partial data and only four areas returned less than 90 per cent, namely Southern Area of North West Region (85%), Eastern Area of Anglian Region and Upper Trent Area of Midlands Region (each 77%) and North Wessex Area of South West Region (70%). The overall average was 95 per cent of the GQA sites.

The complete dataset of roughly 16,700 perceived stress records from the 6,600 sites has been analysed to determine the abundance of stress intensity levels for each stress *type* (Appendix E of Martin and Paisley, 2005). Table 6.2, a summary table of this analysis, shows the forty most abundant stresses, ranked by decreasing abundance. The percentages in the table are the number of records that each specified stress represents as a percentage of the total in the stress dataset

Table 6.1 Completeness of stresses database.

Region	Area	Number of GQA sites	Sites with stress data		Matched 1995 & 2003	
			Number	%	Number	%
Anglian	Central	256	253	98.8	246	96.1
	East	306	237	77.5	221	72.2
	North	191	191	100.0	186	97.4
North East	Dales	203	203	100.0	153	75.4
	Northumbria	249	249	100.0	208	83.5
	Ridings	396	396	100.0	359	90.7
North West	Central*	252	248	98.4	232	92.1
	North	278	272	97.8	238	85.6
	South	357	305	85.4	289	81.0
Midlands	Upper Severn	260	259	99.6	223	85.8
	Lower Severn	319	319	100.0	239	74.9
	Upper Trent	280	217	77.5	212	75.7
	Lower Trent	419	419	100.0	388	92.6
Southern	Hampshire	139	139	100.0	132	95.0
	Isle of Wight	21	21	100.0	19	90.5
	Kent	240	240	100.0	238	99.2
	Sussex	138	138	100.0	128	92.8
South West	Cornwall	320	320	100.0	275	85.9
	Devon	317	315	99.4	301	95.0
	North Wessex	367	258	70.3	246	67.0
	South Wessex	175	166	94.9	146	83.4
Thames	North East	160	160	100.0	114	71.3
	South East	123	123	100.0	118	95.9
	West	289	289	100.0	280	96.9
Wales	North	212	212	100.0	207	97.6
	South East	350	349	99.7	335	95.7
	South West*	299	281	94.0	272	91.0
Totals		6916	6579	95.1	6005	86.8

* Two areas submitted data using the original 2000 sites file, not the updated 2003 sites

The maps and other information produced from this latest survey provide an update to the stresses on aquatic invertebrates in English and Welsh rivers in 2003. Although regarded as an improvement on the 1995 survey, a number of actions are recommended to improve the quality and reliability of future surveys. Attention to these should generate a better appreciation of the effects of stresses on the distribution of aquatic fauna in England and Wales. For further details see Martin and Paisley (2005).

The Environment Agency has changed the way that stresses are recorded. Stresses have been re-categorised into sectors activities, and pressures to align them with systems for managing rivers under the Water Framework Directive. See Section 10 and Appendix E of this report.

Table 6.2 Abundance of most commonly perceived stresses for 2003.

Stress Category and Type	Unkn	Light	Mod.	Svere	Total	%
STW to river - treated STW effluent		676	698	260	1,634	9.78
Run-off (non-agric.)/Leachate - urban/suburban		573	562	167	1,302	7.79
STW to river - combined sewer overflow (CSO)		297	336	89	722	4.32
Farming – fertilisers		310	280	18	608	3.64
No perceived stress*	578				578	3.46
Channel at the site - canalised stream/river (non-navigable)		134	252	162	548	3.28
Farming - other (specify)		203	320	21	544	3.26
Agricultural run-off - intensive arabilisation		128	307	80	515	3.08
Eroded material in channel - inert siltation		139	238	102	479	2.87
Run-off (non-agric.)/Leach. - highway (incl. de-icing salt)		223	154	20	397	2.38
Agricultural run-off - livestock slurry		290	87	12	389	2.33
Eutrophication – agriculture		90	192	49	331	1.98
Bank practices at site - livestock poaching/overgrazing		196	115	14	325	1.95
Run-off (non-agric.)/Leachate - light industry/commercial		100	138	56	294	1.76
Other indicators – <i>Cladophora</i>		95	136	54	285	1.71
Flow-related - regulated flow (lake/reservoir u/s)		116	117	47	280	1.68
Flow-related – weirs		90	135	48	273	1.63
Artificial bank at site - consolidated (stone/brick/concrete)		87	114	60	261	1.56
Channel at the site – bridge		144	90	12	246	1.47
Eutrophication – sewage		63	97	59	219	1.31
Channel at the site - choked channel (>33% plant)		60	111	47	218	1.30
Sampling difficulty - access to one bank only*	205				205	1.23
STW to river - storm sewer overflow (SSO)		61	105	33	199	1.19
Flow-related - other (specify)		71	92	33	196	1.17
Industrial discharge - light industry/commercial		81	82	25	188	1.13
Farming – insecticides		60	118	9	187	1.12
Bank practices at site - mown/managed riparian zone		73	71	26	170	1.02
Farming – herbicides		52	110	3	165	0.99
STW to river - other (specify)		44	71	41	156	0.93
Flow-related - river abstraction		69	63	21	153	0.92
Other indicators – ochre		70	57	25	152	0.91
Mines, quarries & extractions - coal mine drainage		68	66	17	151	0.90
Sampling difficulty - dredge*	148				148	0.89
Natural features – drought		49	78	15	142	0.85
Sampling difficulty - bouldery site		46	72	15	133	0.80
Mines, quarries & extractions - metal mine drainage		69	44	20	133	0.80
STW to river - septic tank		87	27	13	127	0.76
Flow-related - groundwater abstraction		58	51	17	126	0.75
No information*	126				126	0.75
Natural features - moorland drainage		48	63		111	0.66
Total number of perceived stresses in 2003					16,707	

*Indicates that the stress *type* does not require an intensity level to be included.

Data from geographical information systems

The environmental data associated with the biological samples and the diagnostic stress data was supplemented by data derived from GIS. Land cover and simple geological data was available for both Environment Agency and SEPA sites, and land risk data for Environment Agency sites only.

Land cover and geology data comprised the percentage cover of types of land cover and geology in the catchment upstream from the site. The accurate generation of such data clearly required a reliable estimate of upstream catchment area. Some preliminary work was undertaken early in the project to validate estimated upstream catchment areas for some 6,800 Environment Agency GQA sites prior to the generation of the GIS data.

Using simple checks, such as the known relationship of monotonically decreasing upstream catchment area with increasing altitude for sites on the same water course, the upstream catchment areas for nearly 500 sites were identified as being potentially in error. Subsequent investigation by Environment Agency staff confirmed a problem in up to half of these cases, identified as a 'snapping error', where a site might be 'snapped' to a tributary rather than the main river, underestimating the upstream catchment area, or a modelling error, where a site was associated with a canal or loop linked to an artificial drainage grid, neither of which could be accurately represented by the hydrological model used to generate catchment areas.

LowFlow2000 data provided upstream catchment areas for nearly 5,700 GQA sites, and these were compared with the areas for the 6,800 GQA sites provided by the Environment Agency. Removing the 500 potentially erroneous Environment Agency data yielded 5,180 common sites with two estimated values. The criterion used elsewhere, that two values of catchment area were inconsistent if they differed by more than 15 per cent and their absolute difference was greater than three km² (to remove many discrepancies for small catchment areas), showed inconsistencies for 483 of the 5,180 sites. The Environment Agency project manager reported that comparison of these 483 with a third set produced agreement with one or other value in about half of cases.

This preliminary work was superseded by the spatial validation exercise described in Section 5, which was carried out for both GQA and non-GQA sites, and for both Environment Agency and SEPA. In this exercise, locations were confirmed to within 100 m of the river network, checks were carried out on river names, and sites were screened for unwanted water bodies such as dykes, ditches and canals. The complete list of validated biological sites was sent to the Environment Agency and SEPA and the GIS data generated. This was not completely straightforward, however. It transpired that the digital river network used to generate the GIS data was a simplified version compared with the network used to validate the data. Hence the threshold of 100 m could not be applied to all the sites, and in such cases GIS data was not generated.

The geology, land cover and land risk data are described in more detail below.

Geology

The data for Environment Agency and SEPA sites relate to the percentage cover of the geological categories given in Table 6.3 in the catchment upstream.

Table 6.3 Geology categories.

Calcareous
Salt
Siliceous
Peat

To assess how good a proxy calcareous geology was for alkalinity, MIR-max models using both were produced and evaluated - see Section 8.

Land cover

The land cover data included is a mapping generated by Land Cover Map 2000 (LCM2000), a thematic classification of spectral data recorded by satellite images, with external datasets used to add context to help refine the spectral classification. The categories are based on the classifications used in the Countryside Survey 2000 (Fuller *et al.*, 2002) and the coverage for the Environment Agency and SEPA sites is given in Table 6.4.

Table 6.4 Land cover categories.

Land Cover Category	Class	EA	SEPA
Broad leaved/mixed woodland	1.1	Yes	Yes
Coniferous woodland	2.1	Yes	Yes
Arable cereals	4.1	Combined	Mostly separate
Arable horticulture	4.2		
Non-rotational arable	4.3		
Improved grassland	5.1	Yes	Yes
Setaside grass	5.2	Yes	Yes
Neutral grass	6.1	Yes	Yes
Calcareous grass	7.1	Yes	Yes
Acid grass	8.1	Yes	Yes
Bracken	9.1	Yes	Yes
Dense dwarf shrub heath	10.1	Combined	Mostly separate
Open dwarf shrub heath	10.2		
Fen marsh swamp	11.1	No	Most
Bogs	12.1	Yes	Yes
Inland water	13.1	Yes	Yes
Montane	15.1	No	Most
Inland bare ground	16.1	Yes	Yes
Suburban	17.1	Yes	Yes
Continuous urban	17.2	Yes	Yes
Supra littoral rock	18.1	No	Most
Supra littoral sediment	19.1	No	Most
Littoral rock	20.1	No	Most
Littoral sediment	21.1	No	Most
Saltmarsh	21.2	Yes	Yes
Sea/estuary	22.1	No	Most

Land risk scores

Land risk data (available only for Environment Agency sites) comprises indices quantifying the risk associated with the six categories shown in Table 6.5. The scores represent the perceived risk based on a land use categorisation at the resolution of one-km grid squares. Higher scores represent greater risks.

Table 6.5 Land risk categories.

Sheep
Sediment
Pollution
Pesticide
Phosphorus
Nitrogen

The main datasets used to generate these scores were:

- Department for Environment, Food and Rural Affairs (Defra) agricultural census data on land use (2000, which was the last 'full' census available, of 160,000 farmers).
- IACS (Integrated Administration and Control System) register data, including grassland (2004, of 72,000 farmers).

Other supplementary datasets consisted of the following:

- Groundwater vulnerability.
- Nitrate Vulnerable Zone (NVZ) areas.
- Joint Environment Agency/English Nature priority areas for diffuse pollution, developed for Defra's Catchment Sensitive Farming (CSF) Programme.
- Proximity of the land to surface watercourses.
- Soil erosion and slope.
- Nutrient, pesticide and sheep dip usage.
- Pollution data from the Environment Agency's National Incident Recording System (NIRS II).

The component datasets were combined to assess each square kilometre of land. Where 'high risk' combinations occur frequently, the area was designated with a higher risk classification and vice versa for 'low risk' combinations. An example of a high risk score in the 'Nitrogen' category could be: "High dairy cattle and fodder maize density, within an NVZ in a eutrophic river catchment".

7 Construction of project database: flow data

Introduction

Factors relating to the physical flow of water in a river are known to be among the important parameters that determine the composition of the benthic biological community. For example, the flow velocity is a key factor in determining the nature of the substrate of the river bed. Although flow velocity is not recorded directly at sampling sites and so could not be included in the datasets, it has a clear relationship with the slope of the site, which is recorded. The inclusion of slope of the site as one of the physical parameters in the input vector to the clustering models (see Section 8) enables some of the likely effects of flow velocity to be accounted for.

The quantity of flow is also likely to have an impact on the community, especially at the lower end of the range. Water in a river with low flow is likely to be less oxygenated and of generally poorer quality compared to water in the same river in higher flow conditions. In turn, this would make the presence of sensitive taxa less likely in low flow conditions and the presence of tolerant taxa more likely. Because flow was not accounted for in the previous pattern recognition or reasoning models, one of the objectives of the project was to include one or more measures related to flow and to investigate their effects.

Two measures of flow were included. The first is based on estimated duration curves of natural and influenced flow derived from long-term averaged data by LowFlows 2000. The second attempts to quantify directly the condition of the river, prior to the sampling date, compared to a 30-year average, and is derived from time series data. Both measures overcome the difficulty that GQA sampling sites are located in catchments with no measured flow data, and are described in this section.

LowFlows 2000

An Environment Agency contract was set up with Wallingford HydroSolutions Ltd. to model flow data at all GQA sites using LowFlows Enterprise (an update of LowFlows 2000, Young *et al.*, 2003), a suite of modelling techniques designed to estimate natural and artificially influenced river flow at sites that are ungauged. To obtain these estimates, it was necessary to define the boundary of the catchment that drains through the site of interest.

For the natural flow, the ungauged catchment is characterised using the Met Office standard average annual rainfall (SAAR), annual average run-off (ARRO) and the fractional extent of the hydrology of soil types classification (HOST). The long-term standardized natural flow duration curve is then estimated from a group of ten similar catchments from a reference source pool of gauged catchments and taking a weighted mean of observed flow statistics for the selected catchment. The flow duration curve expresses the frequency distribution of flows at a point on the river, and defines the relationship between flow of a given magnitude and the probability of exceeding it. Low flows are exceeded frequently (a high percentage), whilst high flows are exceeded rarely (a low percentage). For example the Q95 value is the flow that will be equalled or exceeded 95 per cent of the time.

The catchment boundary is used to query a geo-referenced database of influence features, including surface and groundwater abstractions, impounding reservoirs and discharges from which the corresponding influenced flow duration curve is estimated.

Because of the large margins of error in flow estimates, there is greater confidence in measures that normalize with respect to flow rather than use absolute flow or influence directly. The recommended statistic is the percentage impact (at 95 per cent exceedence probability) defined as:

$$\frac{\text{Natural Flow} - \text{Influenced Flow}}{\text{Natural Flow}} \times 100$$

This was adopted as a measure of the extent of pressure on the river flow for classification into WFD class intervals.

Development of a method for estimating flow condition

Introduction

Additional efforts were made to acquire actual flow records to determine the flow regime in the river before each biological sample was collected. This would help to indicate whether the sample was taken when conditions were wetter than in an average year, average or drier than average, which in turn would enable links between the biology and flow to be investigated.

Sampling sites were not spatially matched to gauging stations, so there was some discussion of how best to relate the gauging data to the sampling sites. Staff from the Environment Agency Hydrology Team were consulted about the best method to provide reasonable accuracy while keeping the task manageable. The Hydrology Team agreed to provide average monthly flow records over at least a 30-year period for one site in each of the 125 Catchment Abstraction Management Strategy (CAMS) River Basin Districts.

The proposed method was based on ranking the monthly time-series flow data from driest to wettest for each gauging station, and then estimating conditions at GQA assessment points using spatial interpolation. The exercise was treated as a feasibility study and was confined to sites in England and Wales. Refining the method and extending the analysis to cover Scotland (which would require the use of other data to act as a surrogate for flow, such as rainfall) were left for a future project.

Data

Time series data of varying historical lengths were supplied for 217 gauging stations in England and Wales. In 164 cases, the data covered at least the 30-year period January 1976 to December 2005. Of these 164 cases, the data for 81 stations contained complete monthly flow records with no gaps or absences. It was felt desirable to use an additional 43 stations for which the proportion missing was less than five per cent (corresponding to less than 18 records missing from the total of 360) to maximise the amount of data available for spatial interpolation. The grid references were verified and the somewhat uneven distribution of the locations plotted, as shown in Figure 7.1.

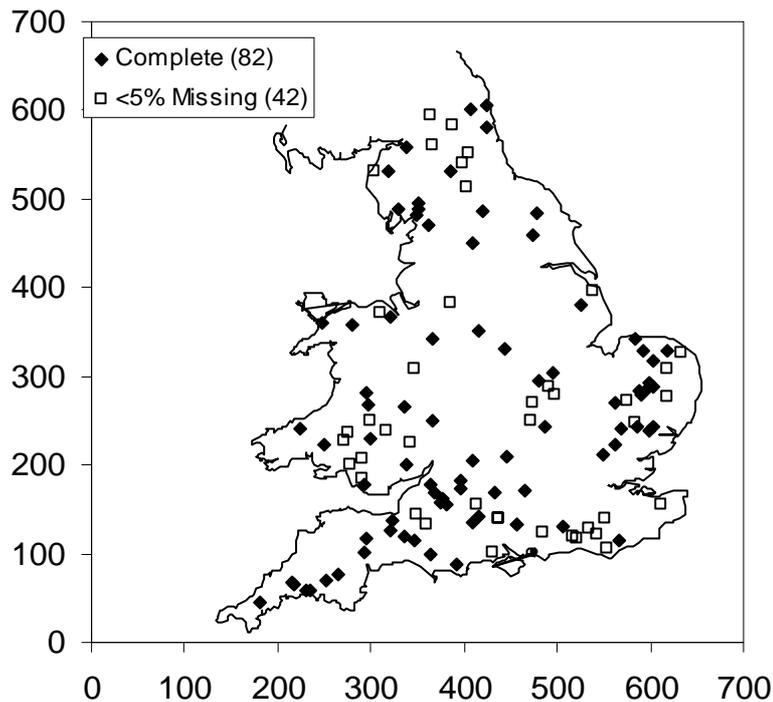


Figure 7.1 Distribution of 124 gauging stations with complete monthly flow records or less than five per cent missing for January 1976 - December 2005.

Data quality

Among the 81 gauging stations with complete data, two stations labelled Drove Lane and River Alre were geographically close together (within 60 m). When the flows were ranked from driest to wettest (see later) quite different rankings were produced. Both rankings were compared with those of the nearest neighbours and the worst in agreement was discarded from the analysis. A further case was found among the 43 stations with missing values (two stations labelled Fullerton and River Anton, 150 m apart).

The gauged data covered the dry summer of 1976, which was useful for screening potential anomalies. According to the ranking produced, Oakley Park on River Dove (of the 43 sites with missing values) experienced its third wettest summer in 1976. This seemed unlikely and this record was removed from the analysis.

The five flow records mentioned above were returned to the Hydrology Team for further investigation, and the analyses proceeded, based on 121 records (80 with complete data, and 41 with less than five per cent missing values).

Each monthly flow value was accompanied by a quality code. Although most of the 121 records had a significant number of months with a quality code other than G ('good') the data were used regardless.

Analysis of data

For each of the 80 gauging stations with complete data, the flow (m^3/s) in each of the 12 months was ranked over the 30-year period from one (driest) to 30 (wettest) and a ranking score given of $(\text{rank}-1)/29$, ranging from zero (driest) to one (wettest). The data for the 41 stations with missing values were ranked similarly and the ranking score given, with the total for months containing missing values adjusted appropriately.

The viability of spatial interpolation depends on the extent of spatial variation in the ranking score for any particular month. This was quantified by determining the standard deviation in ranking scores across gauging stations for that month and averaging over the 30 years. The 30-year averages show that overall, spatial variation in the ranking score is least in the winter months, with a minimum in February, and most in the summer months, with a maximum in July, Table 7.1.

Table 7.1 Extent of spatial variation in ranking score across 121 gauging stations for each month, averaged over 30 years.

Month	Average St Dev in Rank Score
Jan	0.184
Feb	0.172
Mar	0.196
Apr	0.178
May	0.198
Jun	0.211
Jul	0.222
Aug	0.209
Sep	0.219
Oct	0.183
Nov	0.192
Dec	0.194

This variation over the year can be explained in terms of meteorology and soil moisture deficit. Rainfall in the winter months is driven by large scale Atlantic fronts covering large portions of the country, while rainfall in the summer months may be influenced more by localised events such as thunderstorms. Soil moisture deficit (a measure of dryness in the soil) is likely to be least, and most uniform, in months towards the end of winter such as February, whereas greater variation is expected in the summer months.

An indication of how spatial variation varied in a particular month for dry and wet years is provided by plotting the standard deviation the ranking score against mean ranking score for all stations in the analysis. Figures 7.2 and 7.3 show such plots for February and August.

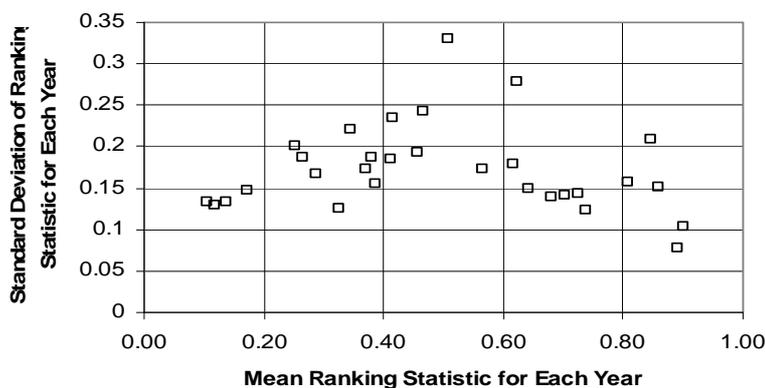


Figure 7.2 Variation in ranking score against mean ranking score across 121 stations for February.

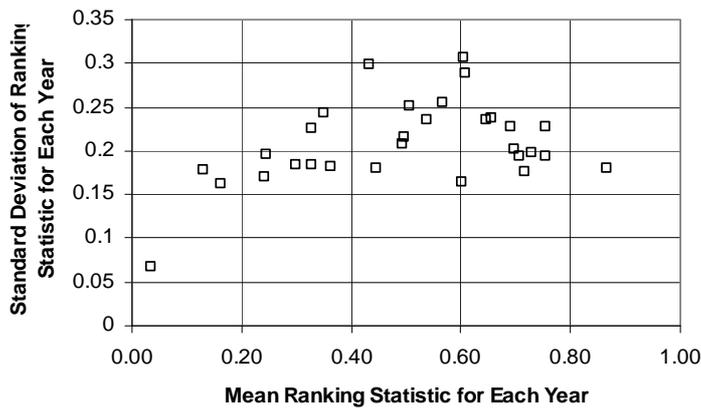


Figure 7.3 Variation in ranking score against mean ranking score across 121 stations for August.

Two observations can be drawn from these plots. Firstly, there is as little spatial variation during a wet February (ranking score close to one) as in a dry August (ranking score close to zero). The meteorology explains this, because conditions in February and August both tend to be driven by large-scale events (Atlantic fronts in February and high pressure in August). Secondly, however, there seems to be less spatial variation in a dry February (ranking score close to zero) than in a wet August (ranking score close to one). This is also explained by the drivers for each set of conditions; even in a dry February, rain is still provided by large scale Atlantic fronts, while in a wet August, rain is more likely to result from more local events such as thunderstorms.

Given these observations, it was likely that spatial interpolation would be most accurate in conditions resembling a wet February or a dry August. Taking the months as a whole, spatial interpolation was likely to be more accurate in February than in August. This was confirmed by the next part of the analysis.

Assessment of validity of spatial interpolation

The validity of spatial interpolation was quantified by predicting the ranking statistic at each point by interpolating from neighbouring points, and comparing the prediction with the actual value. The interpolation was implemented using a 'gravity model', where the ranking statistic \hat{S}_n at station n was predicted from the ranking statistic at all other stations S_i ($i = 1, \dots, N, i \neq n$) according to

$$\hat{S}_n = \frac{\frac{1}{r_{n1}^2} S_1 + \frac{1}{r_{n2}^2} S_2 + \frac{1}{r_{n3}^2} S_3 + \dots + \frac{1}{r_{nN}^2} S_N}{\frac{1}{r_{n1}^2} + \frac{1}{r_{n2}^2} + \frac{1}{r_{n3}^2} + \dots + \frac{1}{r_{nN}^2}}$$

where r_{ni} is the Euclidian distance from station n to station i . The algorithm can be modified so that the calculation includes only those within a certain radius, or a certain number of nearest neighbours, although the results derived were obtained using all possible stations.

This is illustrated in Table 7.2 for a typical gauging station, where the actual and predicted ranks are given, along with the magnitude of the difference $|S_n - \hat{S}_n|$, the mean of which over the 30 months was 0.09. An indication of the accuracy of the prediction is also given in terms of whether the site represents wetter, drier or average conditions. For this, the ranking statistic was broken down into the ranges $[0, 1/3]$, $[1/3, 2/3]$ and $[2/3, 1]$ to represent drier than average, average and wetter than average

conditions. Changes in the predicted condition compared to the actual condition are highlighted in bold. For this station there were ten changes, mostly bunched around the two boundaries separating the three ranges.

The results of repeating this procedure for the 121 stations in a particular month, and then for all twelve months is shown in Table 7.3. For each month, Table 7.3 gives the average and maximum number of changes at a typical station when taken over the 30 years. The grand average number of changes in condition taken over the twelve months was 7.46, representing 25 per cent of the total.

The mean and maximum deviation in the ranking score, and correspondingly the mean and maximum number of changes, is least in the winter months, notably February, and greatest in the summer months, notably July and August. This concurs with the lesser amount of spatial variation in the data noted earlier in February compared with the summer months.

Table 7.2 Actual and predicted ranking scores and conditions for August record for typical gauging station.

	Rank			Condition		
	Actual	Predicted	Error	Actual	Predicted	Change
01-Aug-85	1.00	0.89	0.11	Wet	Wet	No
01-Aug-04	0.97	0.90	0.06	Wet	Wet	No
01-Aug-88	0.93	0.82	0.11	Wet	Wet	No
01-Aug-80	0.90	0.82	0.08	Wet	Wet	No
01-Aug-02	0.86	0.79	0.07	Wet	Wet	No
01-Aug-79	0.83	0.75	0.08	Wet	Wet	No
01-Aug-82	0.79	0.59	0.21	Wet	Ave	Yes
01-Aug-86	0.76	0.73	0.03	Wet	Wet	No
01-Aug-98	0.72	0.67	0.05	Wet	Wet	No
01-Aug-93	0.69	0.60	0.09	Wet	Ave	Yes
01-Aug-78	0.66	0.73	0.08	Ave	Wet	Yes
01-Aug-87	0.62	0.68	0.06	Ave	Wet	Yes
01-Aug-01	0.59	0.61	0.02	Ave	Ave	No
01-Aug-92	0.55	0.57	0.02	Ave	Ave	No
01-Aug-96	0.52	0.28	0.24	Ave	Dry	Yes
01-Aug-94	0.48	0.47	0.01	Ave	Ave	No
01-Aug-77	0.45	0.45	0.00	Ave	Ave	No
01-Aug-00	0.41	0.55	0.13	Ave	Ave	No
01-Aug-89	0.38	0.31	0.07	Ave	Dry	Yes
01-Aug-91	0.34	0.29	0.06	Ave	Dry	Yes
01-Aug-81	0.31	0.45	0.14	Dry	Ave	Yes
01-Aug-05	0.28	0.35	0.07	Dry	Ave	Yes
01-Aug-99	0.24	0.46	0.22	Dry	Ave	Yes
01-Aug-90	0.21	0.15	0.05	Dry	Dry	No
01-Aug-03	0.17	0.29	0.12	Dry	Dry	No
01-Aug-97	0.14	0.28	0.14	Dry	Dry	No
01-Aug-84	0.10	0.18	0.07	Dry	Dry	No
01-Aug-83	0.07	0.22	0.15	Dry	Dry	No
01-Aug-95	0.03	0.07	0.04	Dry	Dry	No
01-Aug-76	0.00	0.05	0.05	Dry	Dry	No
			0.09			10

Table 7.3 Summary for 121 stations with averages etc taken over 30 years.

Month	Summary of averages over the 30 years			
	Mean Ave Deviation	Max Ave Deviation	Mean No Changes	Max No Changes
Jan	0.093	0.237	6.9	18
Feb	0.086	0.214	6.4	16
Mar	0.094	0.255	7.1	19
Apr	0.089	0.226	5.9	16
May	0.098	0.241	6.9	18
Jun	0.118	0.342	8.4	19
Jul	0.137	0.365	9.9	23
Aug	0.120	0.360	8.5	23
Sep	0.129	0.372	9.0	22
Oct	0.10	0.26	6.9	16
Nov	0.096	0.230	6.6	18
Dec	0.097	0.203	7.1	17
Average	0.105	0.276	7.46	18.8

The comparison between the actual and predicted flow condition was broken down further in Table 7.4 to show the proportion of changes in each category for each month. For each table, the percentage of predicted conditions (horizontal) was given for each actual flow condition (vertical). In each case, the greatest proportion (around 25 per cent) represents no change in the predicted condition compared with the actual. The next largest proportion (five to 11 per cent) represents a change to 'average' from either 'dry' or 'wet', while a smaller proportion (three to five per cent) represents a change from 'average' to either 'dry' or 'wet'. The smallest proportion (under one per cent) represents a change of two categories, from 'dry' to 'wet' or vice versa. Although only a small proportion, this represents serious error.

The same data is averaged over the twelve months in Table 7.5, where the effects of the misclassifications are to swell the 'average' category by around a fifth (from 33 to 40 per cent) and shrink the 'dry' and 'wet' categories by around a tenth each (from 33 to 30 per cent).

Table 7.4 Actual (vertical) and predicted (horizontal) flow condition by month.

Jan				Feb				Mar			
%	Dry	Ave	Wet	%	Dry	Ave	Wet	%	Dry	Ave	Wet
Dry	25.5	7.7	0.2	Dry	27.3	5.9	0.2	Dry	25.9	7.3	0.2
Ave	3.0	25.0	5.1	Ave	5.4	23.8	4.0	Ave	3.9	25.0	4.2
Wet	0.2	7.1	26.2	Wet	0.1	5.7	27.6	Wet	0.2	7.9	25.4

Apr				May				June			
%	Dry	Ave	Wet	%	Dry	Ave	Wet	%	Dry	Ave	Wet
Dry	27.0	6.2	0.2	Dry	26.1	7.1	0.2	Dry	24.0	9.0	0.5
Ave	3.7	25.4	3.9	Ave	4.9	24.6	3.4	Ave	4.8	23.2	5.0
Wet	0.2	5.5	27.9	Wet	0.2	7.2	26.1	Wet	0.7	8.3	24.5

July				Aug				Sept			
%	Dry	Ave	Wet	%	Dry	Ave	Wet	%	Dry	Ave	Wet
Dry	21.5	11.1	0.9	Dry	23.7	9.0	0.7	Dry	22.8	10.0	0.7
Ave	4.1	24.0	5.1	Ave	4.3	23.9	4.9	Ave	4.3	24.2	4.6
Wet	0.9	11.0	21.6	Wet	0.8	8.7	24.0	Wet	0.9	9.7	23.0

Oct				Nov				Dec			
%	Dry	Ave	Wet	%	Dry	Ave	Wet	%	Dry	Ave	Wet
Dry	26.0	7.2	0.2	Dry	26.4	6.9	0.2	Dry	26.3	6.9	0.2
Ave	6.1	23.0	3.9	Ave	4.5	25.0	3.5	Ave	4.2	24.1	4.8
Wet	0.4	5.2	27.9	Wet	0.2	6.9	26.4	Wet	0.4	7.2	25.9

Table 7.5 Actual (vertical) and predicted (horizontal) flow condition averaged over all months.

Year Average				
%	Dry	Ave	Wet	Total
Dry	25.2	7.8	0.4	33.4
Ave	4.4	24.3	4.4	33.1
Wet	0.4	7.5	25.5	33.4
Total	30.0	39.6	30.3	

The locations of stations with greatest number of serious prediction errors (that is a prediction of 'wet' when the actual condition is 'dry' or vice versa) were examined to determine the extent to which the density of coverage might be a factor. The stations are plotted with an indication of the number of serious errors in Figure 7.4. The density of coverage does not seem to be a factor, because the seven stations with more than ten serious errors (represented by "+") are not located where the data is sparsest. On the contrary, they tend to be relatively near to other data points, and this is the source of the difficulty.

Examination of the ranking scores of these seven stations show some large differences compared to the ranking scores of stations around them, particularly in July and August, and hence predictions using neighbouring stations agree poorly with actual values. There are two possible reasons for this. Either flow records really are indicative of large spatial variation in flow rankings over short distances in these locations, or flow records at these stations are somewhat unreliable. If the former, there is little that can be done, whereas if the latter, these stations could be removed from the data, as others were removed at an earlier stage.

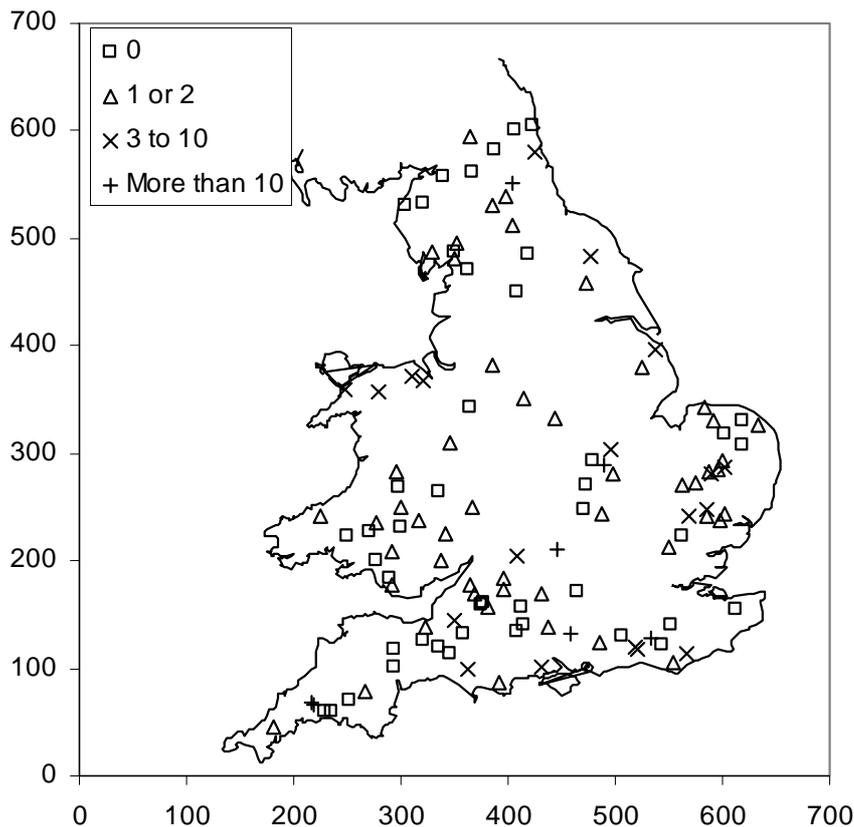


Figure 7.4 Locations of gauging stations and indication of number of serious errors (that is, wet to dry or vice versa).

Analysis of impact of flow condition on taxa

To assess the impact of flow condition on taxa, values of flow condition were calculated for each biological sample for one, two, three, six, 12 and 24 months before the sample date, achieved by interpolation from the ranking statistic given at 121 gauging stations, and averaging over the appropriate number of previous months. As before, values in the ranges $[0, 1/3]$, $[1/3, 2/3]$ and $[2/3, 1]$ were considered 'dry', 'average' and 'wet' respectively.

The data were then split by season into spring and autumn, and by site type into riffles (substrate contained over 70 per cent boulders and pebbles), pools (substrate contained over 70 per cent sand and silt), and riffle/pools (neither riffle nor pool). The abundance distributions (absent and abundance categories 1-4) for each taxon for 'dry', 'average' and 'wet' were then deduced from frequencies of occurrence in the data. Taxa exhibiting a significant change in the probability of absence (greater than 0.05 in magnitude) when the distribution in 'dry' conditions was compared with 'wet' were identified.

Taxa for which the change was positive (that is, the likelihood of absence was greater in 'wet' conditions than 'dry') for riffle sites in spring are given in table 7.6 for a range of 'wet' and 'dry' periods prior to the sampling date, from one month to two years. In general, the change in probability increases with increasing time period. These taxa clearly prefer drier conditions, where the water may be less oxygenated and of poorer quality. The revised BMWP scores (Paisley *et al.*, 2007) are also given in Table 7.6, and the relatively low scores in general confirm the tolerant nature of these taxa.

Table 7.6 Taxa for which probability of absence was generally greater in 'wet' conditions than 'dry', for riffle sites in spring.

Taxon	Score	Change in Absence Prob (W to D) Prior to Sample Date (Months)				
		1	3	6	12	24
Glossiphoniidae	3.2	0.00	0.04	0.11	0.11	0.18
Lymnaeidae	3.3	0.06	0.09	0.12	0.20	0.12
Asellidae	2.8	0.02	0.07	0.08	0.11	0.12
Planorbidae	3.1	0.02	0.07	0.09	0.10	0.12
Erpobdellidae	3.1	0.01	0.04	0.08	0.08	0.11
Sphaeriidae_Pea_mussels	3.9	0.00	0.06	0.09	0.07	0.08
Hydrobiidae	4.2	0.01	0.03	0.04	0.02	0.07
Sialidae	4.3	0.00	0.00	0.02	0.04	0.07
Valvatidae	3.2	0.01	0.02	0.04	0.05	0.06
Coenagriidae	3.5	0.00	0.01	0.02	0.04	0.06
Halplidae	3.6	0.02	0.02	0.02	0.05	0.06
Ancylidae	5.8	0.01	0.02	0.06	0.06	0.02
Caenidae	6.5	0.05	0.08	0.04	0.06	-0.02
Dytiscidae	4.7	0.08	0.02	-0.01	0.06	0.02
Psychomyiidae	5.9	0.02	0.07	0.06	0.05	0.03

Taxa for which the change was negative (that is, the likelihood of absence is less in 'wet' conditions than 'dry') for riffle sites in spring are given in Table 7.8 for the same range of 'wet' and 'dry' periods prior to the sampling date. For these taxa too, the change in probability generally increases with increasing time period. These taxa clearly prefer wetter conditions, where the water may be more oxygenated and of higher quality, and their more sensitive nature is confirmed by their revised BMWP scores which are higher in general than those for the taxa in Table 7.7.

Table 7.7 Taxa for which probability of absence was generally less in 'wet' conditions than 'dry', for riffle sites in spring.

Taxon	Score	Change in Absence Prob (W to D) Prior to Sample Date (Months)				
		1	3	6	12	24
Rhyacophilidae	8.2	-0.04	-0.13	-0.16	-0.21	-0.22
Leuctridae	10.0	-0.02	-0.07	-0.14	-0.23	-0.22
Heptageniidae	9.7	-0.02	-0.10	-0.14	-0.15	-0.20
Ephemerellidae	8.2	-0.04	-0.02	-0.10	-0.16	-0.19
Sericostomatidae	9.1	-0.01	-0.06	-0.06	-0.08	-0.17
Perlodidae	10.8	0.00	-0.12	-0.15	-0.15	-0.17
Goeridae	8.8	-0.05	-0.10	-0.05	-0.11	-0.16
Lepidostomatidae	10.1	0.00	0.00	-0.02	-0.11	-0.16
Elmidae	6.6	0.03	-0.05	-0.06	-0.06	-0.15
Taeniopterygidae	11.3	0.00	-0.13	-0.13	-0.10	-0.14
Leptophlebiidae	8.8	0.02	-0.02	-0.07	-0.08	-0.14
Baetidae	5.5	-0.05	-0.08	-0.10	-0.14	-0.14
Hydrophilidae	7.4	0.05	-0.01	-0.05	-0.05	-0.14
Gyrinidae	8.2	0.03	-0.09	-0.08	-0.05	-0.13
Nemouridae	9.3	0.00	-0.16	-0.16	-0.14	-0.13
Simuliidae	5.8	-0.04	-0.07	-0.07	-0.13	-0.11
Planariidae	5.0	-0.02	-0.06	-0.03	-0.11	-0.11
Chloroperlidae	11.6	-0.01	-0.02	-0.06	-0.11	-0.10
Hydroptilidae	6.2	0.07	0.13	0.07	0.07	-0.10
Hydropsychidae	6.6	0.01	-0.03	-0.04	-0.04	-0.09
Limnephilidae	6.2	-0.07	-0.12	-0.09	-0.11	-0.09
Tipulidae	5.9	0.03	-0.02	-0.05	-0.07	-0.08
Polycentropodidae	8.1	0.02	-0.03	-0.05	-0.05	-0.08
Odontoceridae	11.0	-0.01	-0.01	-0.01	-0.06	-0.06
Ephemeridae	8.4	0.00	0.00	0.00	-0.01	-0.06
Leptoceridae	6.7	0.02	0.02	0.03	0.04	-0.06
Gammaridae	4.5	0.00	-0.05	-0.06	-0.03	-0.04

Corresponding tables summarising the same analysis for riffle sites in autumn, and for the pool sites in both seasons are given in Appendix B.

Summary

The records of 80 gauging stations containing complete data for 30 years, January 1976 to December 2005, were analysed along with a further 41 stations containing less than five per cent of missing data. For each of the twelve months, the 30 flow values for each station were ranked and given a score ranging from zero (driest) to one (wettest).

The standard deviation of the ranking score indicated that the overall spatial variation was less in winter months, and in particular February, than in summer months such as July and August. The spatial variation was least in a wet February or a dry August, and less in a dry February than a wet August. These observations are consistent with explanations based on meteorological drivers and soil moisture deficit.

The validity of the spatial interpolation was assessed by comparing the predicted value of the ranking statistic at each station with the actual value. When the ranking statistic was banded so that [0,1/3], [1/3,2/3] and [2/3,1] represented dry, average and wet conditions, the overall average number of changes in condition for a typical station over 30 years was 7.46, or approximately 25 per cent of cases. The number of changes

was least in the winter months, notably February, and greatest in the summer months, consistent with the spatial variation found in the data.

This level of accuracy is likely to be reflected in predictions of flow condition at a sampling site by interpolation from the gauged data used here. That is, an assessment of whether the site was dry, average or wet is likely to be correct in about 75 per cent of cases, with greater accuracy in the winter months and less in the summer. If the same characteristics applied, it might be assumed that, on average, an assessment of average conditions would be incorrect in about 20 per cent of cases, and an assessment of dry or wet conditions incorrect in around 10 per cent of cases, with the true condition one category removed. It would be possible for an assessment of dry or wet conditions to be in serious error (that is, dry condition when it should be wet or vice versa), but only in less than one per cent of cases.

The interpolation scheme used was the simplest possible, and the effect of alternative schemes could be investigated. Increasing the number and uniformity of the gauging stations may also be beneficial, although an investigation of locations of stations with the greatest number of serious prediction errors suggested that this was not a major factor.

Despite the flaws in the method that was adopted, when effects on taxa were examined it was found that the results accorded entirely with expectation. Taxa more likely to be absent in wet periods compared to dry were those which tolerate poorer quality conditions, confirmed by their generally lower revised BMWP scores. Conversely, taxa less likely to be absent in wet periods compared to dry were those sensitive to poorer quality conditions, confirmed by their generally higher revised BMWP scores. The results confirm the importance of flow condition prior to sampling date as a factor in determining the composition of the community, and suggest it may be a useful addition to variables in pattern recognition and reasoning systems, alongside the percentage impact statistic from LowFlows Enterprise.

Then the mutual information between C and X_j is given by:

$$M(C, X_j) = \sum_{i=1}^n \sum_{k=1}^s \alpha_{ijk} \log_2 \left(\frac{\alpha_{ijk}}{\beta_i \lambda_{jk}} \right)$$

where: α_{ijk} = probability of finding attribute X_j in its k^{th} state in class C_i
 β_i = prior probability of class C_i
 λ_{jk} = prior probability of finding attribute X_j in its k^{th} state.

These probability values are estimated from the current distribution of the sample data between classes (see Figure 8.1) as follows:

$$\alpha_{ijk} = p_{ijk}/T, \quad \beta_i = q_i/T, \quad \lambda_{jk} = r_{jk}/T$$

Summing mutual information values over m attributes gives:

$$G = \sum_{j=1}^m M(C, X_j)$$

and the average value of mutual information between classes and attributes is given by:

$$\bar{M}(C, X) = G/m.$$

The aim of the clustering process is to maximise G – see references above for further details.

Although no changes were made to the clustering procedure, the concept of entropy was used. Entropy is a measure of uncertainty and is closely related to mutual information. Mutual information can be given in terms of the reduction in uncertainty by an alternative formula as follows:

$$M(C, X_j) = H(X_j) - H(C, X_j)$$

where $H(X_j)$ is the entropy (uncertainty) of the attribute X_j :

$$H(X_j) = - \sum_{k=1}^s \lambda_{jk} \log \lambda_{jk}$$

and $H(C, X_j)$ is the conditional entropy of the attribute X_j (that is, the remaining uncertainty given knowledge of the classes):

$$H(C, X_j) = \sum_{i=1}^n \beta_i H(C_i)$$

where $H(C_i)$ is the entropy (uncertainty) of class C_i :

$$H(C_i) = - \sum_{k=1}^s (\alpha_{ijk}/\beta_i) \log_2 (\alpha_{ijk}/\beta_i)$$

$H(C_i)$ is a measure of the quality of the class. Its maximum value would be obtained when all states were equally likely (and uncertainty was greatest), and a large value would represent a 'poor quality' class that contained disparate samples. The minimum value (of zero) would be obtained if the attribute occupied a single state for all samples in the class (and uncertainty was least). A small value would represent a 'good quality' class in which all samples were similar (although care has to be taken – a zero value would also be given for a class containing just a single sample).

Ordering procedure (R-max)

The ordering procedure arranges the output bins (or classes) so that neighbouring bins represent similar patterns (biological communities) and well-separated bins represent different ones. This is achieved by maximising the correlation coefficient between distances in output space (that is, the distances between the bins), and the distances in data space.

Review of variables in input vector

MIR-max requires relatively few input parameters in order to begin training, but these parameters have a strong influence on the model that is eventually produced. Once the set of training variables has been chosen, the bands or states assigned to each training variable need to be defined. This can be straightforward for discrete variables, such as the macro-invertebrates, where the states are already defined (RIVPACS abundance categories). However, for continuous variables it is a more involved process, requiring judgements to be made on the ranges encompassed by each of the bands and the distribution of training samples amongst them.

For the development of a new version of RPDS, the categories used in the old model could have been retained. However, it was decided that the definition of states for the training variables should be revised for the following reasons:

- This approach was recommended in the final report of the previous project (Walley *et al.*, 2002). Although this issue was more closely tied to the definition of states for the Bayesian Belief Network (BBN) element of that project, it would also have an impact on the MIR-max models.
- The new data set differed from the old. It was much larger because it covered a ten-year instead of one-year time period, and included data from Scotland as well as England and Wales. The range of the new data may have required the bands to be updated anyway.
- The availability of Geographical Information System (GIS) data meant that there was the potential to include further environmental training variables, such as upstream catchment characteristics, geology and land cover. Inclusion of any of these would require states to be defined for them.
- The fact that alkalinity is a potential pollutant and was no longer analysed regularly called into question its continued use as a training and classification parameter. However, because the 'natural baseline' alkalinity is such an important factor in river ecology, it was deemed necessary to investigate alternative or surrogate variables.

Reproducing original models based on 1995 data

As a starting point, a model based on the new spring 1995 data was produced and compared to the existing spring RPDS model (also based on data from 1995). It was realised that these models would not be identical because of variations in the training data and the random element inherent in the training algorithm. The difference in the training data was largely because, although the value for alkalinity was based on an average (as described in Section 5), values of other environmental variables were those recorded with individual biological samples. This practice differed from that used

for the original 1995 models, where established values of environmental parameters for each site were used for samples taken in spring and autumn. For this reason, the number of samples in the new model was slightly less than the original (there were 6,039 samples in the original model and 5,339 in the 'clone'). However, the two models were expected to share similar characteristics. Although training of the original RPDS models had taken several days, training time for new models of comparable size had reduced to about half a day, thanks to advances in the power of the PC.

Table 8.1 gives the mutual information values between the 87 input variables and the output classes in the original spring RPDS model, listed in rank order. The two greatest influences on the clustering are the taxa Elmidae and Heptageniidae, followed by the environmental variables alkalinity and the percentage of boulders and cobbles in the substrate. Of the ten highest ranked variables, five are biological and five are environmental (in the corresponding autumn model four are biological and six are environmental). The rest of the environmental parameters are spread throughout the ranking. The strong influence exerted by the macro-invertebrates on the clustering was considered ideal, because the model aims to use the assemblage of the community as the primary means of diagnosing potential pressure.

Table 8.2 gives the corresponding ranking for the 'clone' of the spring RPDS model based on the 'new' 1995 data. The results are not identical but they do show a high degree of consistency. Despite some differences in the order of the rankings, the ten highest ranked variables were the same, with five in the same positions. Furthermore, the top twenties differ by only two variables and the top thirties by only one. The degree of consistency was reassuring and confirmed that the processes used and types of models produced were comparable to those of the earlier project.

Criteria for discretising the continuous variables

Each environmental variable in the original RPDS model was discretised by splitting the range of values into five bands of equal width, regardless of frequency of occurrence in each band.

The distribution of samples between the five equally sized bands is shown in Table 8.3 for each of the 11 environmental variables. For some variables (log slope, boulders and cobbles, pebbles and gravel, and average alkalinity), the distribution of samples between the states is reasonably even, whereas for others, such as width and depth, it is very uneven. Table 8.3 shows a clear tendency for variables with more even distributions to achieve higher rankings.

It was clear that the equal-sized banding scheme could lead to unevenness in the distribution of samples between states, and that this unevenness directly impacted the influence of each variable in the clustering process. Based on these findings, an alternative banding scheme was investigated in which the boundaries of the bands were based on percentiles of the distribution of each variable. An even spread of percentiles was chosen, that is, 20th, 40th, 60th and 80th to produce an even distribution of samples within the bands for each variable regardless of its underlying distribution.

Both models (original RPDS and 'clone') were tested using these bandings and the results are shown in Tables 8.4 and 8.5. There was a high degree of consistency between the two models, as expected, but in each case the environmental variables now dominated the clustering, with environmental variables occupying the first seven places in each case, and the remaining four environmental variables in much higher positions than before. The bias to environmental variables in the model meant that the clustering represented an 'environmental typology', with the characteristics of the

macro-invertebrate community having a lesser impact on the clustering than in previous models. Given that the aim is to produce models which are able to diagnose problems primarily from the biological community, a model biased too much towards environmental variables would be unsuitable. It would be possible to split the biological variables into five abundance levels based on percentage frequency in the dataset. The difficulty with this approach is that it is divorced from practical application because the abundance categories would be different for each taxon. Any changes to the model should represent improvement while retaining practicality.

Table 8.1 Ranking of variables in original spring RPDS model.

Rank	Variable	MI	Rank	Variable	MI
1	Elmidae	0.6209	45	Corixidae	0.2385
2	Heptageniidae	0.609	46	Planariidae	0.2375
3	ALKALINITY	0.588	47	Gyrinidae	0.2343
4	BOULDERS	0.5618	48	Coenagriidae	0.2031
5	Baetidae	0.5395	49	Calopterygidae	0.1965
6	SLOPE	0.5391	50	Dytiscidae	0.1954
7	PEBBLES	0.5373	51	Sialidae	0.1823
8	Hydropsychidae	0.5137	52	DEPTH	0.1802
9	Rhyacophilidae	0.5058	53	WIDTH	0.1802
10	DISCHARGE	0.4945	54	Polycentropodidae	0.1744
11	Perlodidae	0.4855	55	Haliplidae	0.1693
12	Leuctridae	0.4742	56	Psychomyiidae	0.1604
13	Sphaeriidae	0.4718	57	Physidae	0.1573
14	SILT	0.4588	58	Neritidae	0.1299
15	Gammaridae	0.4584	59	Brachycentridae	0.115
16	Asellidae	0.4494	60	Piscicolidae	0.1074
17	DISTANCE FROM SOURCE	0.4322	61	Unionidae	0.0975
18	Leptoceridae	0.4146	62	Perlidae	0.0961
19	Caenidae	0.412	63	Aphelocheiridae	0.0889
20	Nemouridae	0.4006	64	Odontoceridae	0.0871
21	Sericostomatidae	0.4003	65	Dendrocoelidae	0.0856
22	Ephemerellidae	0.398	66	Scirtidae	0.0821
23	Hydrobiidae	0.3889	67	Notonectidae	0.0786
24	Chironomidae	0.3814	68	Molannidae	0.0684
25	Oligochaeta	0.374	69	Viviparidae	0.0577
26	Limnephilidae	0.3655	70	Corophiidae	0.0494
27	Simuliidae	0.3647	71	Platycnemidae	0.049
28	Chloroperlidae	0.3622	72	Cordulegasteridae	0.0386
29	Lepidostomatidae	0.3577	73	Philopotamidae	0.0382
30	Erpobdellidae	0.3379	74	Gerridae	0.0373
31	Glossiphoniidae	0.3337	75	Beraeidae	0.0312
32	Taeniopterygidae	0.3184	76	Capniidae	0.0271
33	Tipulidae	0.3167	77	Astacidae	0.0212
34	Planorbidae	0.3037	78	Hydrometridae	0.0205
35	ALTITUDE	0.2959	79	Naucoridae	0.0141
36	Goeridae	0.289	80	Phryganeidae	0.0138
37	Hydroptilidae	0.288	81	Aeshnidae	0.0126
38	Ancylidae	0.2817	82	Libellulidae	0.0116
39	Hydrophilidae	0.2587	83	Hirudinidae	0.0114
40	SAND	0.2562	84	Nepidae	0.0099
41	Valvatidae	0.2544	85	Dryopidae	0.0092
42	Lymnaeidae	0.245	86	Siphonuridae	0.0084
43	Leptophlebiidae	0.2441	87	Potamanthidae	0.006
44	Ephemeridae	0.242			

Totals

Model MI (Training Variables) 21.638

Taxa MI 17.114

Table 8.2 Ranking of variables in 'clone' spring RPDS model.

Rank	Variable	MI	Rank	Variable	MI
1	Elmidae	0.6517	48	Coenagriidae	0.2307
2	Heptageniidae	0.6308	49	Calopterygidae	0.2052
3	LOG_SLOPE	0.5945	50	Dytiscidae	0.2039
4	BOULDERS_COBBLES	0.5674	51	Sialidae	0.1907
5	Baetidae	0.5585	52	Polycentropodidae	0.1863
6	AVERAGE_ALKALINITY	0.5562	53	Physidae	0.1739
7	PEBBLES_GRAVEL	0.5550	54	Haliplidae	0.1736
8	Rhyacophilidae	0.5260	55	Psychomyiidae	0.1667
9	DISCHARGE_CATEGORY	0.5121	56	Neritidae	0.1439
10	Hydropsychidae	0.5087	57	DEPTH	0.1281
11	Sphaeriidae_Pea_mussels	0.4887	58	Piscicolidae	0.1196
12	LOG_DISTANCE_FROM_SOURCE	0.4825	59	Unionidae	0.1039
13	Perlodidae	0.4737	60	Brachycentridae	0.1038
14	Leuctridae	0.4737	61	Aphelocheiridae	0.1033
15	Asellidae	0.4712	62	Dendrocoelidae	0.1008
16	Gammaridae	0.4537	63	Odontoceridae	0.1004
17	Caenidae	0.4442	64	Perlidae	0.0937
18	SILT_CLAY	0.4274	65	Scirtidae	0.0835
19	Sericostomatidae	0.4258	66	Notonectidae	0.0809
20	Ephemerellidae	0.4229	67	Molannidae	0.0729
21	Leptoceridae	0.4150	68	Corophiidae	0.0636
22	Oligochaeta	0.3922	69	WIDTH	0.0627
23	Simuliidae	0.3901	70	Viviparidae	0.0602
24	Nemouridae	0.3873	71	Platycnemididae	0.0506
25	Chironomidae	0.3859	72	Cordulegasteridae	0.0438
26	Hydrobiidae	0.3795	73	Gerridae	0.0397
27	Lepidostomatidae	0.3772	74	Philopotamidae	0.0379
28	Limnephilidae	0.3760	75	Beraeidae	0.0341
29	Chloroperlidae	0.3659	76	Astacidae	0.0240
30	Glossiphoniidae	0.3578	77	Hydrometridae	0.0234
31	Erpobdellidae	0.3571	78	Capniidae	0.0206
32	Tipulidae	0.3322	79	Phryganeidae	0.0166
33	Taeniopterygidae	0.3250	80	Aeshnidae	0.0135
34	Planorbidae	0.3201	81	Naucoridae	0.0129
35	ALTITUDE	0.3074	82	Nepidae	0.0126
36	Ancylidae	0.3025	83	Libellulidae	0.0111
37	Hydroptilidae	0.2912	84	Dryopidae	0.0107
38	Goeridae	0.2897	85	Pleidae	0.0099
39	Valvatidae	0.2752	86	Hirudinidae	0.0092
40	Ephemeridae	0.2701	87	Potamanthidae	0.0086
41	Hydrophilidae	0.2641	88	Siphonuridae	0.0061
42	Corixidae	0.2579	89	Gomphidae	0.0060
43	Gyrinidae	0.2523	90	Mesoveliidae	0.0060
44	Leptophlebiidae	0.2511	91	Hygrobiiidae	0.0041
45	Lymnaeidae	0.2501	92	Corduliidae	0.0012
46	Planariidae	0.2497	93	Lestidae	0.0000
47	SAND	0.2464			

Totals

Model MI (Training Variables) 22.2486

Taxa MI 17.8090

Table 8.3 Distribution of samples between five equally sized bands for each environmental variable, ordered according to ranking in Table 8.2 (shown in brackets).

Variable	State				
	1	2	3	4	5
LOG SLOPE (3)	399	1,328	2,330	1,182	100
BOULDERS COBBLES (4)	2,452	1,156	925	557	249
AVERAGE ALKALINITY (6)	1787	2,135	1,394	20	3
PEBBLES GRAVEL (7)	916	1,384	1,658	996	385
DISCHARGE CATEGORY (9)	3,340	1,190	529	226	54
LOG DIST FROM SOURCE (12)	6	52	1,457	3,285	539
SILT CLAY (18)	3,733	784	265	166	427
ALTITUDE (35)	3,892	1,147	233	50	17
SAND (47)	3,708	1,231	293	84	23
DEPTH (57)	5,113	181	40	3	2
WIDTH (69)	5,231	100	6	1	1

Table 8.4 Results for original RPDS data using equal percentile bandings.

Rank	Variable	MI	Rank	Variable	MI
1	DISTANCE FROM SOURCE	0.7657	45	Gyrinidae	0.2387
2	DISCHARGE	0.7085	46	Corixidae	0.2383
3	ALKALINITY	0.6953	47	Lymnaeidae	0.2348
4	WIDTH	0.6840	48	Leptophlebiidae	0.2316
5	SLOPE	0.6637	49	Planariidae	0.2288
6	BOULDERS	0.6616	50	Coenagriidae	0.2035
7	SILT	0.6308	51	Calopterygidae	0.1910
8	Elmidae	0.6184	52	Dytiscidae	0.1877
9	Heptageniidae	0.6142	53	Sialidae	0.1757
10	DEPTH	0.5522	54	Haliplidae	0.1689
11	PEBBLES	0.5500	55	Polycentropodidae	0.1679
12	Baetidae	0.5204	56	Physidae	0.1647
13	ALTITUDE	0.5143	57	Psychomyiidae	0.1509
14	Rhyacophilidae	0.5106	58	Neritidae	0.1342
15	Hydropsychidae	0.4991	59	Piscicolidae	0.1161
16	Perlodidae	0.4711	60	Brachycentridae	0.1084
17	SAND	0.4687	61	Perlidae	0.0960
18	Sphaeriidae	0.4642	62	Unionidae	0.0927
19	Leuctridae	0.4613	63	Odontoceridae	0.0897
20	Asellidae	0.4445	64	Aphelocheiridae	0.0883
21	Gammaridae	0.4328	65	Dendrocoelidae	0.0849
22	Caenidae	0.4137	66	Notonectidae	0.0801
23	Leptoceridae	0.4062	67	Scirtidae	0.0786
24	Sericostomatidae	0.3975	68	Molannidae	0.0673
25	Ephemerellidae	0.3857	69	Viviparidae	0.0573
26	Nemouridae	0.3694	70	Platycnemidae	0.0471
27	Hydrobiidae	0.3634	71	Corophiidae	0.0450
28	Chironomidae	0.3619	72	Gerridae	0.0389
29	Simuliidae	0.3605	73	Cordulegasteridae	0.0386
30	Chloroperlidae	0.3601	74	Philopotamidae	0.0366
31	Lepidostomatidae	0.3577	75	Beraeidae	0.0309
32	Oligochaeta	0.3433	76	Capniidae	0.0271
33	Limnephilidae	0.3432	77	Hydrometridae	0.0203
34	Erpobdellidae	0.3343	78	Astacidae	0.0202
35	Glossiphoniidae	0.3183	79	Naucoridae	0.0149
36	Taeniopterygidae	0.3085	80	Phryganeidae	0.0134
37	Tipulidae	0.3081	81	Aeshnidae	0.0124
38	Planorbidae	0.3018	82	Libellulidae	0.0118
39	Ancylidae	0.2957	83	Hirudinidae	0.0112
40	Goeridae	0.2755	84	Nepidae	0.0104
41	Hydroptilidae	0.2710	85	Dryopidae	0.0091
42	Hydrophilidae	0.2614	86	Siphonuridae	0.0083
43	Ephemeridae	0.2456	87	Potamanthidae	0.0055
44	Valvatidae	0.2444			

Totals
 Model MI (Training Variables) 23.6363
 Taxa MI 16.7415

Table 8.5 Results for RPDS 'clone' using equal percentile bandings.

Rank	Variable	MI	Rank	Variable	MI
1	LOG_DISTANCE_FROM_SOURCE	0.7718	48	Leptophlebiidae	0.2468
2	AVERAGE ALKALINITY	0.7572	49	Corixidae	0.2450
3	DISCHARGE_CATEGORY	0.7356	50	Coenagriidae	0.2165
4	WIDTH	0.7141	51	Dytiscidae	0.2033
5	LOG_SLOPE	0.6787	52	Calopterygidae	0.2024
6	SILT_CLAY	0.6749	53	Sialidae	0.1855
7	BOULDERS_COBBLES	0.6632	54	Haliplidae	0.1849
8	Elmidae	0.6187	55	Polycentropodidae	0.1838
9	Heptageniidae	0.6074	56	Physidae	0.1674
10	DEPTH	0.5830	57	Psychomyiidae	0.1510
11	PEBBLES_GRAVEL	0.5601	58	Neritidae	0.1494
12	SAND	0.5451	59	Piscicolidae	0.1203
13	ALTITUDE	0.5327	60	Brachycentridae	0.1034
14	Baetidae	0.5303	61	Unionidae	0.1022
15	Rhyacophilidae	0.5166	62	Aphelocheiridae	0.0981
16	Hydropsychidae	0.5003	63	Dendrocoelidae	0.0968
17	Sphaeriidae_Pea_mussels	0.4674	64	Odontoceridae	0.0953
18	Perlodidae	0.4642	65	Perlidae	0.0916
19	Leuctridae	0.4629	66	Scirtidae	0.0911
20	Gammaridae	0.4569	67	Notonectidae	0.0789
21	Asellidae	0.4476	68	Molannidae	0.0707
22	Ephemerellidae	0.4347	69	Viviparidae	0.0576
23	Caenidae	0.4254	70	Corophiidae	0.0558
24	Leptoceridae	0.4207	71	Platycnemididae	0.0478
25	Sericostomatidae	0.4129	72	Cordulegasteridae	0.0457
26	Nemouridae	0.3709	73	Gerridae	0.0396
27	Lepidostomatidae	0.3686	74	Philopotamidae	0.0381
28	Oligochaeta	0.3682	75	Beraeidae	0.0347
29	Chloroperlidae	0.3679	76	Hydrometridae	0.0222
30	Limnephilidae	0.3563	77	Capniidae	0.0208
31	Hydrobiidae	0.3477	78	Astacidae	0.0207
32	Erpobdellidae	0.3448	79	Phryganeidae	0.0158
33	Chironomidae	0.3443	80	Naucoridae	0.0128
34	Simuliidae	0.3427	81	Nepidae	0.0126
35	Glossiphoniidae	0.3311	82	Aeshnidae	0.0126
36	Taeniopterygidae	0.3138	83	Libellulidae	0.0109
37	Planorbidae	0.3106	84	Dryopidae	0.0106
38	Tipulidae	0.3037	85	Pleidae	0.0103
39	Goeridae	0.2915	86	Hirudinidae	0.0095
40	Ancylidae	0.2861	87	Potamanthidae	0.0071
41	Hydroptilidae	0.2737	88	Mesoveliidae	0.0065
42	Valvatidae	0.2661	89	Siphonuridae	0.0059
43	Lymnaeidae	0.2574	90	Gomphidae	0.0055
44	Hydrophilidae	0.2550	91	Hygrobiiidae	0.0042
45	Gyrinidae	0.2524	92	Corduliidae	0.0011
46	Ephemeridae	0.2487	93	Lestidae	0.0000
47	Planariidae	0.2481			

Totals

Model MI (Training Variables) 24.4248

Taxa MI 17.2084

The effect of state boundaries on the clustering is relatively simple to explain. Suppose the selected boundaries for a particular variable result in the majority of samples falling in one state. The impact of that variable on the clustering process will be minimal because of the inability to discriminate between samples, and therefore between clusters, based on that variable. The clustering algorithm will instead rely on other variables with states that vary more widely between clusters and that provide greater scope for discrimination. The use of equal percentile bandings produced distributions of samples between bands that were practically even, thereby maximising the amount of variation between sample set and clusters. Variables discretised in this way were ideal candidates for good discrimination, and were likely to be variables contributing most to the clustering process.

This posed a difficulty for revision of the RPDS model. A method was required that prevented the environmental variables from dominating the clustering when included with equal percentile banding. The environmental variables needed to be included in a sub-optimal configuration to allow the macro-invertebrates to operate on a 'level playing field'. Two options were suggested:

- Ensure that the distribution of samples between the states was as uniform as possible for all training variables. This would essentially mean modifying the bandings for states of environmental variables until the distribution of samples was 'similar' to that of the macro-invertebrate variables. This was basically what had happened with the original RPDS models, albeit somewhat fortuitously. Although this option had some scope (see below), it was considered to be too *ad hoc* and rejected.
- Reduce the number of environmental variables used in training, many of which were strongly correlated anyway, to allow the macro-invertebrate variables to exert a greater influence. This was considered the more appropriate option.

Modifying bandings for environmental variables

Modifying the bandings is a rather crude method of reducing the influence of environmental variables and it was decided that, in general, this should not be pursued. However, some modification was undertaken to improve the fit to sample distributions and achieve boundary values of more practical use. This would almost certainly weaken the influence of some environmental variables, but as a side effect rather than a goal.

Modifications were made using a piece of software developed for the Bayesian Belief Network in the previous project, designed to show the distribution of samples according to a set of user-defined states. Initially, the user would be able to automatically define a number of equally sized bands, and by using this feature with a larger numbers of states it would be possible to gain some idea of the characteristic of the sample distribution.

Figure 8.2 shows a screenshot of this process for the variable *average alkalinity*, chosen specifically because the distribution appears to be bipolar with a peak around 0-60 mg/l and another around 180-240 mg/l. Bandings based on both equal sub-ranges and equal percentile-banding methods are unsuitable for this kind of feature, and it is more appropriate to define the bandings manually.

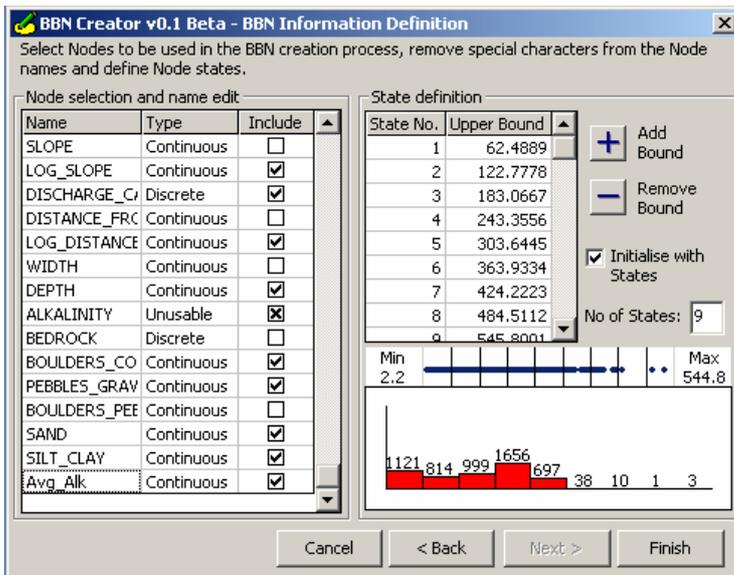


Figure 8.2 Screenshot of division of sample distribution of average alkalinity into nine equally sized bands.

Rationalising environmental variables used in training

The set of environmental variables was reviewed to identify any redundancy, which would allow the set to be reduced with minimal loss of information from the sample. The four variables describing the substrate composition were reduced to two, firstly by combining the values for percentages boulders and cobbles with percentage pebbles and gravel to give a single coarse substrate variable referred to as boulders and pebbles combined; and secondly by removing the percentage sand variable. The reasoning behind this was that the main habitats for the majority of macroinvertebrates are associated with coarse or very fine substrates. Although the removal of sand was a major change, it was reasoned that, being part of a composite value, information on the percentage sand would be represented indirectly in the values of remaining substrate variables. Log distance from source was also selected for removal, given its likely correlation with other variables.

Further reductions were based on rankings derived by three different methods. The first two were the mean of the results of pair-wise testing for correlation and mutual information respectively. The third set of results was the mutual information rankings for environmental variables in an RPDS ‘clone’ model produced using only the macroinvertebrate variables in training (this would indicate the influence on ‘communities typology’ defined by the model, rather than a statistic based on tests with individual taxa as in the other tests). The rankings are given in Table 8.6.

Table 8.6 Three rankings (based on the mean of pair-wise correlation, MI tests and MI values based on a MIR-max model using only macroinvertebrate taxa) and their mean.

	Ranking for mean correlation	Ranking for mean MI	Ranking for MIs for taxa-only model	Mean of rankings
AVERAGE ALKALINITY	1	2	1	1.33
LOG_SLOPE	2	5	4	3.67
BOULDERS_PEBBLES_COMBINED	4	3	7	4.67
DEPTH	3	8	3	4.67
LOG_DISTANCE_FROM_SOURCE*	7	10	2	6.33
SILT_CLAY	5	6	10	7.00
ALTITUDE	8	9	5	7.33
WIDTH	10	11	6	9.00
DISCHARGE_CATEGORY	9	12	12	11.00
BEDROCK	13	13	13	13.00

*Log distance from source is shown because although it was removed from this phase of testing, it was returned later.

Spring 1995 models were produced with all the environmental variables, and with seven, five, four, three and none. Each model was produced initially with average alkalinity in the dataset, and then reproduced with calcareous geology in its place. So that a fair comparison could be made between the two, the training data was reduced from the original RPDS 'clone' dataset of 5,339 samples to a subset of 4,349 samples that contained a value for both variables. The results of the tests of models that included alkalinity in the training set are shown in Table 8.7, and for those that included calcareous geology in Table 8.8. For brevity, only the 30 highest ranked variables are shown, and in each list, the environmental variables that were excluded from the input vector are italicised.

Taking the tests with average alkalinity first, Table 8.7, the first column represents the equivalent model to the RPDS 'clone' of Table 8.5, but for the reduced dataset. The highest rankings are dominated by environmental variables, as before. The other columns in Table 8.7 give the rankings for the models with seven, five, four and three environmental variables. It is clear that macroinvertebrates are not in the majority in the top ten or top three rankings until environmental variables have been reduced to just four.

The corresponding results for models with alkalinity replaced by calcareous geology, Table 8.8, show that calcareous geology performs a similar role in the models to that of alkalinity. Changes in the rankings and MI values achieved for each model are similar.

Table 8.7 Results for environmental training variables reduction tests using alkalinity.

Rank	All RPDS Variables		Seven Environmental Variables		Five Environmental Variables		Four Environmental Variables		Three Environmental Variables	
	Variable	MI	Variable	MI	Variable	MI	Variable	MI	Variable	MI
1	LOG_DISTANCE_FROM_SOURCE	0.7600	BOULDERS_PEBBLES_COMBINED	0.7530	BOULDERS_PEBBLES_COMBINED	0.7794	AVERAGE_ALKALINITY	0.7067	AVERAGE_ALKALINITY	0.6932
2	DISCHARGE_CATEGORY	0.7394	AVERAGE_ALKALINITY	0.6898	SILT_CLAY	0.6966	Heptageniidae	0.6747	Elmidae	0.6737
3	WIDTH	0.7035	LOG_SLOPE	0.6714	AVERAGE_ALKALINITY	0.6887	Elmidae	0.6634	Heptageniidae	0.6714
4	BOULDERS_PEBBLES_COMBINED	0.6926	SILT_CLAY	0.6707	Heptageniidae	0.6752	BOULDERS_PEBBLES_COMBINED	0.6520	BOULDERS_PEBBLES_COMBINED	0.6346
5	AVERAGE_ALKALINITY	0.6879	Elmidae	0.6558	Elmidae	0.6525	LOG_SLOPE	0.6485	LOG_SLOPE	0.6197
6	LOG_SLOPE	0.6848	Heptageniidae	0.6516	LOG_SLOPE	0.6451	Baetidae	0.5900	Baetidae	0.6094
7	SILT_CLAY	0.6616	WIDTH	0.6394	Baetidae	0.5657	DEPTH	0.5893	Hydropsychidae	0.5835
8	BOULDERS_COBBLES	0.6566	DEPTH	0.5995	Hydropsychidae	0.5595	Hydropsychidae	0.5622	Rhyacophilidae	0.5505
9	Elmidae	0.6377	Baetidae	0.5630	DEPTH	0.5575	Rhyacophilidae	0.5364	Sphaeriidae_Pea_mussels	0.5337
10	Heptageniidae	0.6371	ALTITUDE	0.5588	Rhyacophilidae	0.5537	Sphaeriidae_Pea_mussels	0.5269	GEO_CALC	0.5239
11	DEPTH	0.5868	Hydropsychidae	0.5513	GEO_CALC	0.5295	GEO_CALC	0.5237	Gammaridae	0.5203
12	ALTITUDE	0.5686	GEO_CALC	0.5479	Sphaeriidae_Pea_mussels	0.5146	Gammaridae	0.5121	Perlodidae	0.5148
13	GEO_CALC	0.5478	Rhyacophilidae	0.5400	Gammaridae	0.5079	Asellidae	0.5092	Asellidae	0.5081
14	PEBBLES_GRAVEL	0.5428	Sphaeriidae_Pea_mussels	0.5102	Perlodidae	0.4984	Perlodidae	0.5039	Leuctridae	0.5021
15	Baetidae	0.5398	DISCHARGE_CATEGORY	0.5028	Leuctridae	0.4956	Leuctridae	0.4966	Caenidae	0.4731
16	Rhyacophilidae	0.5314	LOG_DISTANCE_FROM_SOURCE	0.4974	Asellidae	0.4891	Ephemerellidae	0.4779	Sericostomatidae	0.4702
17	SAND	0.5192	Gammaridae	0.4963	Ephemerellidae	0.4755	SILT_CLAY	0.4677	Ephemerellidae	0.4700
18	Hydropsychidae	0.5139	Leuctridae	0.4961	BOULDERS_COBBLES	0.4750	Caenidae	0.4661	SILT_CLAY	0.4567
19	Sphaeriidae_Pea_mussels	0.4826	Perlodidae	0.4912	Caenidae	0.4735	Sericostomatidae	0.4639	Leptoceridae	0.4460
20	Leuctridae	0.4792	BOULDERS_COBBLES	0.4786	Leptoceridae	0.4582	Simuliidae	0.4413	Simuliidae	0.4343
21	Perlodidae	0.4776	Caenidae	0.4764	Sericostomatidae	0.4479	Leptoceridae	0.4353	Erpobdellidae	0.4269
22	Asellidae	0.4727	Asellidae	0.4677	LOG_DISTANCE_FROM_SOURCE	0.4272	BOULDERS_COBBLES	0.4269	Oligochaeta	0.4237
23	Gammaridae	0.4691	Ephemerellidae	0.4599	DISCHARGE_CATEGORY	0.4148	LOG_DISTANCE_FROM_SOURCE	0.4202	BOULDERS_COBBLES	0.4177
24	Caenidae	0.4443	Sericostomatidae	0.4509	Erpobdellidae	0.4148	WIDTH	0.4191	Chironomidae	0.4117
25	Ephemerellidae	0.4393	Leptoceridae	0.4412	WIDTH	0.4138	Erpobdellidae	0.4141	Lepidostomatidae	0.4034
26	Sericostomatidae	0.4347	Hydrobiidae	0.4193	Lepidostomatidae	0.4100	DISCHARGE_CATEGORY	0.4141	Hydrobiidae	0.4033
27	Leptoceridae	0.4107	Simuliidae	0.4191	Simuliidae	0.4089	Hydrobiidae	0.4124	Nemouridae	0.4031
28	Lepidostomatidae	0.4031	Oligochaeta	0.4100	Hydrobiidae	0.4072	Oligochaeta	0.4083	Limnephilidae	0.4031
29	Oligochaeta	0.4000	Nemouridae	0.4047	Oligochaeta	0.4011	Nemouridae	0.4016	LOG_DISTANCE_FROM_SOURCE	0.3939
30	Simuliidae	0.3961	Lepidostomatidae	0.4023	Chloroperlidae	0.3956	Lepidostomatidae	0.4008	Chloroperlidae	0.3914
	Totals		Totals		Totals		Totals		Totals	
	Model MI (Training Variables)	25.1088	Model MI (Training Variables)	23.137	Model MI (Training Variables)	22.179	Model MI (Training Variables)	21.603	Model MI (Training Variables)	21.154
	Taxa MI	17.9975	Taxa MI	18.554	Taxa MI	18.812	Taxa MI	19.006	Taxa MI	19.207
	Original RPDS Variables MI	25.1088	Original RPDS Variables MI	24.599	Original RPDS Variables MI	24.246	Original RPDS Variables MI	24.131	Original RPDS Variables MI	23.895

Table 8.8 Results for environmental training variables reduction tests using calcareous geology.

Rank	All RPDS Variables		Seven Environmental Variables		Five Environmental Variables		Four Environmental Variables		Three Environmental Variables	
	Variable	MI	Variable	MI	Variable	MI	Variable	MI	Variable	MI
1	LOG_DISTANCE_FROM_SOURCE	0.7792	BOULDERS_PEBBLES_COMBINED	0.7459	BOULDERS_PEBBLES_COMBINED	0.7778	GEO_CALC	0.7099	GEO_CALC	0.7068
2	DISCHARGE_CATEGORY	0.7536	GEO_CALC	0.6894	GEO_CALC	0.7037	Heptageniidae	0.6757	Elmidae	0.6884
3	WIDTH	0.7265	Heptageniidae	0.6763	SILT_CLAY	0.6816	Elmidae	0.6661	Heptageniidae	0.6785
4	BOULDERS_PEBBLES_COMBINED	0.6859	LOG_SLOPE	0.6747	Heptageniidae	0.6664	BOULDERS_PEBBLES_COMBINED	0.6521	BOULDERS_PEBBLES_COMBINED	0.6250
5	LOG_SLOPE	0.6842	SILT_CLAY	0.6719	Elmidae	0.6465	LOG_SLOPE	0.6306	LOG_SLOPE	0.6205
6	GEO_CALC	0.6759	Elmidae	0.6630	LOG_SLOPE	0.6393	Baetidae	0.5951	Baetidae	0.5979
7	BOULDERS_COBBLES	0.6576	WIDTH	0.6336	Baetidae	0.5943	Hydropsychidae	0.5755	Hydropsychidae	0.5889
8	Heptageniidae	0.6525	DEPTH	0.6001	DEPTH	0.5641	DEPTH	0.5681	Sphaeriidae_Pea_mussels	0.5423
9	SILT_CLAY	0.6438	Baetidae	0.5769	Hydropsychidae	0.5602	Rhyacophilidae	0.5354	Rhyacophilidae	0.5369
10	Elmidae	0.6376	ALTITUDE	0.5740	AVERAGE_ALKALINITY	0.5444	AVERAGE_ALKALINITY	0.5292	AVERAGE_ALKALINITY	0.5353
11	DEPTH	0.5811	AVERAGE_ALKALINITY	0.5499	Rhyacophilidae	0.5320	Sphaeriidae_Pea_mussels	0.5237	Gammaridae	0.5184
12	AVERAGE_ALKALINITY	0.5547	Hydropsychidae	0.5422	Sphaeriidae_Pea_mussels	0.5137	Asellidae	0.5101	Perlodidae	0.5115
13	Baetidae	0.5456	Rhyacophilidae	0.5347	Perlodidae	0.5108	Gammaridae	0.5086	Leuctridae	0.5043
14	ALTITUDE	0.5421	DISCHARGE_CATEGORY	0.5028	Leuctridae	0.5055	Leuctridae	0.5069	Asellidae	0.4984
15	PEBBLES_GRAVEL	0.5386	Leuctridae	0.5019	Gammaridae	0.4975	Perlodidae	0.4990	Ephemerellidae	0.4832
16	Hydropsychidae	0.5206	LOG_DISTANCE_FROM_SOURCE	0.4980	Asellidae	0.4908	Caenidae	0.4684	Caenidae	0.4708
17	SAND	0.5175	Gammaridae	0.4972	Ephemerellidae	0.4713	Sericostomatidae	0.4675	Sericostomatidae	0.4618
18	Rhyacophilidae	0.5157	Sphaeriidae_Pea_mussels	0.4956	BOULDERS_COBBLES	0.4669	Ephemerellidae	0.4659	SILT_CLAY	0.4487
19	Leuctridae	0.4898	Perlodidae	0.4936	Caenidae	0.4637	SILT_CLAY	0.4521	Leptoceridae	0.4371
20	Sphaeriidae_Pea_mussels	0.4887	Asellidae	0.4875	Sericostomatidae	0.4547	Leptoceridae	0.4341	Oligochaeta	0.4326
21	Perlodidae	0.4800	Caenidae	0.4754	Leptoceridae	0.4343	Simuliidae	0.4340	Hydrobiidae	0.4298
22	Gammaridae	0.4796	BOULDERS_COBBLES	0.4697	LOG_DISTANCE_FROM_SOURCE	0.4320	BOULDERS_COBBLES	0.4334	BOULDERS_COBBLES	0.4273
23	Asellidae	0.4644	Ephemerellidae	0.4634	Simuliidae	0.4295	Erpobdellidae	0.4249	Simuliidae	0.4251
24	Ephemerellidae	0.4485	Sericostomatidae	0.4455	WIDTH	0.4243	LOG_DISTANCE_FROM_SOURCE	0.4230	Erpobdellidae	0.4112
25	Caenidae	0.4473	Leptoceridae	0.4213	Erpobdellidae	0.4201	Hydrobiidae	0.4162	Lepidostomatidae	0.4068
26	Sericostomatidae	0.4324	Simuliidae	0.4175	DISCHARGE_CATEGORY	0.4186	Oligochaeta	0.4139	Chironomidae	0.4012
27	Leptoceridae	0.4144	Erpobdellidae	0.4072	Hydrobiidae	0.4148	Lepidostomatidae	0.4029	Nemouridae	0.3972
28	Nemouridae	0.3966	Oligochaeta	0.4046	Lepidostomatidae	0.4005	WIDTH	0.4003	Limnephilidae	0.3954
29	Lepidostomatidae	0.3945	Hydrobiidae	0.4023	Oligochaeta	0.3999	DISCHARGE_CATEGORY	0.3996	LOG_DISTANCE_FROM_SOURCE	0.3954
30	Hydrobiidae	0.3885	Lepidostomatidae	0.4006	Nemouridae	0.3969	Nemouridae	0.3976	DISCHARGE_CATEGORY	0.3935
Totals			Totals			Totals			Totals	
	Model MI (Training Variables)	25.0976	Model MI (Training Variables)	23.141	Model MI (Training Variables)	22.178	Model MI (Training Variables)	21.611	Model MI (Training Variables)	21.121
	Taxa MI	17.9976	Taxa MI	18.551	Taxa MI	18.811	Taxa MI	19.05	Taxa MI	19.169
	Original RPDS Variables MI	24.9764	Original RPDS Variables MI	24.457	Original RPDS Variables MI	24.111	Original RPDS Variables MI	23.899	Original RPDS Variables MI	23.765

Analysis of the models by ranking was complemented by an analysis designed to measure the extent of variability of the macroinvertebrate community represented by the clusters. The cluster variability of macroinvertebrates was measured using:

- The mean standard deviation of the clusters (the standard deviation of each variable in the cluster then averaged over all the variables).
- The mean entropy of the clusters (a measure of the uncertainty associated with the distribution of each variable in the cluster, and then summed over all the variables. Its theoretical maximum is when all states are equally likely for each variable and its theoretical minimum (of zero) when each variable occupies a single state).

Variability in the variables themselves was measured using:

- The mean standard deviation of each variable (the standard deviation of each variable when averaged over the clusters).
- The mean entropy of each variable (a measure of the uncertainty associated with the distribution of each single variable, summed over the clusters. Its theoretical maximum is when all states are equally likely, and its theoretical minimum (of zero) when the variable occupies a single state).

Three models were produced for each set of test parameters and the best model chosen. Table 8.9 shows the results for tests when alkalinity was included and when it was replaced by calcareous geology, along with the model based on macroinvertebrates alone ('taxa only').

As expected, both sets of tests indicated that the summed mutual information for the taxa ('taxa MI') increases as the number of environmental variables decreases, reflecting the 'optimisation' of clustering for macroinvertebrates with successive removal of environmental parameters. Measures of variability (entropy, standard deviation) improved correspondingly with each removal, both within the clusters and among the variables themselves.

Models that included alkalinity tended to have marginally higher MI values than their counterparts with calcareous geology instead. They also tended to have clusters with slightly lower variability as measured by entropy and standard deviation. These results could not confirm that the calcareous geology variable was performing exactly the same role as alkalinity, but similar characteristics were certainly displayed. When alkalinity was substituted by calcareous geology, the overall quality of the model diminished slightly, indicating that it would be better to retain alkalinity if possible, but calcareous geology was a viable alternative.

Table 8.9 Mean entropy and standard deviation for each cluster and each variable based on analysis of macroinvertebrate variables, plus MI for whole model.

	Mean Cluster Entropy	Mean Variable Entropy	Mean Cluster SD	Mean Variable SD	Taxa MI
AllEnv_Alk	26.2041	0.3276	21.5875	0.2698	17.9975
7Env_Alk	25.8940	0.3237	21.4355	0.2679	18.5544
5Env_Alk	25.6608	0.3208	21.3153	0.2664	18.8120
4Env_Alk	25.5237	0.3190	21.2148	0.2652	19.0063
3Env_Alk	25.3680	0.3171	21.1741	0.2647	19.2071
AllEnv_Calc	26.2599	0.3282	21.6151	0.2702	17.9976
7Env_Calc	25.8243	0.3228	21.3937	0.2674	18.5512
5Env_Calc	25.6505	0.3206	21.3255	0.2666	18.8112
4Env_Calc	25.4565	0.3182	21.2517	0.2656	19.0505
3Env_Calc	25.3957	0.3174	21.2139	0.2652	19.1689
Taxa only	25.1106	0.3139	21.1278	0.2641	19.5279

Final test model

The results of the preliminary tests suggested that a model with four or five environmental variables would represent a reasonable compromise between the opposing goals of maximising the representation of environmental characteristics of the site and maximising the influence of macroinvertebrates in the eventual clusters.

A final test model was adopted with five environmental variables. The aim was to employ environmental variables that covered the widest range of influences on habitat, including chemical composition, flow, substrate composition, river dimension and temperature. Hence, the final five were alkalinity/calcareous geology (indicative of chemical conditions), slope (indicative of flow velocity), boulders and pebbles (indicative of substrate composition), distance from source (indicative of river size) and altitude (indicative of temperature).

Table 8.10 shows the results of tests undertaken with the final model containing these five environmental variables, both with alkalinity and when replaced by calcareous geology. Compared to the previous results for five environmental variables (Tables 8.7 and 8.8), the most noticeable change is the improved ranking of Heptageniidae and Elmidae, although Baetidae has dropped. These changes could signify a slight improvement in the influence of macroinvertebrates in the model. The fall of boulders and pebbles appears to confirm that its previous high ranking was a result of its strong correlation with the other substrate training variables, silt and clay.

Table 8.10 Results for revised five environmental variables using alkalinity (left) and calcareous geology (right).

Rank	Variable	MI	Rank	Variable	MI
1	LOG_SLOPE	0.6977	1	LOG_SLOPE	0.7071
2	AVERAGE_ALKALINITY	0.6956	2	GEO_CALC	0.6945
3	Elmidae	0.6661	3	Elmidae	0.6707
4	Heptageniidae	0.6574	4	Heptageniidae	0.6674
5	BOULDERS_PEBBLES_COMBINEI	0.6493	5	BOULDERS_PEBBLES_COMBINEI	0.6401
6	LOG_DISTANCE_FROM_SOURCE	0.6321	6	LOG_DISTANCE_FROM_SOURCE	0.6193
7	ALTITUDE	0.5958	7	Baetidae	0.5828
8	Baetidae	0.5711	8	Hydropsychidae	0.5677
9	Hydropsychidae	0.5587	9	ALTITUDE	0.5630
10	GEO_CALC	0.5445	10	AVERAGE_ALKALINITY	0.5470
11	Rhyacophilidae	0.5378	11	Rhyacophilidae	0.5468
12	Sphaeriidae_Pea_mussels	0.5108	12	Sphaeriidae_Pea_mussels	0.5069
13	Leuctridae	0.5048	13	Perlodidae	0.5064
14	Perlodidae	0.5018	14	Asellidae	0.5020
15	Gammaridae	0.4967	15	Leuctridae	0.5000
16	Asellidae	0.4933	16	Gammaridae	0.4973
17	Ephemerellidae	0.4879	17	DISCHARGE_CATEGORY	0.4691
18	Caenidae	0.4787	18	Caenidae	0.4675
19	DISCHARGE_CATEGORY	0.4770	19	Ephemerellidae	0.4596
20	SILT_CLAY	0.4683	20	Sericostomatidae	0.4589
21	WIDTH	0.4524	21	SILT_CLAY	0.4581
22	Leptoceridae	0.4435	22	WIDTH	0.4528
23	Sericostomatidae	0.4420	23	Simuliidae	0.4400
24	BOULDERS_COBBLES	0.4309	24	Leptoceridae	0.4362
25	Simuliidae	0.4189	25	BOULDERS_COBBLES	0.4255
26	Nemouridae	0.4142	26	Oligochaeta	0.4145
27	Oligochaeta	0.4071	27	Hydrobiidae	0.4073
28	Limnephilidae	0.4069	28	Erpobdellidae	0.4054
29	DEPTH	0.4062	29	Nemouridae	0.4024
30	Lepidostomatidae	0.4044	30	Lepidostomatidae	0.3997
Model MI (Training Variables)		22.1137	Model MI (Training Variables)		22.1276
Taxa MI		18.8432	Taxa MI		18.9036
Original RPDS Variables MI		24.3575	Original RPDS Variables MI		24.2111

Final models with full dataset based on average alkalinity

The results of tests obtained with all preliminary models were presented at a meeting with the Environment Agency project manager in May 2007, after which it was confirmed that the full models should be trained on the basis of the five parameters of the final preliminary model.

Spring and autumn models with the five parameters in the input vector were then trained using the full dataset, incorporating SEPA data and the years 1995-2004. Two models were produced for each season, one based on average alkalinity and one on calcareous geology.

Tables 8.11 and 8.12 show the distribution of samples by region and year in the spring and autumn models based on average alkalinity. As with the preliminary models, although the value for alkalinity was based on an average, values of the other four environmental variables were those recorded with individual samples, differing from the practice used for the original 1995 models. For this reason, the number of samples in the new models was slightly fewer (5,604 in spring, and 5,794 in autumn) than the 6,039 samples in each season for the original 1995 models. Virtually none of the SEPA samples prior to 2003 could be used because environmental data was missing.

Table 8.11 Distribution of samples by region and year for spring MIR-max model based on average alkalinity.

Year	EA									SEPA				Grand Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	Total	E	N	W	Total	
1995	679	668	277	1125	435	1098	510	812	5604	0	0	0	0	5604
1996	636	245	2	877	198	190	43	57	2248	0	0	0	0	2248
1997	671	201	29	629	213	376	126	13	2258	0	0	0	0	2258
1998	624	242	262	1023	235	170	202	53	2811	0	0	0	0	2811
1999	627	238	235	572	272	139	282	17	2382	0	0	0	0	2382
2000	505	719	709	1064	472	1128	497	815	5909	0	0	0	0	5909
2001	14	22	13	80	84	10	86	12	321	0	0	0	0	321
2002	374	312	343	622	220	378	242	292	2783	0	54	0	54	2837
2003	366	311	369	552	214	354	227	345	2738	629	262	0	891	3629
2004	338	290	392	415	198	338	208	299	2478	679	456	498	1633	4111
Total	4834	3248	2631	6959	2541	4181	2423	2715	29532	1308	772	498	2578	32110

Table 8.12 Distribution of samples by region and year for autumn MIR-max model based on average alkalinity.

Year	EA									SEPA				Grand Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	Total	E	N	W	Total	
1995	673	671	274	1355	413	1100	497	811	5794	0	0	0	0	5794
1996	669	291	15	657	207	159	108	3	2109	0	0	0	0	2109
1997	630	227	106	980	234	170	83	6	2436	0	0	0	0	2436
1998	641	250	146	954	242	152	197	2	2584	0	0	0	0	2584
1999	496	233	314	539	273	120	282	3	2260	0	0	0	0	2260
2000	541	512	492	802	320	978	484	697	4826	0	0	0	0	4826
2001	231	210	148	581	159	133	105	32	1599	0	0	0	0	1599
2002	374	299	356	477	196	335	217	302	2556	0	43	0	43	2599
2003	367	302	351	527	212	328	238	311	2636	608	234	0	842	3478
2004	330	257	387	423	201	300	195	289	2382	610	365	415	1390	3772
Total	4952	3252	2589	7295	2457	3775	2406	2456	29182	1218	642	415	2275	31457

The total number of samples in the spring and autumn models was 32,110 and 31,457 respectively, representing more than a five-fold increase compared to the original model. Data in the models represent roughly 74 per cent of the total number of biological samples available (Tables 4.4 and 4.5) for spring and autumn. The corresponding number of sites in each season at which samples were taken is shown in Table 8.13, representing 67 per cent of the total number of sites available (Table 4.6) for spring and autumn.

Table 8.13 Distribution of sites for spring and autumn MIR-max models based on average alkalinity.

	EA									SEPA				Grand Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	Total	E	N	W	Total	
Spring	697	797	859	1424	593	1397	540	966	7273	716	620	496	1832	9105
Autumn	698	823	836	1469	571	1342	547	918	7204	711	526	407	1644	8848

Final models with full data based on calcareous geology

Although the new version of RPDS was to incorporate the model above based on average alkalinity, the corresponding model based on calcareous geology is summarised in Tables 8.14-8.16. Geological data was acquired as part of the GIS data, and this was only requested for biological sites with spatial locations validated using the procedures described in Section 5.4. The number of samples and the number of sites represented in this model is therefore slightly less than in the model based on average alkalinity.

Table 8.14 Distribution of samples by region and year for spring MIR-max model based on calcareous geology.

Year	EA								SEPA			Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	E	N	W	
1995	507	548	202	876	389	958	432	723	0	0	0	4635
1996	513	243	2	646	170	175	70	59	0	0	0	1878
1997	518	265	23	499	194	307	160	13	0	0	0	1979
1998	461	307	182	768	210	156	195	28	0	0	0	2307
1999	465	326	174	440	240	157	260	19	0	0	0	2081
2000	378	640	516	844	430	987	402	727	0	0	0	4924
2001	11	48	9	53	76	18	83	11	0	0	0	309
2002	298	334	265	534	193	352	239	271	0	7	0	2493
2003	287	304	287	432	187	337	250	300	329	85	0	2798
2004	266	304	318	333	178	285	232	263	333	197	176	2885
Total	3704	3319	1978	5425	2267	3732	2323	2414	662	289	176	26289

Table 8.15 Distribution of samples by region and year for autumn MIR-max model based on calcareous geology.

Year	EA								SEPA			Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	E	N	W	
1995	515	607	200	971	373	967	408	721	0	0	0	4762
1996	509	295	11	486	189	117	127	14	0	0	0	1748
1997	501	255	73	731	205	112	84	6	0	0	0	1967
1998	487	298	101	725	216	145	196	4	0	0	0	2172
1999	375	308	226	431	249	147	253	15	0	0	0	2004
2000	404	460	364	623	286	847	406	646	0	0	0	4036
2001	184	222	108	463	136	155	120	30	0	0	0	1418
2002	295	320	277	385	174	315	219	291	0	6	0	2282
2003	288	296	278	425	196	306	263	279	320	79	0	2730
2004	269	252	314	338	188	254	226	261	292	145	149	2688
Total	3827	3313	1952	5578	2212	3365	2302	2267	612	230	149	25807

Table 8.16 Distribution of sites for spring and autumn MIR-max models based on calcareous geology.

	EA									SEPA				Grand Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	Total	E	N	W	Total	
Spring	587	905	656	1093	522	1276	675	849	6563	347	212	176	735	7298
Autumn	588	892	638	1112	512	1226	637	881	6486	345	173	146	664	7150

9 Revision and evaluation of RPDS (River Pressure Diagnostic System)

Introduction

The original scope of the project anticipated incorporating the new MIR-max and BBN models into the existing RPDS and RPBBN systems respectively, and then developing a combined system with new functionality for each component. However, because of the time taken to construct the project database, the combined system was abandoned and the systems were developed as separate entities. As a result, both software systems were modified to incorporate their respective new model, and some limited testing and evaluation was completed. It was not possible to implement the additional functionality for either system within the original project because of the severely shortened timescales. This section describes the modifications made to the RPDS system to upgrade it to RPDS 3.0, the quantitative evaluation undertaken and the additional functionality required in the future.

Development of RPDS 3.0 software

Following the generation of clusters from the spring and autumn datasets described in the previous section, the MIR-max algorithm was used to order the clusters for each season in two-dimensional space to produce hexagonal output maps of side-length 10, 15 and 20 clusters. Each map was rotated to align as closely as possible with the originals in RPDS 2.0, for easy comparison.

A number of modifications were made to the MIR-max output, project database and RPDS 2.0 software to accommodate the new spring and autumn models:

- Data for the diagnostic variables (chemical statistics, stresses, GIS and flow data) was appended to the clustered spring and autumn datasets.
- The datasets and MIR-max model data required modification to fit into the RPDS 2.0 database.
- Streamlining of the database querying operations was required to reduce excessive response times.
- The geographical map panel was revised to incorporate Scotland as well as England and Wales.

Qualitative evaluation of RPDS 3.0

Qualitative evaluation of RPDS 3.0 was undertaken firstly by visual comparison with RPDS 2.0 of the output maps for several variables in the spring and autumn models,

secondly by examination of the geographic locations of samples in particular clusters, and thirdly by interpreting the output maps for the new variables.

Comparison with RPDS 2.0

Figure 9.1(a) shows the hexagonal output maps (known as Hex 10 because of the 10 locations on each edge) of RPDS 2.0 for three of the variables in the spring model: number of families, BOD and Elmidae. The corresponding maps of RPDS 3.0 are shown in Figure 9.1(b). Comparing the shapes of the two sets of maps, it is immediately clear that the ordering produced for the new model is almost identical to the old, with the 'gaps' appearing in virtually identical locations. From the colouring of the maps it is also evident that the distribution of the three variables across the new clusters is also similar to the distribution across the old. Note that the scales adopted in RPDS 3.0 are different to those used in RPDS 2.0. Different ranges have been used for the linear scales for number of families and BOD, while for taxa the logarithmic abundance scale used in RPDS 2.0 has been replaced by a nonlinear scale indicating actual counts. Comparison of the maps shows that variation across the 'quality' gradient in the models of RPDS 2.0 has been reproduced in the new models of RPDS 3.0.

The models in RPDS 2.0 also exhibited variation across a 'site type' gradient, and Figure 9.2(a) shows the Hex 10 output maps from the autumn model for the three variables: ASPT, pH and Heptageniidae. Their counterparts in RPDS 3.0 are shown in Figure 9.2(b). A high degree of similarity is again apparent between the ordering of clusters in the two models. After allowing for changes in scale, the colouring of the maps indicates that the distributions of these three variables across the clusters have also been reproduced well.

Examination of geographical locations

Because the dataset for the new model contained samples from SEPA as well as the Environment Agency, the geographic map panel in RPDS 3.0 was amended to incorporate Scotland as well as England and Wales. The map panels of RPDS 2.0 and RPDS 3.0 are shown in Figure 9.3(a) and (b) respectively. The Hex10 output map displays number of families for the spring model in each case, and a cluster has been chosen in a similar location in each that contains samples with a low number of families. The geographical distribution of the sites is illustrated in each map panel, from which it is clear that samples in the old model are an approximate subset of those in the new. This is more apparent from Figures 9.4(a) and (b), where ASPT is displayed on the Hex10 maps and the clusters chosen contain samples with high ASPT values. The high altitudes of the corresponding sites are clearly shown on each map panel, with a large proportion of Scottish sites evident in the map panel of RPDS 3.0.

Interpretation of maps of new variables

Simple geology (Section 6), land cover (Section 6) and flow variables (Section 7) were incorporated in the new models and some qualitative evaluation is given in Figures 9.5-9.9.

Figure 9.5 shows output maps for the geological variables (percentage of upstream catchment categorised as calcareous, siliceous, peat or salt) for the spring model. Only maps for the spring model are shown because maps for the autumn model are similar. Calcareous and siliceous geology dominate the categories, with over 90 per cent occurring in many clusters for calcareous and a few clusters for siliceous. Comparison of the output map for calcareous geology with that of pH (Fig 9.2(b)(i)) indicates a good correlation between the percentage of calcareous geology and

alkaline conditions. There is a similarly good correlation between the percentage of peat (up to around 15 per cent) and acidic conditions. The percentage categorised as salt is usually very low and reaches a maximum of around one per cent in only a few clusters.

The new dataset includes percentage cover in the upstream catchment of more than twenty land cover categories. Output maps for a selection of six of the categories in the autumn model are shown in Figure 9.6. Only maps for the autumn model are shown because maps for the spring model are similar. Catchments with a high proportion of urban (up to around 7.5 per cent) and suburban (up to around 15 per cent) cover tend to be those with poorer quality water; there is a good correlation with maps of the number of taxa (Figure 9.1(b)(1)), BOD5 (Fig 9.1(b)(ii)) and ASPT (Fig 9.2(b)(i)). Catchments with a high proportion of arable land (up to around 60 per cent) tend to be those with calcareous geology (Figure 9.5(i)) rather than any other. The map of improved grassland shows some clear structure, but further investigation is needed to identify correlations. While the map of broad-leaved and mixed woodland shows little structure, the map of coniferous woodland shows clear correlation with those for siliceous (Fig 9.5(ii)) and peaty (Fig 9.5 (iii)) geology.

Figures 9.7 to 9.9 provide qualitative information about the two flow variables incorporated in the new model. Figures 9.7(a) and (b) show Hex10 maps for percentage impact at Q95 in the spring and autumn models respectively. Percentage impact at Q95 is defined as

$$\frac{\text{Natural Flow} - \text{Influenced Flow}}{\text{Natural Flow}} \times 100,$$

so that positive values imply a reduction from natural flow (for example by abstraction), while negative values imply an increase (for example from discharges). Preliminary interpretation of both maps suggests a weak negative correlation with distance from source, which would be expected because abstractions are likely to occur nearer the source than discharges.

While the percentage impact at Q95 provides information about the site, being based on long-term average data, the other flow variable incorporated into the new model was designed to provide information about flow conditions at the time of sampling. Based on 30 years' monthly flow data at a network of gauged sites, flow condition was calculated on a scale of [0,1] with zero for driest and one for wettest, based on a time period preceding the sample of one, two, three, six, 12 or 24 months. The distributions of the cluster averages of these values are shown for the spring and autumn models in the Hex10 output maps of Figures 9.8 and 9.9. Increasing structure is evident in both sets of maps for time periods of more than two months prior to the sample date. Preliminary interpretation suggests a positive correlation with ASPT: for comparison see Figure 9.2(b)(i). This would be expected following the analysis in Section 7, which indicated the increased prevalence of high scoring taxa in wetter conditions and low scoring taxa in drier conditions.

An initial qualitative evaluation suggests that of the two flow variables incorporated, quantification of flow condition may have a stronger relationship to biology than percentage impact at Q95.

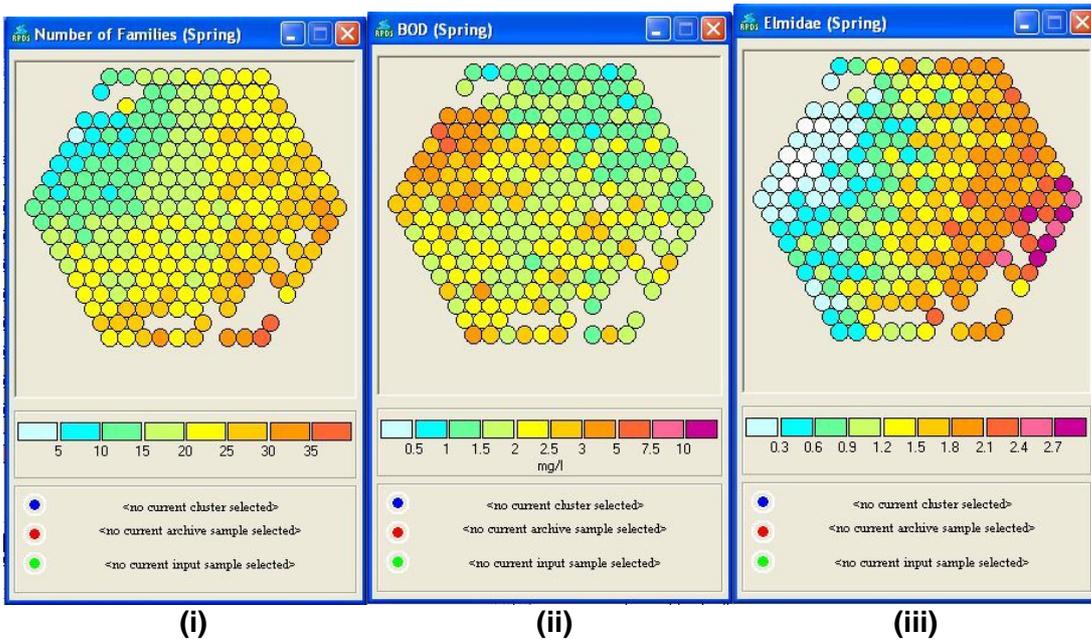


Figure 9.1(a) Output maps from RPDS 2.0 (spring 1995 model): (i) Number of families, (ii) BOD and (iii) Elmidae.

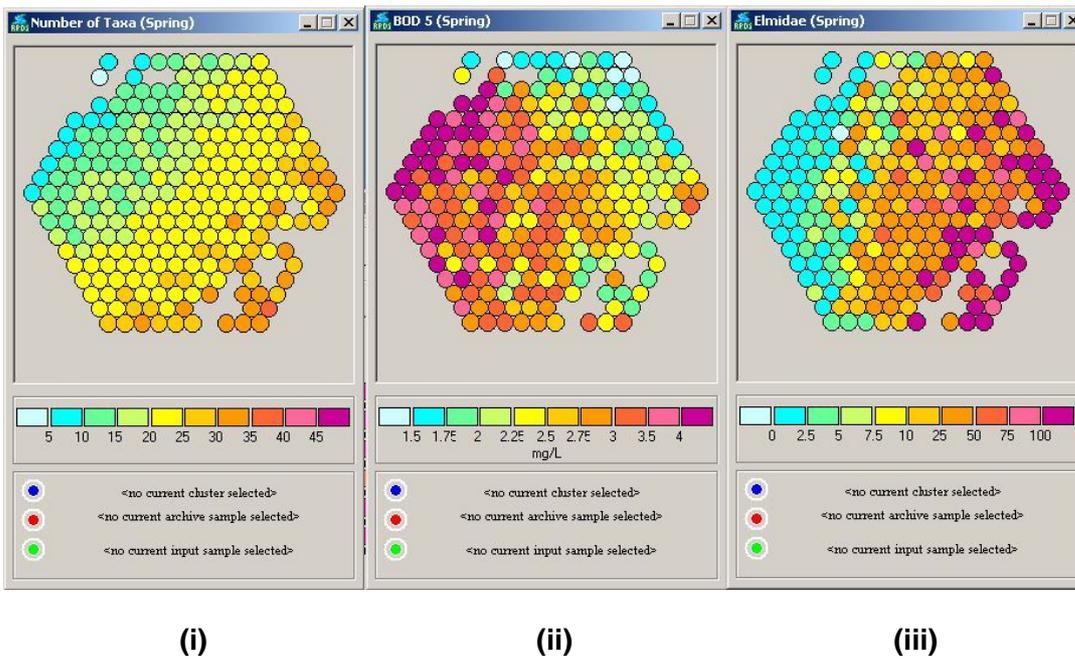
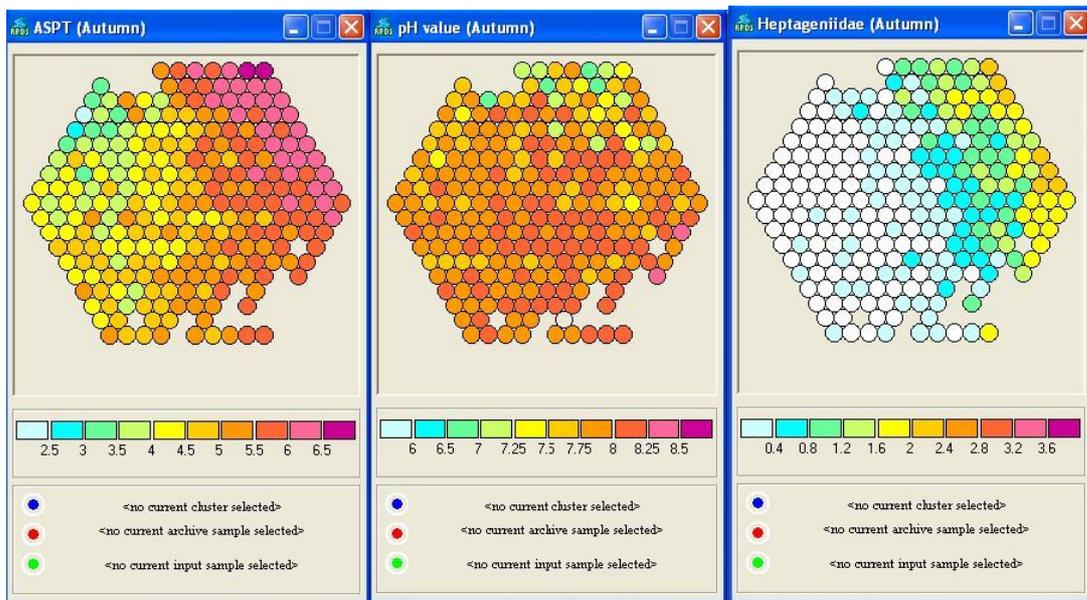


Figure 9.1(b) Output maps from RPDS 3.0 (spring model): (i) Number of families, (ii) BOD and (iii) Elmidae.

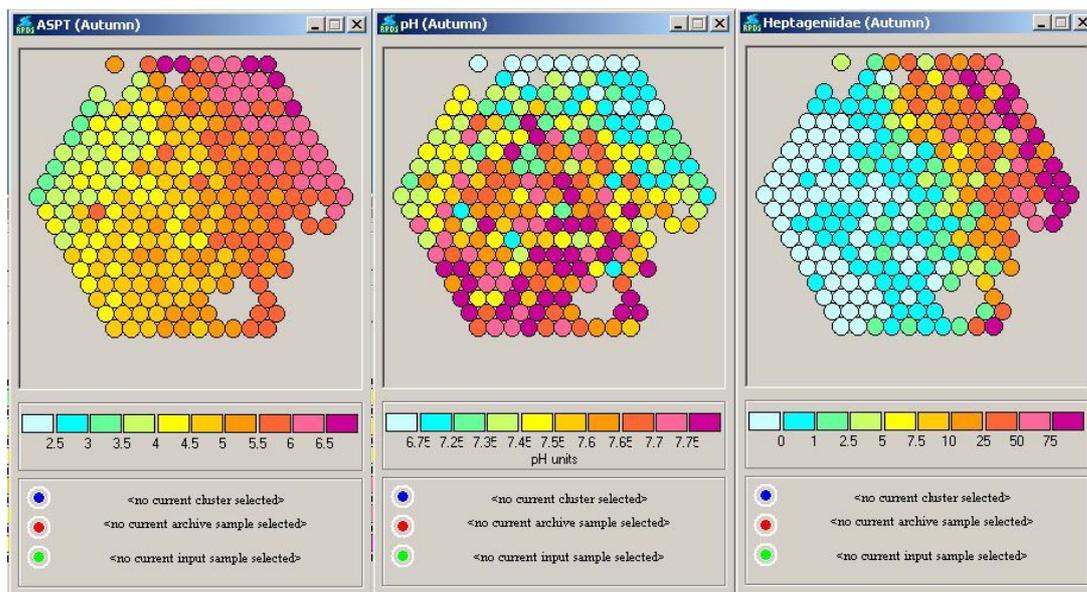


(i)

(ii)

(iii)

Figure 9.2(a) Output maps from RPDS 2.0 (autumn 1995 model): (i) ASPT, (ii) pH and (iii) Heptageniidae.



(i)

(ii)

(iii)

Figure 9.2(b) Output maps from RPDS 3.0 (autumn model): (i) ASPT, (ii) pH, (iii) Heptageniidae.

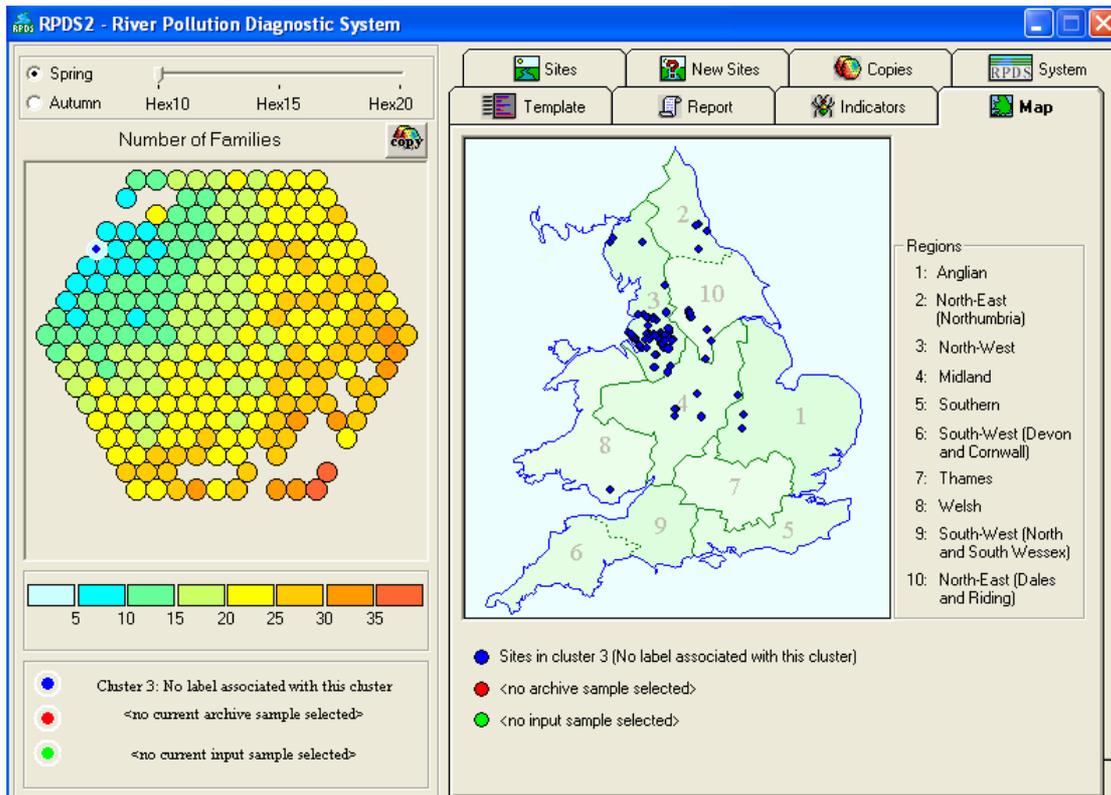


Figure 9.3(a) Geographic map panel on RPDS 2.0 showing spatial locations of samples in cluster containing least diverse sites.

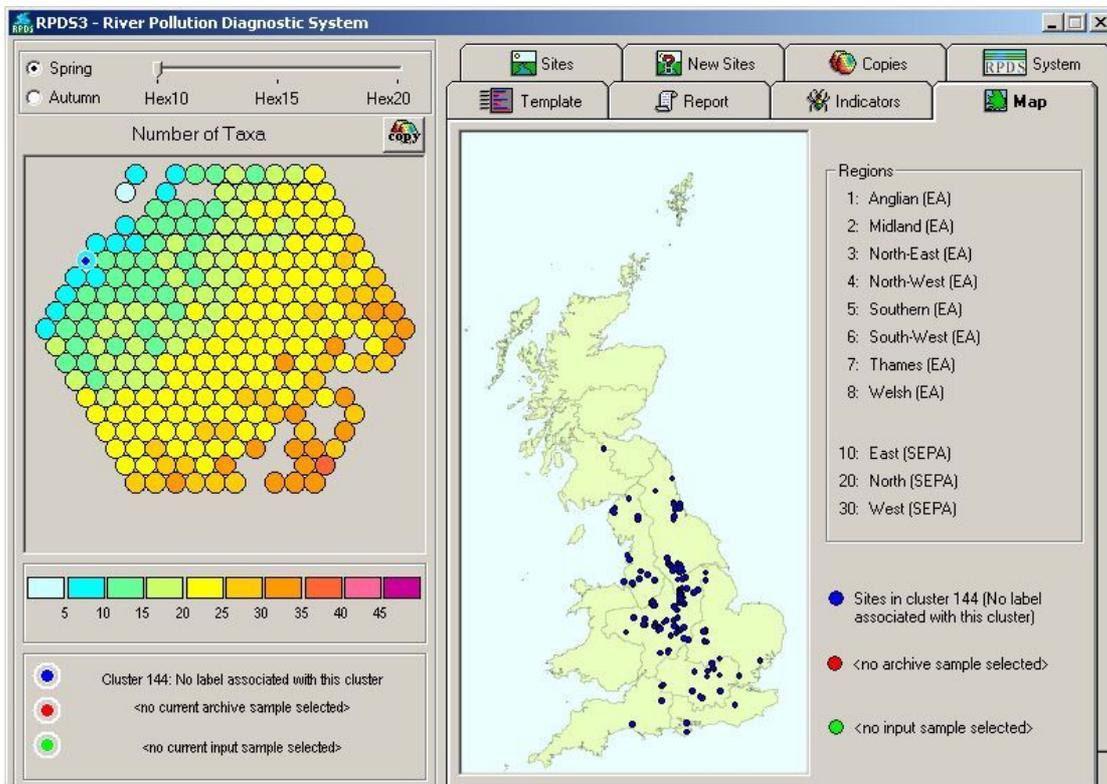


Figure 9.3(b) Geographic map panel in RPDS 3.0 showing locations of samples in cluster corresponding closely to that illustrated in Figure 9.3 (a).

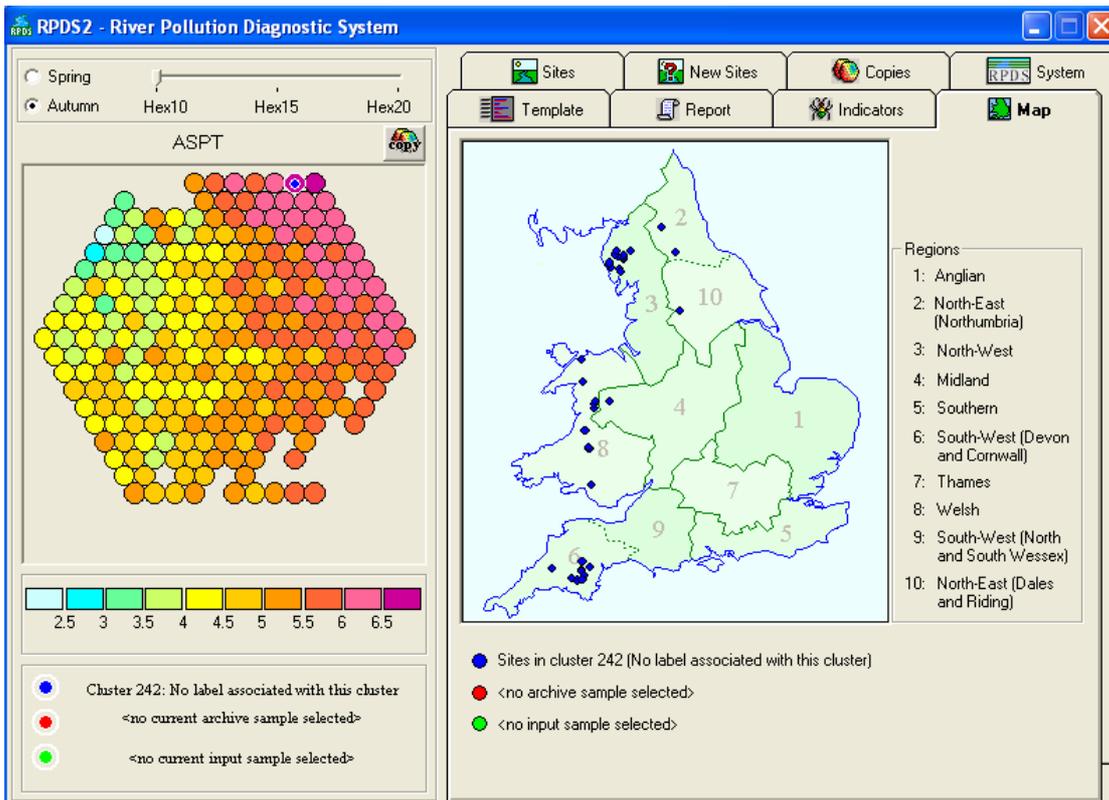


Figure 9.4(a) Geographic map panel on RPDS 2.0 showing spatial locations of samples in cluster containing samples with high ASPT values.

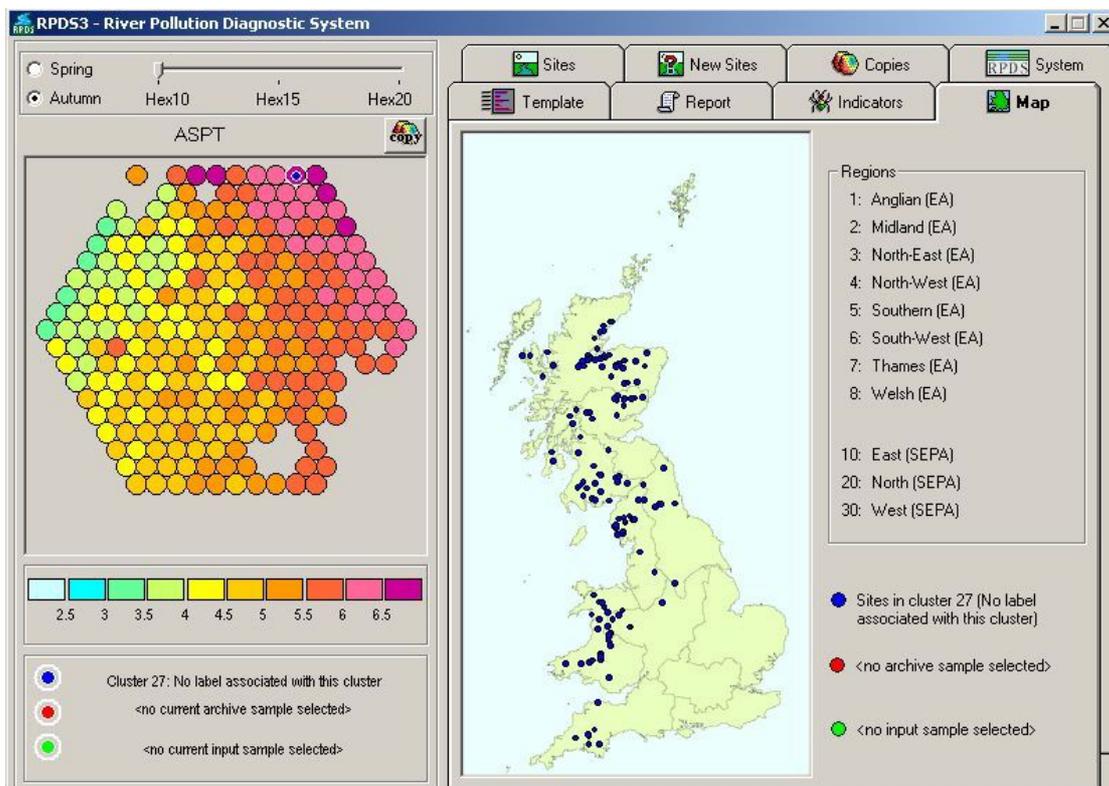
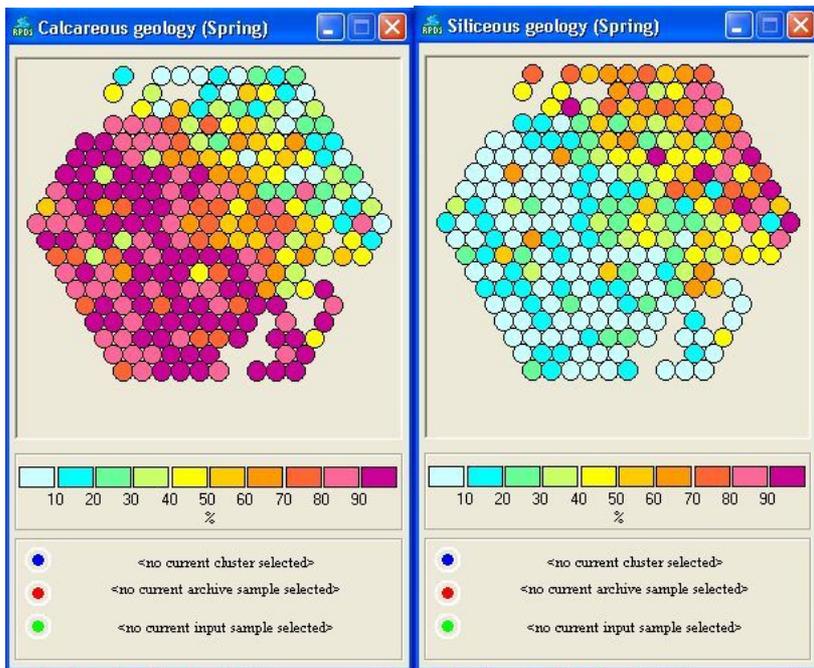
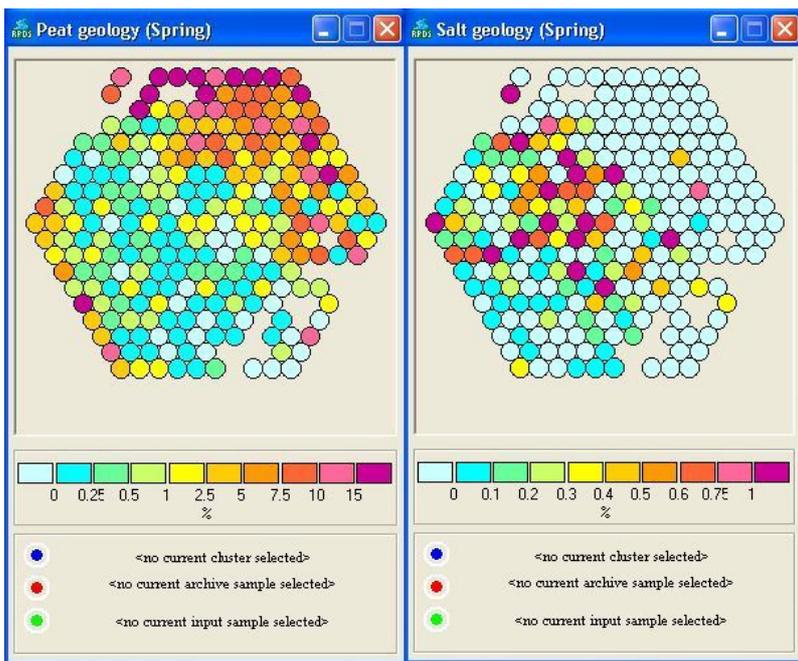


Figure 9.4(b) Geographic map panel in RPDS 3.0 showing locations of samples in cluster corresponding closely to that illustrated in Figure 9.4(a).



(i)

(ii)



(iii)

(iv)

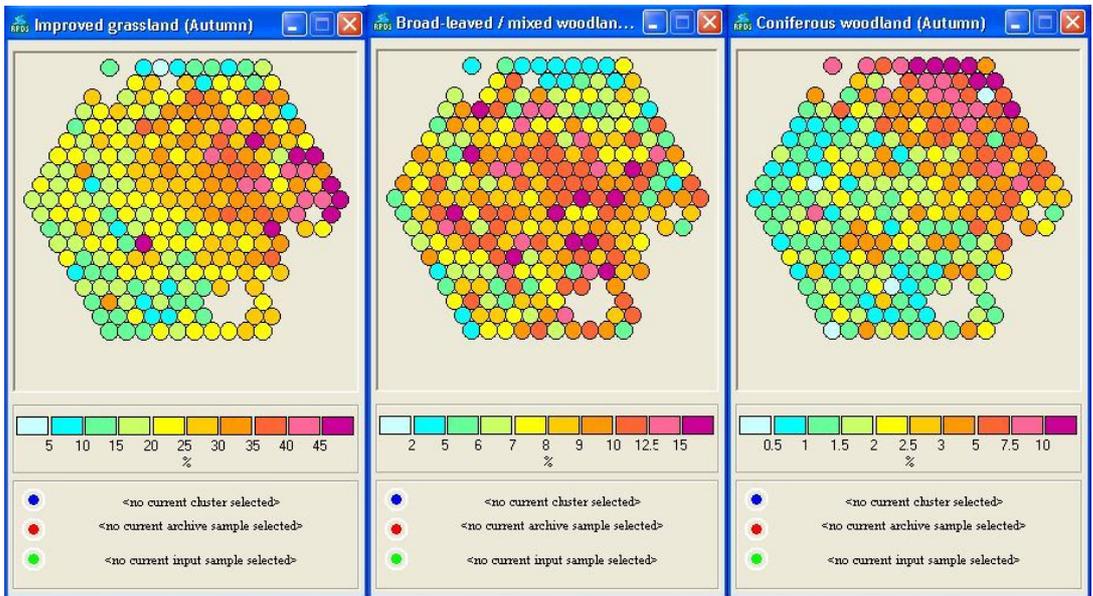
Figure 9.5 New geological variables in RPDS3. Shown are maps of percentage of upstream catchment area in spring model categorised as (i) calcareous; (ii) siliceous; (iii) peat; and (iv) salt.



(i)

(ii)

(iii)

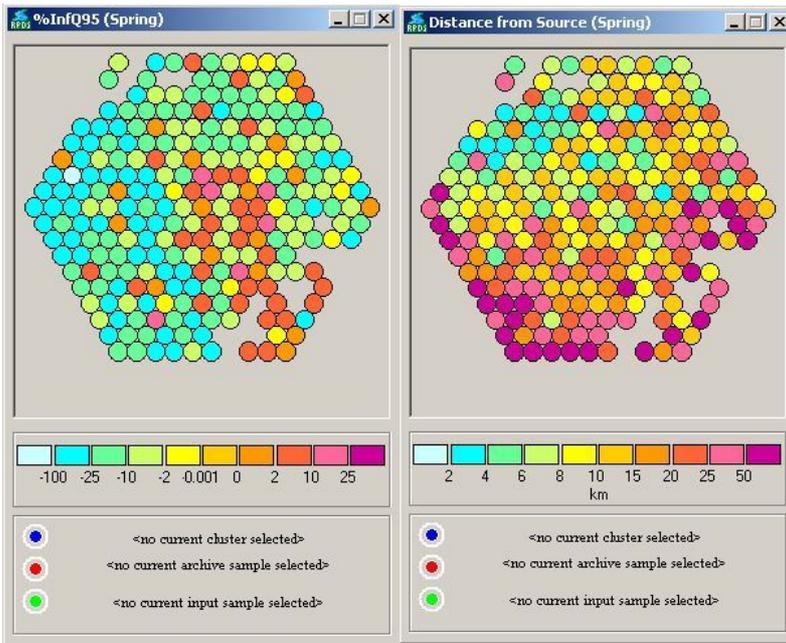


(iv)

(v)

(vi)

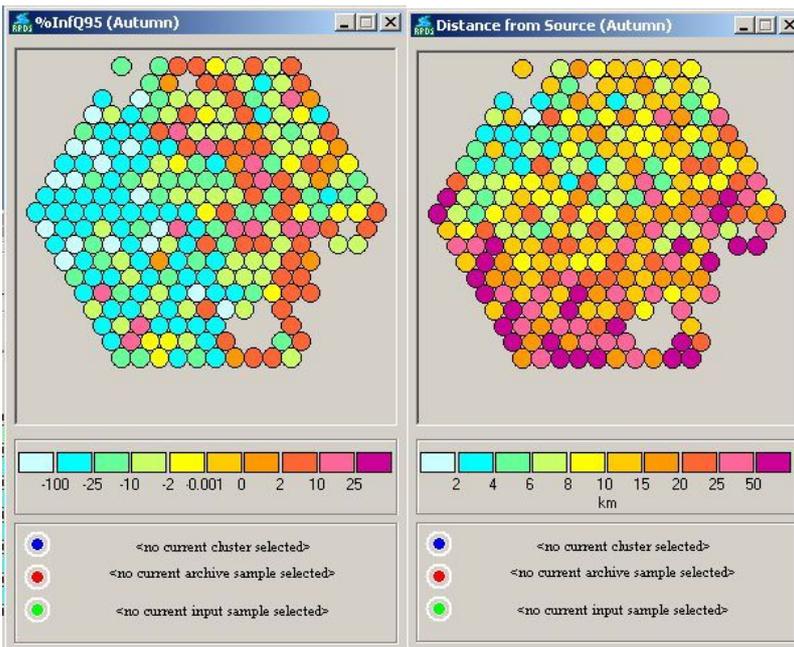
Figure 9.6 New land cover variable in RPDS3. Shown are maps of percentage of upstream catchment area in autumn model categorised as (i) urban; (ii) suburban; (iii) arable; (iv) improved grassland; (v) broadleaved woodland; and (vi) coniferous woodland.



(i)

(ii)

(a)

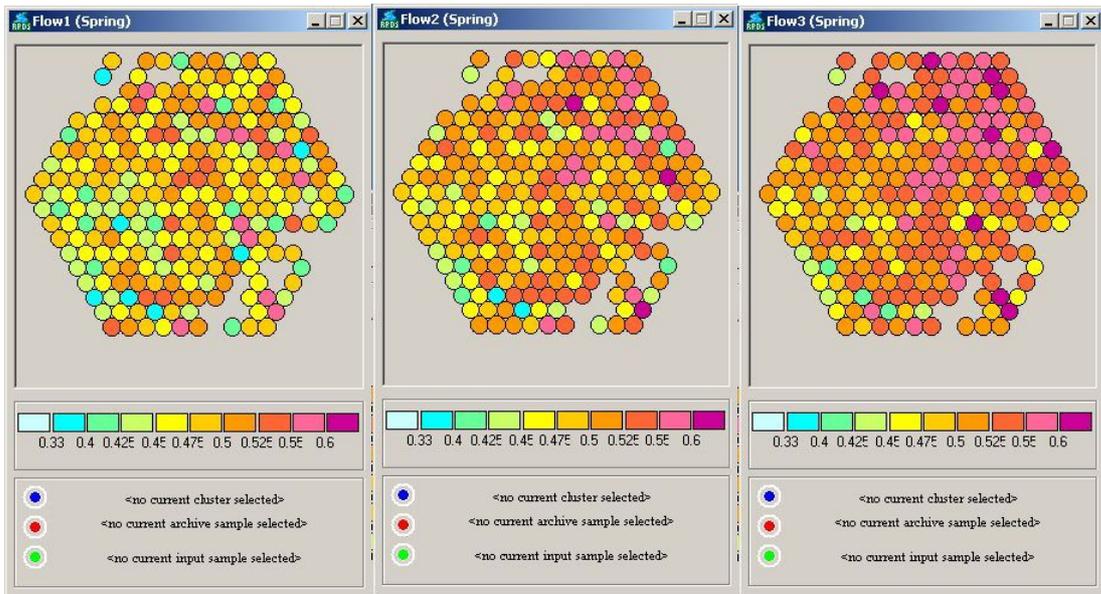


(i)

(ii)

(b)

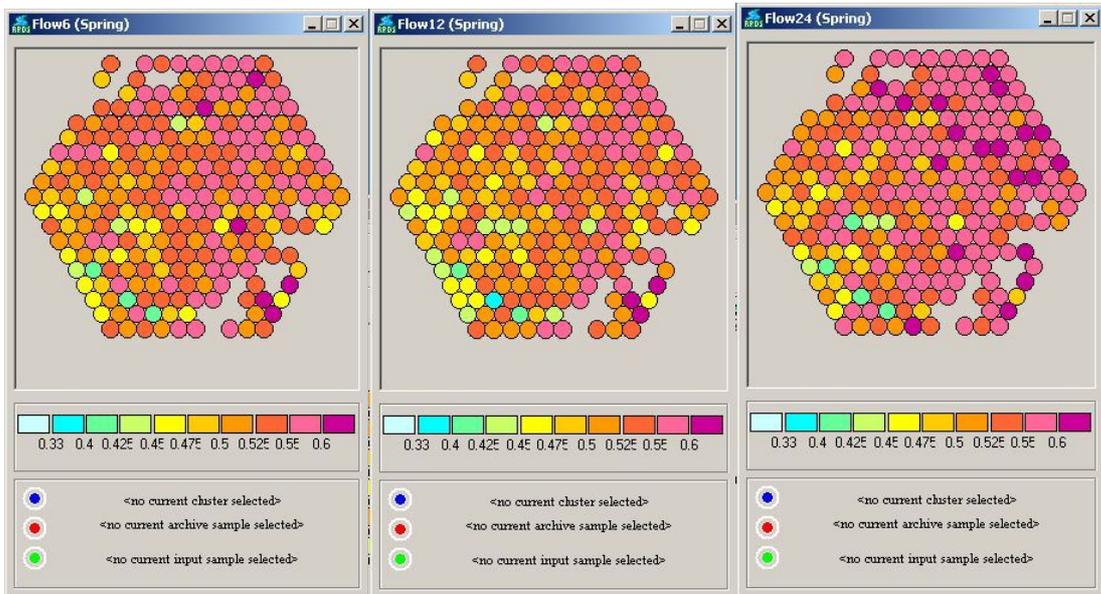
Figure 9.7 (a) Output maps in spring model for (i) percentage impact at Q95 and (ii) distance from source. (b) Output maps in autumn model for (i) percentage impact at Q95 and (ii) distance from source. (Note that a positive value of percentage impact at Q95 corresponds to a reduction from natural flow, and a negative value corresponds to an increase from natural flow).



(i)

(ii)

(iii)

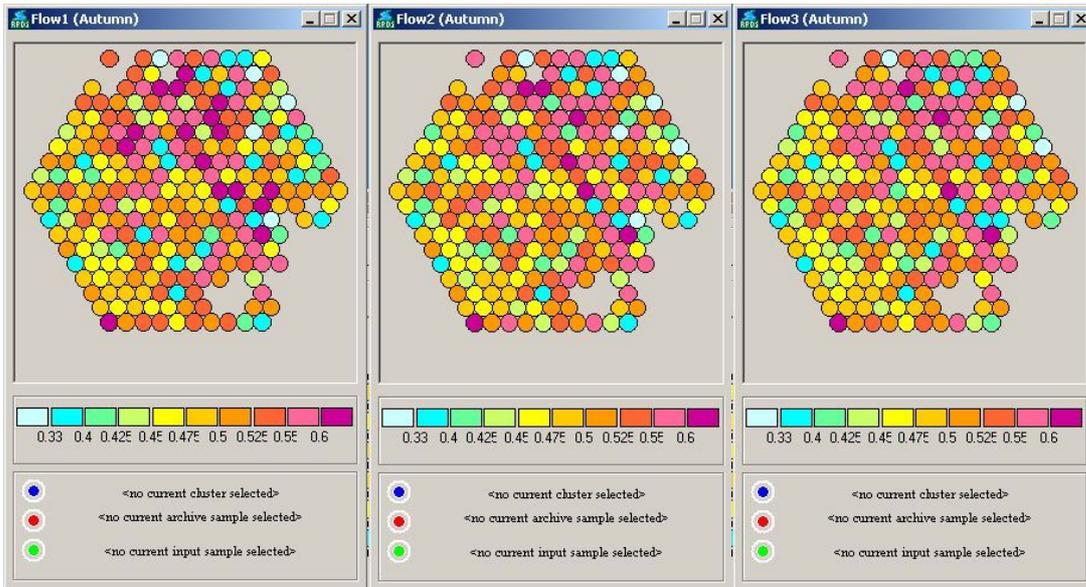


(iv)

(v)

(vi)

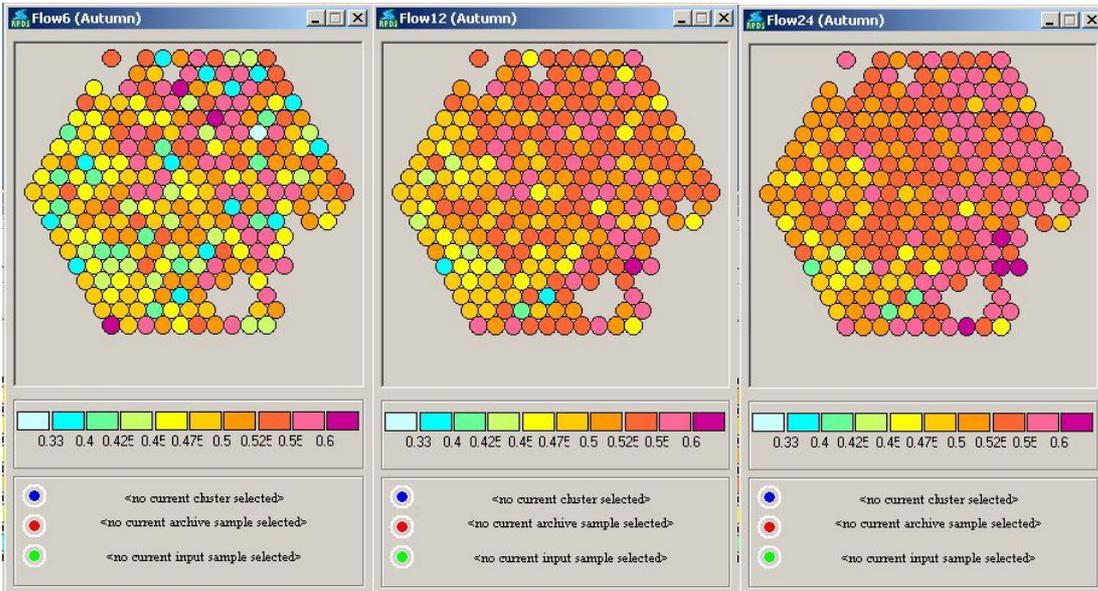
Figure 9.8 Flow condition variable in RPDS 3.0 spring model. On a scale of zero to one (where zero is driest and one is wettest) over a time period prior to the sample date of: (i) one month; (ii) two months; (iii) three months; (iv) six months; (v) 12 months; and (vi) 24 months.



(i)

(ii)

(iii)



(iv)

(v)

(vi)

Figure 9.9 Flow condition variable in RPDS 3.0 autumn model. On a scale of zero to one (where zero is wettest and one is driest) over a time period prior to the sample date of: (i) one month; (ii) two months; (iii) three months; (iv) six months; (v) 12 months; and (vi) 24 months.

Extensions to RPDS functionality

Following a meeting with potential users, a list of requirements was drawn up (Table 9.1). Some requirements relate to easier use of the software while others refer to additional functionality needed for the Water Framework Directive. Following refinement, these requirements (and a corresponding set for RPBBN) were implemented during two extensions to the original project. The subsequent work to modify RPDS is described in the next section.

Table 9.1 List of requirements for additional functionality to RPDS.

Deliverable
1. Include a batch mode option – batch input and output. In the output report, the stresses with highest occurrence at a given probability. Standard format for data files (to link with RIVPACS & BIOSYS).
2. Needs an indicator confidence based on amount of data within the cluster, and the relative certainty of the classification of a sample to the cluster.
3. Add high, good, moderate, fair and poor ecological WFD Status.
4. Include a one-line summary of pressures in the output (along with bin label).
5. Include predictions of both new and old BMWP scores.
6. Include LIFE score in the diagnostic info for RPDS, calculated as abundance-related ASPT.
7. Include the Flow Q95 statistic for the six months prior to the sample date in the diagnostic info.
8. Under the hexagon, expand the key to say exactly what is displayed (this may repeat information in indicators description box). For example, where a taxon is shown in the hexagons, the wording under the scale bar should say the name of the taxa and that the value is the average abundance category of samples in that cluster as this isn't clear at the moment. I've found that people have no idea what is being represented so need it being spelt out to them!
9. Facility to output bitmaps into Word and to copy/paste screens.
10. Comparisons between years via tabbing facility.
11. Include crayfish as stress.
12. Populate image windows in the indicators tab.
13. Make the expansion button work so screen is filled.
14. User Guide amended and to include worked examples.
15. In template please include an explanation of what the number is at the end of the graphs (e.g. that the '4' is the scale of the graph, that it is the maximum abundance category present).
16. Put a scale bar under each graph – at least little dashes to show where the values are.
17. We're not including any RHS data or ANC values.
18. Keep Hex 10 as the default.
19. Define a default template (not just a blank one).

10 Modifications to RPDS software

Introduction

This section covers work undertaken to enhance the functionality of RPDS based on the list of requirements given in Table 9.1. These tasks and the corresponding work on RPBBN (Section 12) were undertaken during two extensions to the main project, following some refinement of the requirements. The sections below reflect the revised work items..

Data Tasks

Add LIFE, BMWP, WHPT and EQI/EQRs from the new version of RIVPACS (RIVPACS IV, as implemented in the River Invertebrate Classification Tool (RICT) software for WFD classification) to the database. Generate diagnostic information for clusters (high, good, moderate, fair and poor WFD ecological status).

The LIFE (Lotic-invertebrate Index for Flow Evaluation), BMWP and WHPT (Walley Hawkes Pailsley Trigg index) values for the samples could be generated relatively easily because they only required the existing biological sample data and list of indices. As a result, these values were generated in-house using a generalized score-calculating algorithm. As values were associated with each of the existing database samples, generation of the diagnostic information was simplified because it could be achieved using existing algorithms.

For WFD, RIVPACS IV (implemented in River Invertebrate Classification Tool software, RICT) and the WFD classification of ecological status replaced RIVPACS III and the GQA biological classification. The river invertebrate classification of ecological status was not finalised until late 2007 and so it could not be incorporated in RPDS and RPBBN before this extension.

Predictions from RIVPACS IV differ slightly from RIVPACS III because of the addition of reference sites from the highlands and islands modules to the GB module and the removal of reference samples from about 40 reference sites deemed to be of insufficient environmental quality. These changes are reported in Davy-Bowker *et al.* (2008), based on analyses reported in Davy-Bowker *et al.* (2007).

Whereas the GQA classification is based on EQIs, the WFD classification is based on ecological quality ratios (EQRs). EQIs are simply the raw predictions of the classification metric (ASPT or N-taxa) from RIVPACS divided by the value observed in samples collected from the site. EQRs are based on predictions of the classification metric at WFD reference state. WFD reference state is the value of the classification metric at the site if it was in WFD reference state, in which there are no more than minor ecological changes caused by human activity. Reference values of the classification metric are those observed at reference sites that are in reference state. In the UK, reference values were based on RIVPACS predictions. These were adjusted, to remove variations caused by the varying quality of RIVPACS reference sites, so that the predictions and EQIs related to the quality represented by the boundary between WFD high and good status. This was then converted to a reference value by multiplying by a factor based on the median value of the classification metric at all RIVPACS reference sites and the subset in WFD reference condition. RIVPACS reference sites that were also in WFD reference condition were

identified by screening them against the criteria for defining reference devised by pan-European Geographical Intercalibration Groups.

Bias (systematic error caused by laboratory analysis of samples) was taken into account for N-taxa, but not for ASPT (other than the default that RICT implements) because it was also taken into account in the WFD classification of ecological status.

In order to obtain the EQR data from RICT, it was first necessary to extract the sample input data from the project database. This data consisted of several environmental parameters for each site, such as altitude, slope, discharge category, distance from source, width, depth alkalinity and substrate composition, and the biological sample data to provide the means of generating the observed sample values. The main problem faced in supplying this data was ensuring that all the necessary input values were present and that there was a spring and an autumn sample in each year, which could be combined to produce the 'annual' sample used for WFD classification. These data requirements meant that a fifth of samples of the project database were ineligible for inclusion in the RICT input dataset. The total number of samples containing the full range of parameters and having both spring and autumn samples within a year was just under 52,000, which produced roughly 26,000 combined season samples.

This input dataset was reclassified using RICT. This was the first time that a large number of sites was classified using RICT in batch mode, which had not been released for general use at the time. Mark Caulfield (RICT programmer) copied RICT to a fast internet server and made numerous modifications to the input and output files and user interface during the course of this work. These changes are listed below.

- Filenames of input CSV files generated by RICT were altered so that the year followed rather than preceded the rest of the name. This enabled large batches of input files to be ordered sensibly rather than by year. Input files with the same bias could then be listed together in Windows directories, making it easier to set up each run (each analysis of each pair of input files comprising data for up to 50 samples).
- The batch process was modified so that the user could chain runs together (RICT would start to analyse data for the run (from a pair of input files) as soon as it had completed the analysis of the previous run). This enabled multiple runs to be loaded into RICT (up to 20) without overloading the programme.
- The time and date that RICT completed each batch run was added to RICT's run menu, so that the user could track the progress of each run.
- The display for setting up each run was modified to indicate the bias value that would be used – this gave a quick visual check that bias had been altered from the default.
- The probability of the site belonging to each of the five WFD ecological status classes was added to the output.

The EQRs and WFD classification added to the project database were derived in an identical manner to those used by the UK's regulatory agencies for WFD classification. The only minor difference was that, in England and Wales, WFD classification accidentally ignored Pediciidae (a component of Tipulidae) in classifications to date (2009), although this error is expected to be corrected soon.

The results from RICT and those produced by the Environment Agency contained matches for 50,945 of the 63,565 samples in the RPDS database. RICT produced

values for Biological Monitoring Working Party score (BMWP)⁹, Average Score Per Taxon (ASPT), Number of Taxa (NTaxa) and the WFD status class based on the minimum of NTaxa and ASPT (MINTA). Several values were produced for each parameter including basic observed and expected values, the values adjusted for sample bias and confidence measures for both the predicted values and the quality classification. From this range of values, the observed, the reference-adjusted expected and the most common face value bias EQI were included for BMWP, ASPT and NTaxa. The class and class confidence was included for MINTA.

Once these results were created, it was then relatively straightforward to match them to their original RPDS samples and use existing algorithms to generate the necessary diagnostic information for RPDS.

Replace stress categories with revised categories used in river basin management and described in PISCES database

Stress categories were simplified and split into sector, activity, and pressure for use in biological outcomes (predicting WFD ecological status at the end of the six-year river basin management cycle for WFD) and stored on the Environment Agency's PISCES (Pressure Information Supporting Classification Elements for the Water Framework Directive) database. The PISCES codes for sector, activity and pressure are given in Appendix E. Because these categories were more familiar to Environment Agency staff, stress data in the project database were converted to the new categories.

The closest match between stress categories and new categories was identified in a meeting between the Environment Agency project manager (John Murray-Bligh) and PISCES database manager (Graeme Storey) in early 2009 and is reproduced in Appendix F. This table provided the means to convert existing project stress data to the new PISCES codes. Because the existing 1995 stress data had been converted to the 2000 stress codes when it was incorporated into the project database, it was also converted to the PISCES codes.

The change in data format, however, raised the issue of what data should be included in the RPDS samples. In the 2000 stresses survey, stresses were categorised by source, category and type. With over 130 types of stresses, it was impractical to include 'type' data as this would require the addition of hundreds of extra fields to the samples. Therefore only 25 stress category fields were included in the original RPDS. Analysis revealed that in the revised data there were 47 different types of activity and 26 types of pressure. Despite the fact that including both activity and pressure would require 73 fields, an additional 48 fields from the original 25, it was felt that the benefits of including the additional information would outweigh the drawbacks of increased database size and complexity of outputs.

Changes to the user interface

Replace displays of GQA classification by WFD classification

References to the GQA classification were replaced by references to the WFD classification in reporting of the diagnostic information for clusters. Other references to GQA were replaced on the Sites tab and in the printed output available under the Print Options of the System tab.

Under the hexagon, expand the key to say exactly what is displayed

⁹ The Biological Monitoring Working Party score is a method of quantifying biological quality. The quality for a sample is denoted by the sum of the scores for each macroinvertebrate taxon in the sample.

An indicator description box was added under the hexagon to provide information on the selected indicator. This is shown in Figure 10.1 for the indicator 'total ammoniacal nitrogen'.

Populate Image windows in the Indicators tab

Where the chosen indicator is taxonomic, a coloured image of the taxon is displayed. Descriptive text providing details of distinguishing features, habitat and life cycle were added. This is illustrated in Figure 10.2 for the indicator Caenidae.

Add scale bar to graphs

Scale bars were added to make it easier to interpret data on the indicators selected in the Template panel.

Figure 10.3 illustrates this for indicators in the current cluster (blue dot in the hexagon and corresponding blue bars). Similar scales are displayed in two other cases: firstly, when comparing data in the current cluster to that of an archive sample in the model (Figure 10.4, red dot on the hexagon and red bars); secondly, when comparing data in the current cluster to that of a input sample (Figure 10.5, green dot in hexagon and green bars).

The ranges of the scale in the bar depend on the minimum and maximum values in the selected cluster, and are defined automatically.

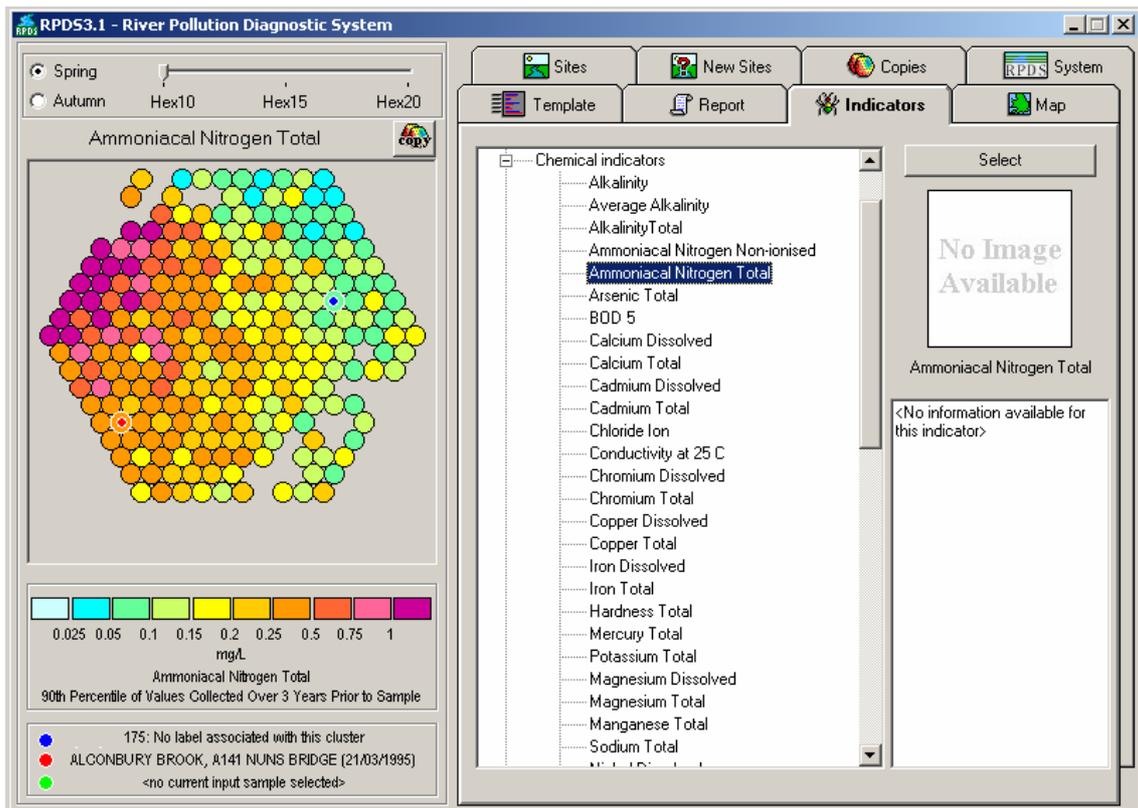


Figure 10.1 Description box below hexagon for indicator 'total ammoniacal nitrogen'.

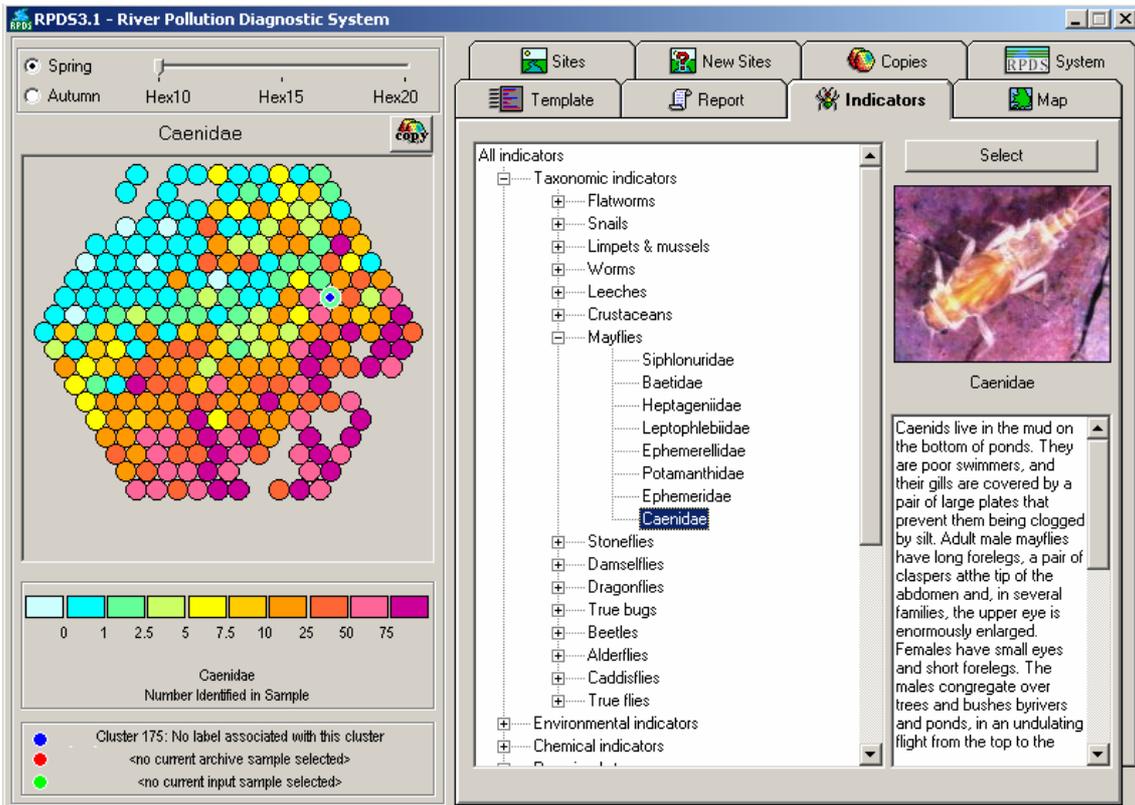


Figure 10.2 Image and descriptive text for indicator Caenidae.

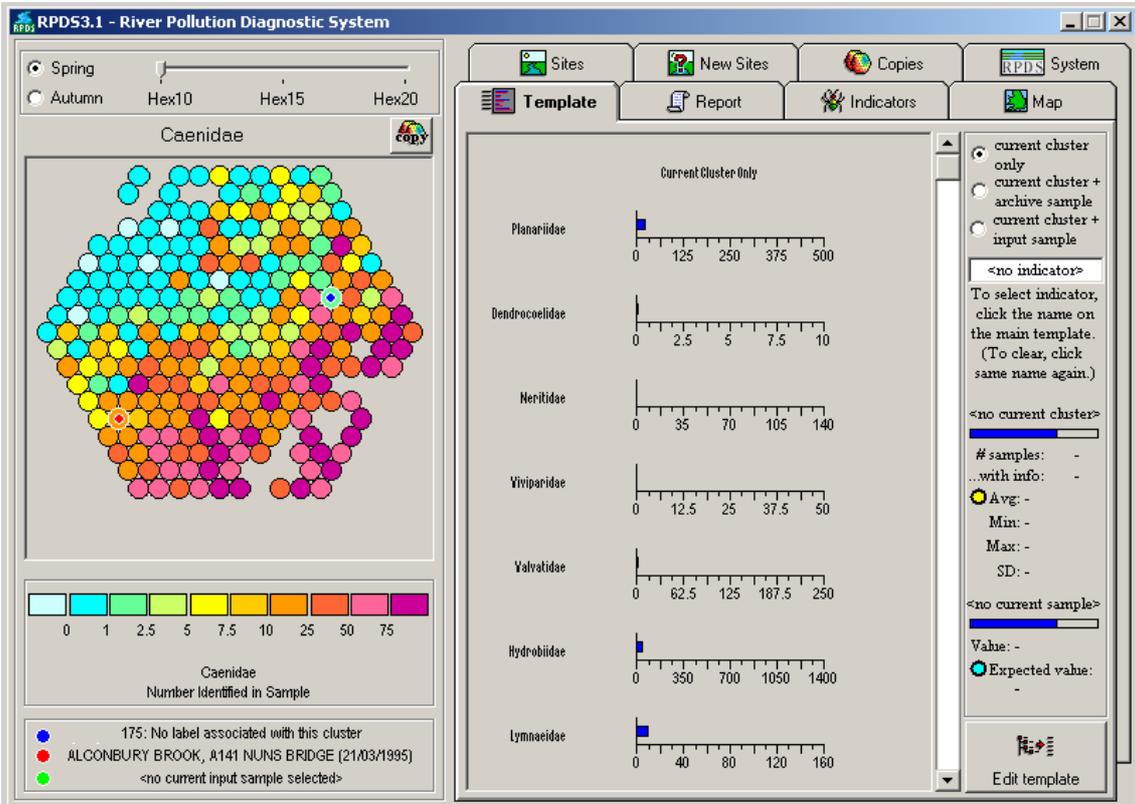


Figure 10.3 Range of scale values for selected indicators on current cluster only.

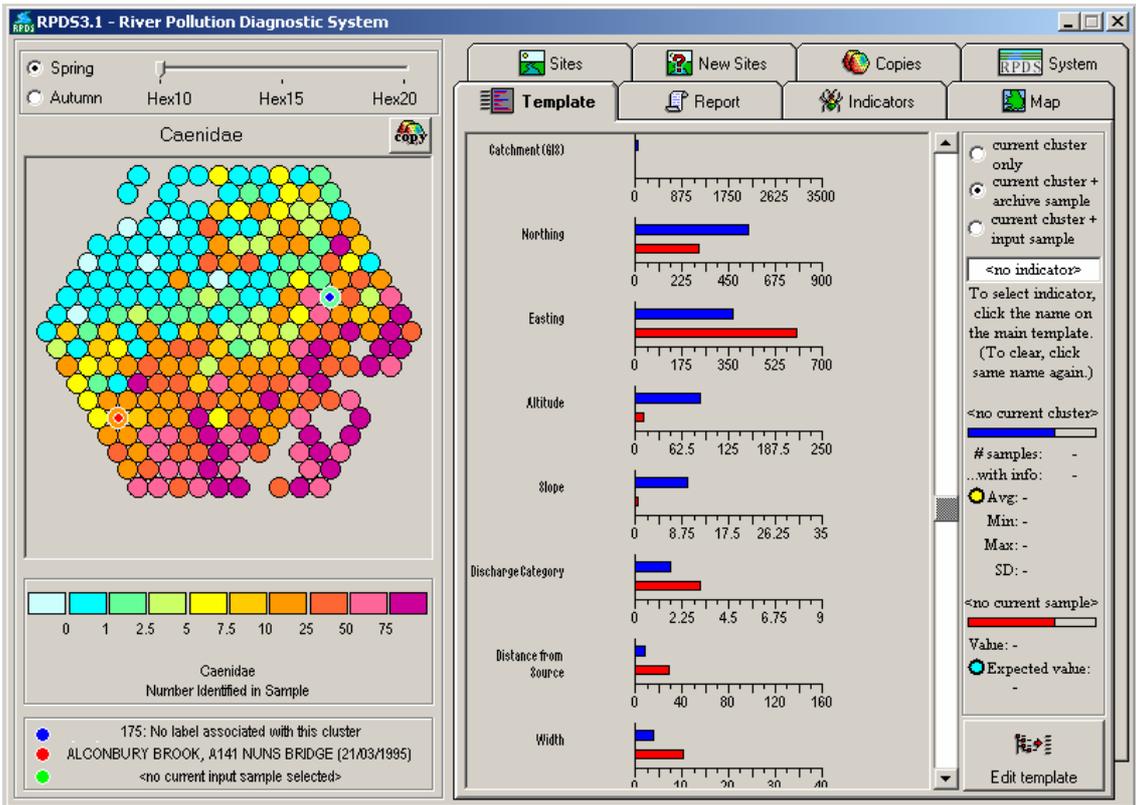


Figure 10.4 Range of scale values for selected indicators on current cluster and archive sample data.

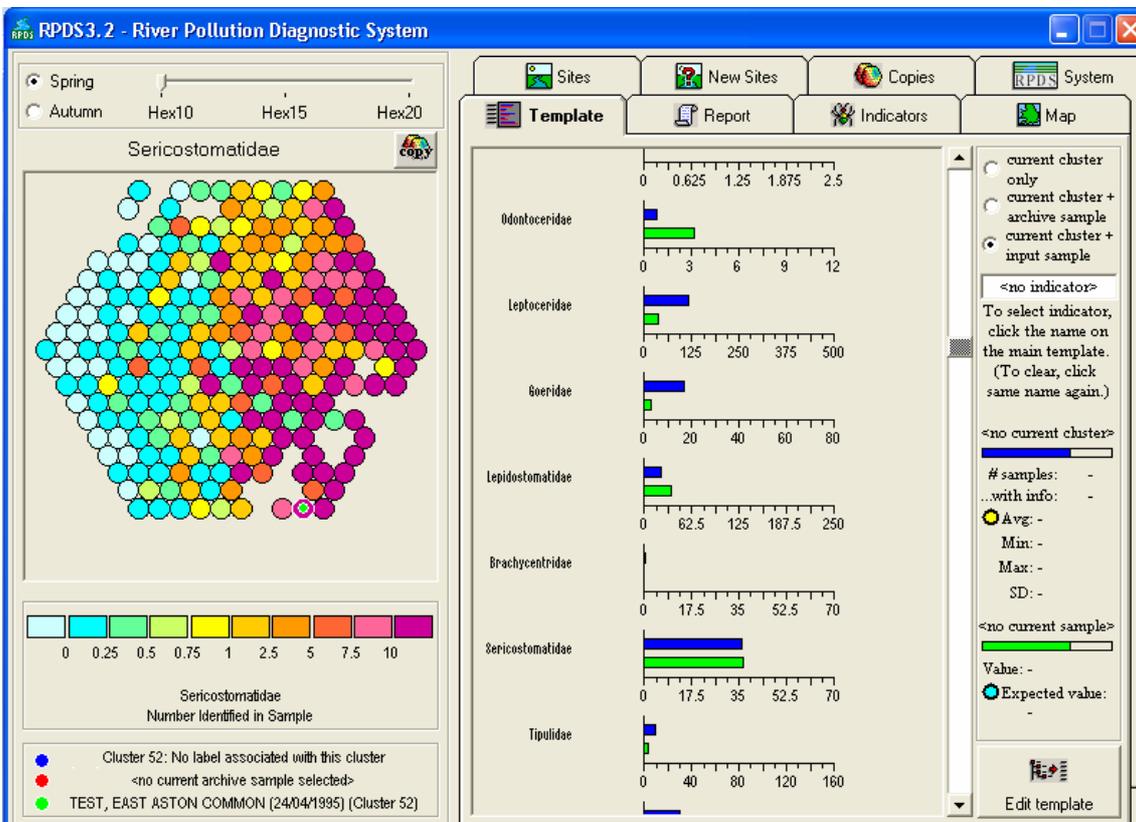


Figure 10.5 Range of scale values for selected indicators when comparing current cluster and input sample data.

Batch Mode

Introduction

In the existing version of RPDS, it was possible to load new samples from a file, classify a single selected sample and check the results within the application. Although this feature enables the diagnostic capabilities of RPDS to be used on new sample data, the fact that only one sample could be processed at a time made the analysis of multiple samples time-consuming. To solve this problem, a 'batch mode' was added to RPDS to enable large numbers of samples to be classified and potential pressures reported quickly in a single operation.

The main issue in creating the 'batch mode' was to standardise and automate the analysis and interpretation. In the interactive mode of RPDS, the user interprets the results. Typically, the user would analyse the data for the cluster to which the new sample had been classified to identify potential pressures. To implement the batch mode, it was necessary to develop an algorithm capable of mimicking this process. To provide a clear understanding of 'batch mode', how batch mode functions and the range of parameters available to the user, a brief description of the process of identifying potential pressures is provided.

The fundamental premise of the pattern recognition and diagnosis in RPDS is that a subset of parameters is sufficient to identify a particular type and that, given an existing body of knowledge about that 'type', it is possible to infer information about additional parameters from this knowledge. So for example, given a set of symptoms a doctor may be able to identify a particular disease based on the body of knowledge of diseases that occurred when similar symptoms were found in the past. Using knowledge of that disease, the doctor is able to infer the type of infection, how the disease will progress and potential cures.

In RPDS, the initial subset of classifying parameters consists of biological and environmental sample parameters. The 'body of knowledge' is chemical and stress data collected along with biological and environmental data used to create the model. The inferred parameters are the chemical, flow and stress values typical of a cluster. Analysis and interpretation involves identifying the inferred parameters that are 'normal' and those that are 'potentially problematic'. To do so, it is necessary to have an idea of what is 'normal' and what constitutes sufficient deviation from this norm to be considered 'potentially problematic'.

Deriving the predicted values

The first stage of the batch mode process is to classify the new sample data, and two methods have been provided for this (selected under the System tab): mutual information or Mahalanobis distance.

In simple terms, classification by mutual information is done by iteratively adding the sample to each of the clusters and recording the change in the mutual information value for the whole model. Any addition that results in an improvement of the model is highlighted as a potential classification and these are scaled from best to worst.

Mahalanobis distance is a natural measure for quantifying the 'distance' between a new sample and samples in a particular cluster. The inversion and storage of large covariance matrices is difficult for data management (since it needs to be done for each cluster), and this is circumvented by assuming that each of the attributes used for clustering is independent so that covariances are zero. The Mahalanobis distance then simplifies to the Euclidian distance D_j between the site and the centre of the cluster (after each attribute has been normalised to unit variance).

Both methods identify one potential classification as the best, although neither method produces a definitive answer because there may be several other clusters to which the sample might feasibly belong. In the original RPDS application, the full set of results was simply shown to the user and it was left to them to analyse the results and draw their own conclusions. For a batch mode though, it would be necessary to derive a single definitive solution on which the rest of the analysis could be based.

In previous projects, two methods of obtaining predicted values from a classification have been used. Having selected the approach to use (mutual information or Mahalanobis distance), predicted values can be derived from the best cluster or weighted mean of several clusters. The batch mode offers the user the option of either method to derive predictions.

- Best cluster – in this method the ‘best’ cluster solution is taken to be the definitive answer and all parameter values of that cluster are considered to be the predictions for the new sample.
- Weighted value – in this method, pseudo-probabilities of belonging to each cluster are derived and predicted values are then calculated as a weighted mean with the pseudo-probabilities used as the weights. In the case of mutual information, the pseudo-probabilities are given by MI values (normalised by their sum) of all clusters where addition of the new sample improved the MI (that is, the change in MI was positive). In the case of Euclidian distance, they are given by the quantities $h_j = e^{-D_j^2/2}$ (normalised by their sum). A similar approach is adopted in RIVPACS, where distance is used to determine the probabilities of group membership.

Confidence in the predicted values

Both of these methods provide predicted values for new sample data, generated using the cluster values for the variable. However, the following issues affect the reliability of the values.

- Amount of source data – the sparseness of data for particular variables means that some values may be based on few actual values, making bias a potential problem.
- Effects of data distillation – the cluster values are mean values. The mean is used because it is representative of all sample values assigned to the cluster. However, it does not contain information on the spread and skew of the sample set. Measures of spread and skew are important parameters for evaluating cluster values because they indicate whether the underlying sample values are consistent or more random and widely spread. Providing an indication of how representative the distilled values are is important.

Identifying outlying values

To mimic the analysis and interpretation, the ‘batch mode’ algorithm can use one of two methods based on simple statistic measures: standard deviation and percentiles. Using these measures of distribution, it is possible to identify ‘outlying’ values that deviate from the norm by more than a predefined threshold value. The standard deviation method allows the user to define the number of standard deviations a value must deviate by. The percentile method provides two predefined levels, 10th/90th percentiles and 5th/95th percentiles.

Sources of the statistical data

To identify outliers, it is necessary to derive values for the mean, standard deviation and percentiles. Two potential sources of data could provide these values, the sample data itself and cluster values, which are the values generated to describe the cluster from the sample data assigned to it. In RPDS, these are mean values. The mean, standard deviation and percentile were generated using both sources of data and the option to use either is provided.

Activating the batch mode

The batch mode is activated by clicking the 'Start Batch Mode' button at the bottom of the New Sites tab. A dialogue box asks the user to select the options described above, see Figure 10.6.

Report

Following a run in batch mode, a .csv file is generated with a report, Figure 10.7. Each line of the report gives details of the sample and its outlying attributes (if any) according to the user-specified criteria described above.

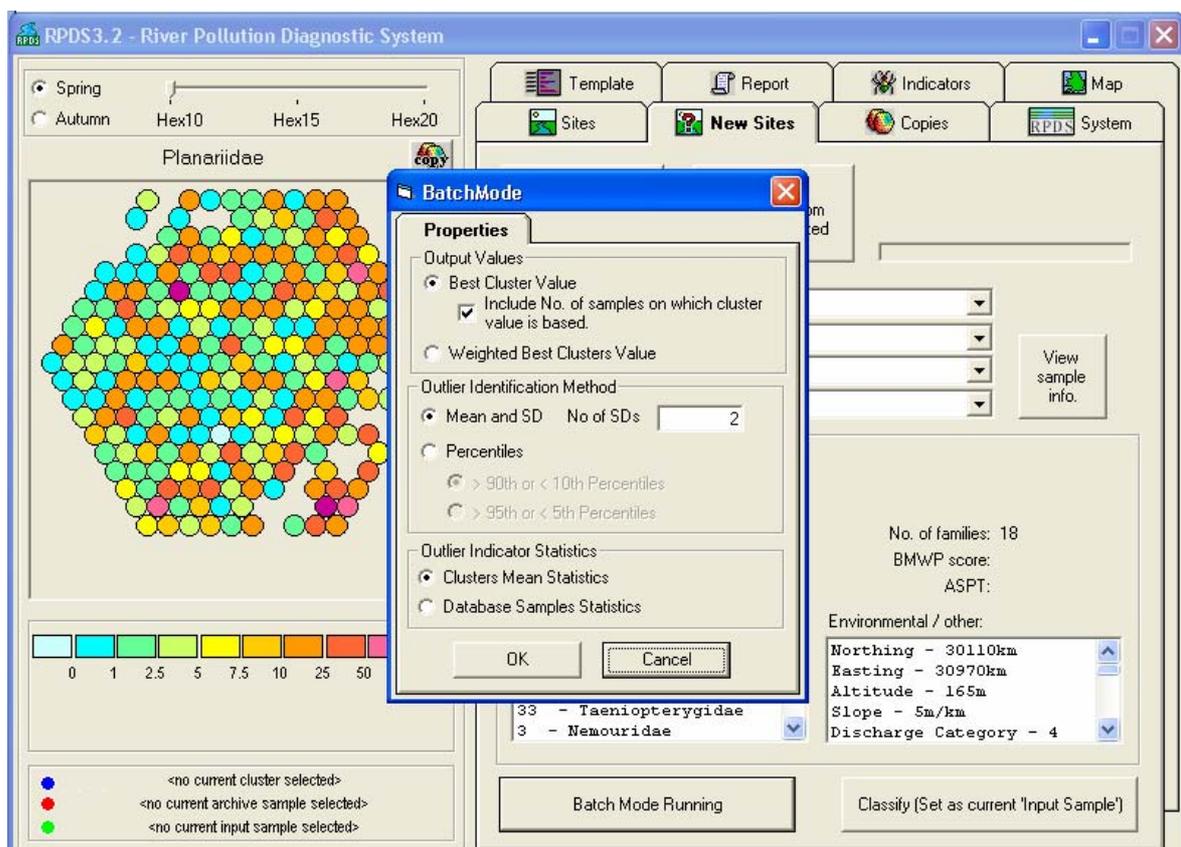


Figure 10.6 Activation of batch mode with dialogue box for selecting options.

SampleID	Date	Region	BestCluster	Pressure	Value	ClusterSar	10%	90%	Pressure	Value	ClusterSar	10%	90%	Pressure
19500627	#####	Anglian	160	Sand	25.70624	105/105	4.019108	19.85246						
19501033	#####	Anglian	220	Sand	25.59452	113/113	4.841667	20.60674						
19500458	#####	Anglian	31	Magnesiur	29.72039	146/146	3.281778	25.30769	Sand	25.70624	9/146	4.019108	19.85246	
19501508	#####	Anglian	16	Silt & Clay	63.60782	181/181	1.420635	54.50476	Phosphate	14.42852	3/181	3.22E-02	1.163447	
19500585	#####	Anglian	31	Magnesiur	29.72039	146/146	3.281778	25.30769	Sand	25.70624	9/146	4.019108	19.85246	
19500625	#####	Anglian	50	Magnesiur	19.20711	2/132	2.405751	15.03553						
19500046	#####	Anglian	231	Hardness	493.8317	11/122	56.50911	390.7496						
19501063	#####	Anglian	98											
19500022	#####	Anglian	231	Hardness	493.8317	11/122	56.50911	390.7496						
19501068	#####	Anglian	99	Sand	25.59452	140/140	4.841667	20.60674						
19500020	#####	Anglian	20											
19501129	#####	Anglian	49											
19500586	#####	Anglian	17	Silt & Clay	59.92633	148/148	1.328244	55.36449						
19500951	#####	Anglian	169											
19500205	#####	Anglian	231	Hardness	493.8317	11/122	56.50911	390.7496						
19501256	#####	Anglian	191											
19500704	#####	Anglian	54	Manganes	311.3038	3/271	15.52183	204.2972	Sulphate	292.3473	2/271	9.2775	177.7952	
19501594	#####	Anglian	201	Iron Dissol	510.1445	5/178	36.84365	336.5581	Manganes	341.6512	2/178	19.62259	200.5194	Zinc Disso
19500169	#####	Anglian	202	Hardness	493.8317	6/170	56.50911	390.7496	Calcium T	156.6335	10/170	14.3504	116.8532	Magnesiur
19501220	#####	Anglian	56	Iron Total	1431.075	141/141	245.5025	1037.831	Silt & Clay	63.60782	1/141	1.420635	54.50476	
19500646	#####	Anglian	232											
19501536	#####	Anglian	110	Magnesiur	31.77655	2/101	3.933041	25.87028	Potassium	12.07711	8/101	1.575023	9.340227	
19500330	#####	Anglian	171											
19501381	#####	Anglian	155	Sulphate	327.4052	78/143	11.24451	198.1857	Sodium To	231.993	2/143	8.816591	57.22475	Chloride lo
19500179	#####	Anglian	48	Nickel Tot	18.58994	46/141	1.828125	8.786378	Nickel Dis:	17.84954	5/141	1.688284	7.740284	Phosphate
19501230	#####	Anglian	75	Iron Total	1431.075	2/196	245.5025	1037.831						
19500596	#####	Anglian	31	Magnesiur	29.72039	146/146	3.281778	25.30769	Sand	25.70624	9/146	4.019108	19.85246	

Figure 10.7 Report from run in batch mode. Fields will depend on options selected. This example shows best cluster and outliers defined by 90th and 10th percentiles; columns G & L show total number of samples in best cluster and number of values for that variable.

User feedback workshop

A user feedback workshop was held at Staffordshire University on 7 May 2009. Six staff from the Environment Agency were present: Dr John Murray-Bligh (project manager), Christine Moore (project administrator), Caroline Howarth, Ian Humphreyes, Ben McFarland and Collette Sales. Progress made in the project was presented followed by testing of the new RPDS and RPBBN systems. Feedback on the new version of RPBBN is dealt with in Section 12 while feedback on RPDS is reported here. A similar workshop was held with David Colvill and Mark Hallard, Scottish Environmental Protection Agency, on 9 July 2009.

The following suggestions for improvements were received:

Data

- Use the same colour coding and number scale as used for WFD.

Interface

- The scale bar is helpful but some of the percentages need checking (there are instances where silt/clay is over 100 per cent).
- Text explanations are needed for the land risk scores.
- Text explanations are needed for indices and WFD classification (MINTA, LIFE).
- In the report page, proportions in WFD classes are headed by 'GQA Classification'.
- Sites tab has a field entitled 'RIVPACS GQA Class'.

- When printing the contents of a cluster, classification data output is GQA not WFD.

Batch mode and reporting

- When setting up the batch mode, suggest a default option and an edit facility for advanced users.
- Need to consider how much data formatting needs to be done to BIOSYS reports before being suitable for input to the system (the most simple and straightforward way is needed for the average ecologist).
- I need to be able to upload multiple sites together. Is there a way of uploading biological and environmental data directly from BIOSYS into RPDS? This would save an enormous amount of time, which from my previous experience of using the system is a fundamental barrier.
- Need to make the choices (such as stats elements) as straightforward as possible for the average ecologist. If it is overly complicated, the average ecologist won't have the time or confidence to use it. It's a great source of information, so I want as many ecologists as possible to be happy using it.
- Good guidance is needed, including how to interpret data.
- Present the number of samples with data/number of samples in cluster as a percentage.
- Maybe replace 'count' by 'number of samples in cluster'.
- For most parameters 'value' is a 'mean value in cluster'.
- Suggested cluster statistics: n, 10 per cent, 90 per cent, units.
- Maybe alter the descriptions of some of the chemicals (average alkalinity = three-year average alkalinity (CaCO₃, mg/l) and maybe provide information in a text box.
- The estimated time to completion is somewhat haphazard.

The more straightforward suggestions for improvement were implemented quickly, while some of the others such as compatibility of data formatting need coordination with other project teams. A suggested data format is given in Appendix C.

11 Revision and testing of Bayesian Belief Network Model

Review of the original network

An introduction to Bayesian Belief Networks (BBN) can be found in Walley *et al.* (2002). Briefly, a BBN is composed of two components: a network of cause-effect (or parent-child) links between variables and a set of conditional probability matrices that define the relationship of each variable to its causal variables. Each variable in the network has two or more possible states, and the conditional probability matrices define the probability of a variable being in each of its possible states, given the states of each of its causal variables. The probability matrices for variables that have no causal variables are just vectors of prior probabilities.

Both the causal network and probability values can be derived by expert opinion or objectively by analysing data. However, experts are generally more competent in understanding causality than in estimating the probabilities of a large number of combined events. Therefore, when developing the RPBBN model, expert advice was usually used to develop the causal structure and sample data was used to derive probabilities. The data requirement for probabilities was the limiting factor when building the model and only the most frequently sampled variables could be included in the final model, Figure 11.1. The prior and conditional probabilities were derived from the 1995 survey of rivers in England and Wales, consisting of spring and autumn samples for 3,615 sites with biological, environmental and chemical data.

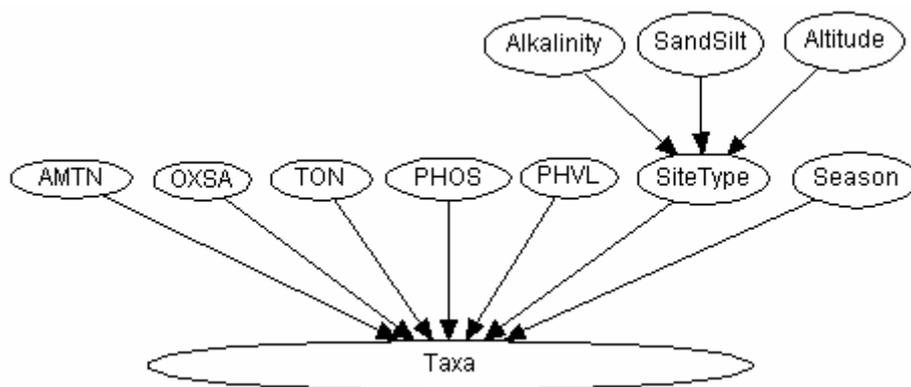


Figure 11.1 Causal belief network of original model.

After imposing a general cause-effect structure to the data, which assumed that the environmental and chemical variables are causal factors (parent nodes) of the states of biological variables (child nodes), the strengths of relationships between variables were assessed using multiple linear regression and mutual information. The strongest relationships with chemical variables were found to be those with total ammoniacal nitrogen (AMTN), dissolved oxygen (OXSA), phosphorus (PHOS), pH (PHVL) and total oxidised nitrogen (TOXN), all of which were well represented in the database in

terms of frequency of occurrence. The environmental variables were combined into one variable, site type, as in Walley *et al.* (1998), and together with season, these seven parameters formed the set of causal variables.

Before the construction of the network, several constraints were imposed on the variables:

1. Each child node (biological family) would have five parent nodes (causal factors).
2. Each chemical variable would have five possible states, with boundaries defined by the 15th, 35th, 65th and 85th percentiles of its distribution.
3. Each biological family would have four possible states (zero for absent, 1, 2 and 3+, except for two highly abundant families which were zero, 1+2, 3 and 4+).
4. Site type would have three possible states.
5. Season would have two possible states.

Season and site type were considered to be such key causal variables that they should be two of the five parents of each biological family, leaving three of the five chemical variables to be the remaining three parents. This resulted in a maximum size of the joint probability matrices of 3,000 (that is $4 \times 2 \times 3 \times 5 \times 5 \times 5$), with the values to be derived from 7,230 samples. The ratio of sizes of the dataset and joint probability matrices was an important consideration, and experience with earlier models suggested that it should not be less than 4:1. However, given that around 50 per cent of the combinations of states were highly improbable, a ratio of 2:1 was thought to be justified for RPBBN.

Although no restriction was placed on the number of child nodes that a parent variable can have (this doesn't affect the size of the conditional probability matrices required), the choice of which three of the five chemical variables to link to each biological family was not straightforward. Optimising the capabilities for diagnostic reasoning favoured connecting each chemical node to the 60 per cent of biological families with which the relationships were strongest. Optimising predictive reasoning, on the other hand, favoured connecting each biological family to the three chemical nodes with which the relationships were strongest. These two objectives were not compatible, for allocation on the basis of diagnostic capability resulted in some biological families being connected to all five chemical parents (and thereby breaking the limit on the number of connections permitted), while others were connected to none at all. The solution was based on a compromise, where the biological families were ranked separately for each chemical, based on the strength of the relationship between them, and the allocations made on the basis of the three highest rankings.

BBN Creator

BBN Creator is a software system created to automate some of the tasks involved in the construction of Bayesian Belief Networks and works in collaboration with the HUGIN Bayesian Belief Network development software. The application itself is a more refined version of software used as part of an early research project concerning river pollution in England and Wales.

The original software was used to produce a general pollution diagnostic BBN, which is the basis for the software package RPBBN (River Pressure Bayesian Belief Network). The software consisted of a number of VBA (Visual Basic for Applications) functions that were designed specifically to be used with the project data, Microsoft[®] Excel[®] and the HUGIN API (Application Programmers Interface). These functions automated

construction, data analysis and testing tasks, and significantly reduced network development time.

The need to use these functions on other data sets and third party interest led to the functions being made generic and grouped into a module. This BBN creation module could be imported into an Excel® workbook and used in conjunction with the data. Although the module was efficient and robust, it did require some familiarity with computer programming. Therefore, the next step in this development process was to provide a graphical user interface to simplify interaction with the existing module. The result is the BBN Creator application, which provides the benefits of the BBN creation module without the need for computer programming skills. Further details are given in the User Guide in Appendix G.

Revisions to causal network

In the course of this project, we anticipated making improvements to both the causal structure and the conditional probability matrices. The main constraint in the design of RPBBN was the limitation of 7,230 samples in the 1995 database. A database of five times that size would permit the probabilities of the current network to be estimated more accurately, or a further parent node to be added to each child node. Tests with the flow condition data (Section 7) showed significant effects on presence/absence probabilities, and its inclusion as a causal factor could improve the performance of the network.

During meetings with the Environment Agency project board, several improvements to the structure of the original causal belief network were suggested. The network agreed upon is shown in Figure 11.2, and the improvements are discussed below.

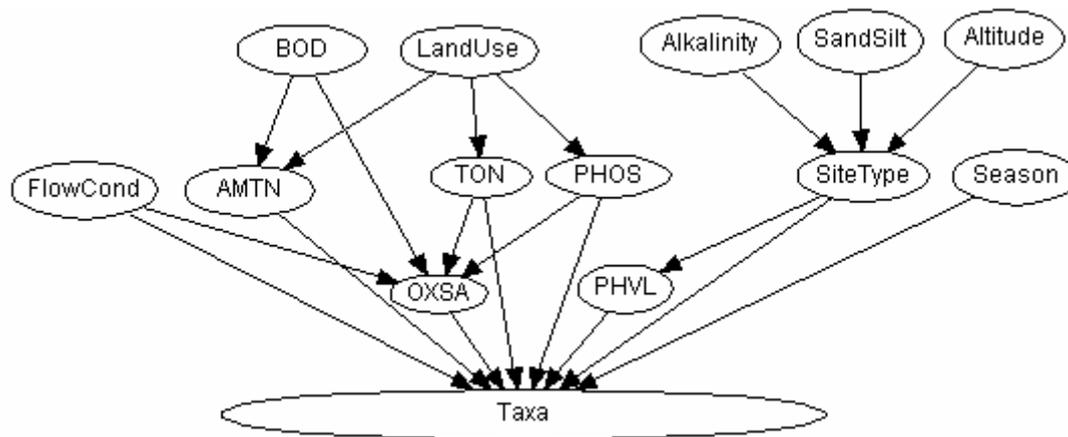


Figure 11.2 Revised causal belief network.

Addition of flow condition node

The addition of flow condition as a sixth parent node was possible because of the greater amount of data. As described earlier in this report, a value for flow condition was assigned to each sample by interpolation from monthly gauged flow records ranked over a thirty-year period from zero (driest) to one (wettest). A value was produced for periods of one, three, six, 12 and 24 months prior to the sample date. Values in the ranges $[0, 1/3]$, $[1/3, 2/3]$ and $[2/3, 1]$ were denoted 'dry', 'average' and 'wet' respectively.

Two solutions for adding flow to the network were considered.

1. Add a single flow node, which would be a parent to all taxa.
2. Add multiple flow nodes and connect each taxon to the one with which it has the strongest relationship.

There were two main areas in which these solutions differed; complexity of implementation and main beneficiaries from the relationships that would be introduced. The addition of any node would increase the complexity of the RPBBN causal network and the range of evidence required to fully exploit the model. Adding multiple nodes would therefore be the most complex option and would require more evidence to use them.

The prediction of a variable/node would benefit from having its strongest relationships modelled. A model with flow conditions for just one time period would mean all the taxa, including those with which it has the strongest relationships, would be linked to that node. This would be beneficial for prediction of the flow node, but not all taxa. The multiple flow node option, by contrast, would benefit the taxa, since they could be linked to the flow node with which they have the strongest relationship. However, prediction of individual flow variables would be expected to be worse than for their single node equivalents. The range of flow conditions that could be predicted would, of course, be much wider.

Both options had their advantages and disadvantages, so selecting an option became dependent on the initial criterion for including flow condition in the model. In our case, flow was introduced to provide additional supporting information when making predictions. Therefore, the impact of flow condition on the prediction of other variables is more important than on the prediction of the variable itself. Given this criterion, the multiple node option would be the better choice. It would be more complicated to implement but would provide better predictions for taxa, and other nodes in the network as a result.

Introduction of BOD and improved modelling of dissolved oxygen

The weakest variable in RPBBN, in terms of its predictive ability, was the dissolved oxygen, percentage saturation node (OXSA), and several changes were made to improve the performance of this variable. BOD was introduced and incorporated as a parent to AMTN and OXSA. TOXN and PHOS were also made parents of OXSA to account for the effects of eutrophication. One problem was that OXSA at night-time was thought to be the most important for invertebrates, whereas the data relates to measurements made during the daytime. However the extent of diurnal variation in oxygen concentration is determined partially by flow. In conditions of high flow, the effects of turbulence attenuate the variation through aeration (raising the concentration when low) and release (lowering when high), whereas in conditions of low flow these effects are much reduced. For this reason, flow condition was included as a parent of OXSA. The availability of at least three years' data prior to sampling enabled the use of the 10th percentile value for OXSA, which was also expected to improve performance. This was not possible with the original RPBBN, when all chemical values were based on mean values over the three months preceding the sample.

Site type and effect on pH

The environmental characteristics at a sampling site have a profound effect on the biological community found there, and they need to be taken into account to ensure that classifications of water quality from biological data are accurate. The modelling of site type in RPBBN was taken from Walley *et al.* (1998), where a value of unpolluted

ASPT was derived for each of the 1995 GQA sites and the range split into five roughly equal bands (1-5), where site type 5 represented fast flowing upland stream and site type 1 slow flowing lowland rivers. The classification was based upon the consensus of four different methods.

Three environmental variables dominated the prediction of ASPT: alkalinity, altitude and the percentage sand+silt. Once the 1995 GQA sites had been classified, the dataset was used to derive conditional probability matrices for a BBN version of the site type classifier, with the three variables as parent nodes. This was then used as an integral part of RPBBN.

It was proposed that the 'site type' variable in the new BBN should be based on the same method, with unpolluted ASPT generated by the new RIVPACS model based on the recently revised BMWP scores (Paisley *et al.*, 2007). Once the site types had been defined, a BBN model of site type could be derived using the three variables as the parent nodes, as before. However, delays in agreeing the revised scores in SEPA and Environment Agency meant this approach could not be adopted in the time available, and so the previous model was used.

The 'site type' node was made a parent to the 'pH' node to account for the causal relationship between the two variables (upland sites are often characterised by thinner, peaty soils which are more acidic). The use of the fifth percentile over the preceding three years for pH rather than its mean over three months was also expected to improve performance.

Inclusion of land cover

Although this was to be included in the improved model, land cover was omitted from the final model because of the constraints of time. The effects on the model were likely to be of secondary importance, and given the difficulties of defining variables and states, we decided that this aspect of the model required further research and should be left for future work.

Summary of dataset used

The original BBN model was based on a dataset of matched biological and chemical samples taken from 3,615 sites in the spring and autumn of 1995. Following the matching procedures required for chemical data, as well as the three environmental variables used as the basis for the 'site type' node, corresponding data for the revised spring and autumn BBN models is summarised in Tables 11.1 and 11.2. Note that the models include no SEPA data because of incorporation of the flow node, for which data was available for Environment Agency only.

The totals of 16,244 and 15,856 samples represent around 44 and 43 per cent of the total Environment Agency biological samples available (Tables 4.4 and 4.5), although the NW region seems particularly under-represented before 2000. The amount of data is more than four times that used in the original model. Table 11.3 summarises the corresponding number of sites at which samples were taken in each season. The totals are 56 and 57 per cent of those available (Table 4.6).

Table 11.1 Distribution of matched biological and chemical sample data by region and year available for revised BBN model, spring.

Year	Environment Agency								Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	
1995	431	176	10	345	283	628	322	638	2833
1996	413	123	1	425	122	83	25	33	1225
1997	440	106	0	390	131	161	72	2	1302
1998	410	134	1	668	155	61	117	7	1553
1999	411	131	9	357	164	43	173	5	1293
2000	287	343	265	598	315	623	305	615	3351
2001	7	4	4	30	43	5	49	2	144
2002	220	145	179	392	125	196	130	218	1605
2003	219	146	211	356	125	197	123	206	1583
2004	211	148	202	252	117	158	54	213	1355
Total	3049	1456	882	3813	1580	2155	1370	1939	16244

Table 11.2 Distribution of matched biological and chemical sample data by region and year available for revised BBN model, autumn.

Year	Environment Agency								Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	
1995	426	260	12	548	282	635	315	638	3116
1996	435	148	0	387	136	60	69	2	1237
1997	419	114	1	603	141	64	52	0	1394
1998	421	136	0	648	150	60	112	1	1528
1999	250	125	90	66	165	45	173	0	914
2000	224	242	167	492	206	547	268	430	2576
2001	122	74	63	344	76	57	61	11	808
2002	208	146	156	289	112	170	101	216	1398
2003	219	145	200	305	124	183	112	212	1500
2004	205	131	219	252	124	159	80	215	1385
Total	2929	1521	908	3934	1516	1980	1343	1725	15856

Table 11.3 Distribution of sites for matched biological and chemical sample data for revised BBN.

Season	Environment Agency								Total
	ANG	NE	NW	MID	SO	SW	TH	WEL	
Spring	451	369	442	799	352	695	328	672	4108
Autumn	451	378	437	798	347	692	328	664	4095

Derivation of conditional probability matrices

Conditional probability matrices were derived from the datasets described in Tables 11.1 and 11.2. However, two potential difficulties can arise if raw values are used: they generally contain some zero-valued probabilities which need to be eliminated and the distributions can be 'lumpy' if the total number of cases in the dataset is not large compared to the number of elements in the matrix.

Zero-valued probabilities correspond to states that are impossible, and in the worst case can prevent the BBN algorithm from functioning. It was found previously that, even when the algorithm continued to function, its performance improved when the value of the probability for such states was set to a small non-zero value. This was achieved in the original BBN model by the addition of small residual values, and this technique was used again here.

Probability distributions were smoothed in the original BBN model by adjusting the parameters of a redistribution algorithm so that the predictive performance of the model was optimised. Although a similar technique was tested for the new model, the use of smoothed distributions consistently resulted in worse performance so the raw distributions were used instead. The most likely reason for this is that, as well as reducing inconsistencies, smoothing also diminishes the characteristics of the distribution and the distribution flattens out as the amount of smoothing increases. It is likely that the greater amount of data produced better original distributions that only deteriorated when smoothed.

Testing and evaluation

RPBBN 2.0 differed from the original RPBBN 1.0 in several respects, namely:

- Conditional probability matrices were based on more than four times as much data.
- Chemical statistics were based on percentile values over the three years prior to the sample date rather than mean values over the prior three months.
- The structure of the model was changed.
- Five states were used for the taxonomic variables rather than four.

Preliminary dependent testing

Tables 11.4-11.6 present the results of preliminary dependent tests of RPBBN 2.0, incorporating all the changes outlined above, against the original RPBBN 1.0. Dependent testing refers to testing based on the use of the same dataset used to build the model, whereas independent testing refers to testing based on a separate dataset. Although these results give an indication of the combined impact of the changes, independent testing was used to evaluate the impact of each separately, and these results are described later.

Spearman rank correlation coefficients in Table 11.4 indicate major overall improvements in the performance of the network for total ammoniacal nitrogen and dissolved oxygen, a lesser but notable deterioration for total oxidised nitrogen, and minor changes for phosphorus and pH. Predictions of BOD5, which was not present in the earlier network, are not as good as any of the other variables in the new network, but are still better than predictions of total ammoniacal nitrogen in the earlier version. Predictions of flow condition variables are all much poorer than those for the chemical variables, but this is unsurprising given that each flow condition node is connected to much fewer biological taxa.

Table 11.4 Results of dependent tests on RPBBN 2.0 against RPBBN 1.0 expressed in terms of Spearman rank correlation coefficients between predicted and recorded values of chemical variables. Note that BOD5 and flow condition nodes were not included in RPBBN 1.0.

	RPBBN 1.0	RPBBN 2.0	Change
Total Ammoniacal Nitrogen	0.5620	0.6629	0.1009
Phosphorus	0.7008	0.6947	-0.0061
Dissolved Oxygen	0.6794	0.7409	0.0615
pH	0.6889	0.6992	0.0103
Total Oxidised Nitrogen	0.7412	0.6991	-0.0421
BOD5	-	0.5843	-
Flow Condition (3 months)	-	0.2542	-
Flow Condition (6 months)	-	0.3248	-
Flow Condition (12 months)	-	0.2330	-
Flow Condition (24 months)	-	0.3374	-

Tables 11.5 and 11.6 provide further quantitative data on performance. Table 11.5 quantifies the accuracy and certainty of the predictions and should be interpreted with reference to the bandings of variables and their prior probabilities given in Table 11.6. This indicates that prior probabilities are approximately evenly distributed for all variables except flow condition, for which the distributions are increasingly biased toward the middle of the three bands as the time period prior to sampling increases.

Table 11.5 indicates that for five of the chemical variables (those in the previous network) the predicted band with highest probability agrees with the actual band in more than half the cases, while for BOD (not in the previous network) agreement is achieved in more than a third of cases. The match is best for the flow condition nodes, although this is as a result of the bias toward one state ('average') in the prior probabilities, which increases as the time period prior to sampling increases.

Table 11.5 Performance characteristics of RPBBN 2.0.

	Spearman's Rank	% with Correct Highest Prob	Mean Highest Prob	SD Highest Prob
Total Ammoniacal Nitrogen	0.6629	52.17	0.6748	0.2253
Phosphorus	0.6947	54.33	0.6943	0.2194
Disolved Oxygen	0.7409	56.64	0.7239	0.1954
pH	0.6992	52.80	0.6824	0.2097
Total Oxidised Nitrogen	0.6991	53.89	0.7029	0.2170
BOD5	0.5843	36.76	0.3867	0.1081
Flow Condition (3 months)	0.2542	56.04	0.6214	0.1700
Flow Condition (6 months)	0.3248	60.66	0.7140	0.1827
Flow Condition (12 months)	0.2330	74.43	0.7944	0.1438
Flow Condition (24 months)	0.3374	79.38	0.8763	0.1366

Table 11.6 States of variables in Tables 11.4 and 11.5 together with their prior probabilities.

Total Ammoniacal Nitrogen	0-0.09 0.2737	0.09-0.17 0.2256	0.17-0.29 0.1900	0.29-0.6 0.1664	0.6-43.89 0.1444
Phosphorus	0-0.05 0.2549	0.05-0.16 0.2108	0.16-0.43 0.1959	0.43-1.2 0.1784	1.2-15 0.1600
Disolved Oxygen	0-67 0.1682	67-78 0.2142	78-86 0.2704	86-92 0.2280	92-119.2 0.1192
pH	0-7 0.0874	7-7.4 0.2239	7.4-7.65 0.2594	7.65-7.85 0.2537	7.85-9.3 0.1756
Total Oxidised Nitrogen	0-2 0.1727	2-5 0.2452	5-7.5 0.2261	7.5-10.5 0.2032	10.5-28.06 0.1528
BOD5	0-1.85 0.1459	1.85-2.45 0.2476	2.45-3 0.2238	3-4.15 0.2042	4.15-154 0.1785
Flow Condition (3 months)	0-0.3333 0.2583	0.3333-0.6666 0.5043	0.6666-1 0.2373		
Flow Condition (6 months)	0-0.3333 0.2445	0.3333-0.6666 0.5414	0.6666-1 0.2142		
Flow Condition (12 months)	0-0.3333 0.1735	0.3333-0.6666 0.7056	0.6666-1 0.1208		
Flow Condition (24 months)	0-0.3333 0.1379	0.3333-0.6666 0.7441	0.6666-1 0.1181		

Independent testing

For the independent tests, the dataset was split so that 90 per cent of the data was used to produce the models while the remaining tenth was used as a test dataset. Tests with the original network with the original dataset, original chemical statistics and four states for the biological variables provided baseline data (Test 1). Following this, three further sets of tests were undertaken designed to quantify changes caused by the greater amount of data (Test 2); varying the chemical statistics used (Test 3) and increasing the number of states (from four to five) for biological variables (Test 4). Details of the tests are provided in Table 11.7.

Table 11.7 Description of independent tests.

Test	Purpose	Network	Dataset	Chemical Statistics	No of Bio States
1	Baseline	Original	Original	Three-month means	4
2	Effect of increased dataset	Original	New	Three-month means	4
3	Effect of increased dataset and new chemical statistics	Original	New	Three-year percentiles	4
4	Effect of increased data, new chemical statistics and five biological states	Original	New	Three-year percentiles	5

The results of each test are given in Table 11.8. The apparent deterioration between the results of Test 1 with those for RPBBN 1.0 in Table 11.4 is typical of the difference obtained when employing independent and dependent testing. Compared to the results of Test 1, on the other hand, the results of Tests 2, 3 and 4 indicate that each of

the changes outlined in Table 11.7 produced an overall increase in predictive performance of the network.

Changes caused by the greater amount of data (Test 2) are relatively modest; the largest is an increase of 0.0338 in the prediction of phosphorus (although there is a decrease for total ammoniacal nitrogen, the change is so small as to be insignificant).

Two of the changes caused by variation in the chemical statistics used (Test 3), on the other hand, are much more significant, the largest of which is an increase of 0.2413 for dissolved oxygen. Predictions of dissolved oxygen in the original network were the poorest of the five chemicals, whereas in the new network they appear to be among the best. A large factor is likely to be that the statistic used in the new network (value of the 10th percentile over the three years prior to sampling) is of much greater ecological significance than that used in the original network (mean value over the preceding three months). In addition, the new network has more relationships with OXSA, and with the connection to BOD there are more pathways for the propagation of evidence. The smaller but significant increase for total ammoniacal nitrogen of 0.1203 is also likely to be caused by the change in statistic used, now the value of the 90th percentile rather than the mean value, as well as the connection with BOD.

Another factor may explain the improved performance of these two chemicals over the other three. This is the way that connections are made from biological families. The strength of the relationship between each chemical variable and each taxon is quantified by mutual information, and each taxon is then linked to the three chemicals with which it has the strongest relationships, based on ranking. A modification to the algorithm used for this resulted in dissolved oxygen and ammoniacal nitrogen being allocated links from more taxa than the other chemicals. A further test is required to quantify this effect.

By contrast, changes caused by greater numbers of bands for the biological families (Test 4) are rather modest, with the largest change of 0.0347 recorded for phosphorus.

Table 11.8 Results of independent tests on RPBBN 1.0 expressed in terms of Spearman rank correlation coefficients between predicted and recorded values of chemical variables.

Test	Total Ammon. Nitrogen	Phosphorus	Dissolved Oxygen	pH	Total Oxid. Nitrogen
1	0.4720	0.6058	0.3824	0.4825	0.5848
2	0.4662	0.6397	0.4029	0.4844	0.6002
Difference	-0.0058	0.0338	0.0205	0.0020	0.0155
3	0.5865	0.6463	0.6442	0.5284	0.6168
Difference	0.1203	0.0067	0.2413	0.0440	0.0165
4	0.6157	0.6810	0.6506	0.5585	0.6444
Difference	0.0291	0.0347	0.0064	0.0301	0.0276

Extensions to functionality of RPBBN

Table 11.9 lists requirements drawn up following consultation with potential users of the system. As with the additional functionality required for RPDS (Table 9.1), some of the requirements relate to easier use of the software while others refer to tasks needed for the Water Framework Directive. These requirements (along with those for RPDS) were implemented during two extensions to the original project. The work relating to the modification of the RPBBN software is described in the next section.

As well as improvements to the functionality of the software, there is considerable scope for the future development of the model itself. Initial steps have been taken to include a variable indicative of flow, but further research is needed to optimise the use of this variable, as well as other variables such as land cover.

Table 11.9 List of requirements for additional functionality to RPBBN.

Deliverables
1. Changes to the chemical bands to cover the WFD and River Ecosystem classification boundaries. The suggested bands have been included in a table at the end of this document.
2. Include a list of predicted taxa (plus their predicted abundance category) as an extra page in RPBBN. Predicted taxa list need to be exportable either as CSV or Excel format. Below each list, output the calculated biotic indices. Headers to include sample ID/date and possibly waterbody/catchment. The report screen could have the option to include environmental parameters alongside the predicted taxon.
3. Batch input and output processing – particularly batch input. It would be a massive help to be able to re-input a site quickly (by downloading from an input file) without having to go through all the charts every time. Similarly, when reporting predicted indices (such as WFD class) to batch, output these – so you can get a list of, say, 30 sites with their predicted class. Standard format for data files (to link with RIVPACS, BIOSYS).
4. Show the changes graphically in taxon and physical/chemical plots from default position of probabilities to new predicted values. It would be better if the report screen showing the predicted list of taxa/abundance and indices could do the same, comparing the change in score and taxa.
5. Helpful to be able to toggle views between the report screen or taxon probabilities view panes when adjusting environmental data.
6. Include confidence of predictions based on the number of samples used to generate the probabilities.
7. Change the abundance categories from a four group system to at least a five (0, 1, 2, 3, 4+).
8. Include predictions of both new (WHPT) and old BMWP scores. Also calculate Number of scoring taxa, average score per taxon, and family LIFE scores. These scores need to be validated against observed scores derived from the taxon samples held in BBN.
9. Allow the user to input a predicted single season score from RIVPACS so RPBBN can produce an EQI, and a predicted 'classification' - using high, good, moderate, fair and poor ecological WFD status. For scenario testing.
10. At end of March produce a stand-alone CD that will run on a non-agency laptop of the RPBBN program and associated runtime modules. This will initially form an important component of the Water Quality Toolkit developed in Southern Region and produce test bed for future development.
11. Add WFD classification so that effect physical and chemical changes on WFD class can be observed. This will require a field for entering reference values (predictions) for classification metrics ASPT, N-taxa and WHPT and a facility to specify the classification metrics (swap from ASPT to WHPT) and to change the class boundaries – maybe via a configuration page.
12. A one-day workshop to explain the systems developed in this project and the EMCAR project to managers in water resources, river basin planning and monitoring policy.

12 Modifications to RPBBN software

Introduction

This section covers work to enhance the functionality of RPBBN, the requirements for which are given in Table 11.9. Some were modified during the project. Further refinements were implemented during two extensions to the main project. The sections below reflect the revised work items.

Changes to RPBBN model

These work items dealt mainly with changing the underlying RPBBN model to improve the range of its predictive abilities. In both the previous Environment Agency R&D project (E1-056) and the original EMCAR project, one definitive version of the RPBBN model was envisaged. However, whilst a single model might have the broadest appeal, it is not well suited to performing any one task. Therefore, when an application for the RPBBN system is found it is often necessary to adapt the RPBBN model to improve performance.

This section covers modification work in the two extension projects. Due to the creation of additional RPBBN models, a model-naming scheme is introduced to simplify and clarify the referencing of models. The following are names and descriptions of the models referred to:

1. RPBBN-P1 (Project 1): Model developed in the initial Environment Agency R&D project (E1-056).
2. RPBBN-P2 (Project 2): Initial model developed for this project, EMCAR Project EMC/WP06/077.
3. RPBBN-S (Southern Region): Model developed for the Environment Agency's Southern Region to help in their assessment of the impacts of developments in Ashford, Kent.
4. RPBBN-A (Annual): Model developed to produce annual, combined seasonal (spring/autumn) predictions compatible with values used to derive WFD quality classifications.

In what follows, note that RPBBN-(number) refers to a version of the software, whereas RPBBN-(letter) refers to a version of the model. Hence, software versions RPBBN1.x (created during project E1-056) all use model RPBBN-P1, whereas software versions RPBBN 2.x (created during this project) all use models RPBBN-P2, RPBBN-S and RPBBN-A.

Changes to abundance categories

The RPBBN-P1 and early versions of RPBBN-P2 used just four abundance categories: absence and three categories of abundance, which differed between taxa depending on the ranges of abundances in which they usually occur. In the revised versions used in the first extension project, the number of abundance states for each taxa were increased to five. These five states were zero or absent, 1-9, 10-99, 100-999 and 1,000-99,999, which correspond to the abundance categories 0, 1, 2, 3 and 4+.

As with the changes in the number of states for chemical variables, there were concerns about the effect of this increase in states on the quality of probabilities. However, it was felt that with the sevenfold increase in amount of data available, it was possible to make these changes and retain some confidence in the quality of generated probability values.

Changes to chemical bands

BBNs are limited in their ability to represent continuous variables so all the variables in the RPBBN model need to be discretised, that is, to have a set of discrete states defined. In RPBBN-P2 states were defined using percentiles to ensure that samples in the project database were well distributed over the selected states. The aim of this approach was to ensure the quality of probabilities derived from the sample data. A drawback to this approach was that the boundary values chosen for the RPBBN model did not correspond to chemical boundaries of WFD or River Ecosystem (RE) classifications used for river management.

The purpose of this work was to assess the feasibility of using WFD and RE chemical bands and if possible, modify the states of relevant chemical variables. Before modifying the states of chemical variables, it was necessary to check there would be sufficient numbers of samples within the new bands to maintain an adequate precision in the probability estimates that would be derived.

Table 12.1 shows the chemical bands that were suggested initially and the distribution of samples in the project database across them. There are wide variations in the distribution of samples of orthophosphate, oxygen percentage saturation and biological oxygen demand (five-day), all having around 18,000 samples in one of the bands, usually at one end of the distribution, and much fewer samples in the remaining bands. The distribution of samples for total ammonia is a little better but there are still over 10,000 samples in the 0.1-0.199 band. Samples of total oxidised nitrogen appear to have the best distribution, despite the variation in the size of first four bands, which causes a fluctuation in the values.

The main problem with using the initial sets of states was that there were simply too many. In a BBN, probability values are required for every possible combination of states of the variable itself and all its causal/parent variables. Any increase in the number of states and/or parent variables leads to a corresponding geometric increase in the number of probabilities.

Table 12.1 Chemical bands suggested initially and distribution of samples across them; red are WFD limits and yellow are RE boundaries beyond WFD bands.

Orthophosphate (OPhos)		Total ammonia (AmTN)		Oxygen percentage saturation (OxSa)		Five-day biological oxygen demand (BOD5)		Total oxidised nitrogen (TOxN)	
Bands	No. of Samps	Bands	No. of Samps	Bands	No. of Samps	Bands	No. of Samps	Bands	No. of Samps
0 – 19.99	3,158	0 – 0.049	3,703	0 – 19.99	22	0 – 2.49	16,701	0 – 0.99	3,240
20 – 29.99	1,972	0.05 – 0.099	7,655	20 – 49.99	1,227	2.5 – 2.99	7,684	1 – 2.49	5,242
30 – 39.99	1,815	0.1 – 0.199	11,029	50 – 59.99	1,557	3 – 3.99	8,501	2.5 – 2.99	1,577
40 – 49.99	1,516	0.2 – 0.249	3,530	60 – 69.99	3,526	4 – 4.99	3,981	3 – 3.99	3,101
50 – 79.99	3,209	0.25 – 0.299	2,669	70 – 74.99	3,004	5 – 5.99	2,114	4 – 4.99	3,110
80 – 99.99	1,611	0.3 – 0.599	6,777	75 – 79.99	4,058	6 – 7.99	1,606	5 – 7.49	8,851
100 – 119.99	1,526	0.6 – 1.299	3,526	80 – 119.99	18,559	8 – 14.99	761	7.5 – 9.99	6,880
120 – 199.99	4,117	1.3 – 2.499	1,540	120 +	0	15 – 49.99	60	10 – 24.99	7,033
200+	18,989	2.5 – 4.99	716			50+	9	25 – 99.99	12
		5 – 9.99	219					100+	0
		10+	101						

In RPBBN-P2, chemical variables only have five states and the number of possible states for macroinvertebrate families is 37,000 (AmTN (5 states) x TOxN (5) x OPhos

(5) x site type (5) x season (4) x flow (3) x taxon (5) = 37,000). Using the states suggested, the smallest number of possible probability values required for a macroinvertebrate variable would be 297,000 (AmTN (11 states) x TOxN (10) x OPhos (9) x site type (5) x season (4) x flow (3) x taxon (5) = 297,000). This is eight times as many as that of the existing model and eight times the number of samples in the database that would be needed as the source of their values. We therefore decided that the number of suggested states would need to be reduced.

Table 12.2 gives a list of modified bands used in the revised model. The number of states was reduced by merging bands, usually those at the extremes. All variables were reduced to seven states, except total ammonia, which was reduced to eight. This process removed a total of 11 bands; of which two were WFD limits for orthophosphate and three were RE boundaries for total ammonia, oxygen percentage saturation and BOD5. The merged bandings tended to be those with lower numbers of samples and so would be those that generate probability values with the lowest confidence.

The reduction in the number of states meant that the new maximum number of probability values had reduced to 117,600, still three times as many as those in the current model and requiring three times more than the total number of samples in the project database. Although this was far from ideal, it offered a reasonable compromise between providing some confidence in the probabilities generated and its operational usability. These changes, however, imply a greater possibility of inconsistent probabilistic predictions in extreme conditions. Smoothing the probability distributions may resolve some of the problems with inconsistency as it did in the development of the RPBBN-P1 model (see Environment Agency R&D Technical Report E1-056/TR). However, there was insufficient time to investigate this.

Table 12.2 Chemical bands used in RPBBN-S and distribution of samples across them; red are WFD limits and yellow are RE boundaries beyond WFD bands.

Orthophosphate (OPhos)		Total ammonia (AmTN)		Oxygen percentage saturation (OxSa)		Five-day biological oxygen demand (BOD5)		Total oxidised nitrogen (TOxN)	
Bands	No. of Samps	Bands	No. of Samps	Bands	No. of Samps	Bands	No. of Samps	Bands	No. of Samps
0 - 29.99	5,130	0 - 0.1	11,358	0 - 49.99	1,249	0 - 2.49	16,701	0 - 0.99	3,240
30 - 49.99	3,331	0.1 - 0.199	11,029	50 - 59.99	1,557	2.5 - 2.99	7,684	1 - 2.49	5,242
50 - 79.99	3,209	0.2 - 0.249	3,530	60 - 69.99	3,526	3 - 3.99	8,501	2.5 - 4	4,678
80 - 99.99	1,611	0.25 - 0.299	2,669	70 - 74.99	3,004	4 - 4.99	3,981	4 - 4.99	3,110
100 - 199.99	5,643	0.3 - 0.599	6,777	75 - 79.99	4,058	5 - 5.99	2,114	5 - 7.49	8,851
200 - 299.99	3,194	0.6 - 1.299	3,526	80 - 99.99	18,513	6 - 7.99	1,606	7.5 - 9.99	6,880
300+	15,795	1.3 - 2.499	1,540	100+	46	8 +	830	10+	7,045
		2.5 +	1,036						

The principle motivation for making changes was to make RPBBN better suited for setting ecological standards, initially for use in the Environment Agency Southern Region's work on a major development in Ashford, Kent. As a result, the network became known as the 'Southern Region' or just 'Southern' RPBBN model and was named RPBBN-S.

Two-season BBN

Environment Agency staff had been asked to predict the biological outcomes of programmes of measures and other activities on the WFD ecological status during the implementation of the first river basin management cycle. A preliminary exercise to predict ecological status in 2015 was undertaken in 2008, based on measures and activities identified in the first draft River Basin Management Plans. This was repeated in 2009 for the definitive River Basin Management Plan.

The quality classification of rivers in the UK is based on two samples taken in spring and autumn. These samples are combined to produce a pooled composite, two-season or 'annual' sample, which is used for classification. The RPBBN model, on the other hand, is based on individual samples and may include several samples taken from a site during a year. These differences between the two data sources meant that the existing RPBBN model was unable to predict the appropriate 'annual' value¹⁰, and that a two-season version was required.

The two-season BBN was based on RPBBN-P2, with five categories for most environmental variables. This better matched the precision of predictions of chemical concentrations. Also, for biological outcomes, predictions for a few thousand sites would need to be generated and the RPBBN-P2 model was much quicker than the RPBBN-S model. Data from many sites would normally be input into this model in a table and the output would also be a table.

To enable the RPBBN system to predict two-season values, the existing project data needed to be revised to produce two-season samples and the probabilities for a revised, 'seasonless' RPBBN model generated from them. Fortunately, much of the necessary data preparation had already been completed during the creation of the RICT classification dataset (see the 'Data Tasks' section of Section 10). This dataset contained individual spring and autumn samples for sites sampled in both seasons. The following procedures were carried out to modify the existing dataset into that needed to create the two-season model.

- For every site, for each year that there was data, the maximum abundance value over the two seasons was calculated for each taxon.
- The autumn chemistry values were attached to annual biology data.
- Finally, existing annual and fixed values, like mean annual substrate, stresses survey data and environmental parameters were also attached.

The final dataset contained over 15,200 records for 4,100 sites. Biological data, including abundances and biotic indices, were based on two-season combined samples.

The main change required to the causal structure of the model was removal of the season variable. Flow categories were also reduced to the two longest periods (conditions over previous 12 and 24 months) to simplify and speed up the model, but all remaining biological and environmental variables were retained. The revised model was then populated with probabilities taken from the two-season dataset. Although the two-season dataset contained a little less than half the number of samples in the original dataset, removal of the season variable halved the size of probability matrices of biological variables. This maintained the ratio of probability matrix size to training samples and confidence in the quality of the resulting probability distributions. Once the new probability distributions had been generated, the two-season RPBBN was packaged with the RPBBN 2.2 software. This version of the RPBBN model was named RPBBN-A, with the 'A' standing for annual¹¹.

Changes to the user interface

The aim of these changes was to organise the information produced by the RPBBN model more effectively and to simplify interaction with the model. The original RPBBN software lacked any reporting facilities other than on-screen bar charts. Introduction of

¹⁰ The existing RPBBN model can predict 'annual' values, by simply excluding evidence on season. However, the predictions would still be based on the 'wrong' underlying data.

¹¹ The suffix 2S for two-season was considered but it was felt that this might be confused with RPBBN-S, the acronym for the southern RPBBN model.

additional report information and a report panel were therefore key developments for the new version.

Include a list of predicted taxa

Producing a list of predicted taxa and predicted biological indices are two different processes and both raise different types of problems. However, they are both caused by the outputs of a BBN being probability distributions, which are multi-valued and do not provide a prediction of the value of any parameter. Reporting this information in an accessible and easy to understand format was one of the main challenges in producing outputs for the RPBBN 2.1.

One of the first problems in generating a list of predicted taxa was how to reduce the information output from RPBBN 2.1 into a more manageable format. The RPBBN model can output up to 500 probability values every time it is updated and it was clear early in the project that the number would have to be reduced to keep the report succinct and not to overwhelm the user with information.

The solution that was chosen was to reduce each probability distribution to the most likely state. This was perhaps the most important probability value, and this approach had been used in the evaluation of the original RPBBN model (see Environment Agency R&D Technical Report E1-056/TR). Obviously, this considerably reduced the amount of available information being reported, but it made the report more succinct and accessible. Reporting only the most likely state reduced the number of values that had to be output from a maximum of 500 to 97.

Include predictions of biological indices

Predicting values of biological indices raised a different set of problems, as they required the probability distributions to be converted into a set of biotic index values. Initially, we considered simply using the most likely state as the predicted state for the taxa. Problems justifying the selection of this state as a firm prediction when its probability was low led to this approach being abandoned.

The solution eventually chosen was to use a weighted value, similar to that used in testing the original RPBBN model (see Environment Agency R&D Technical Report E1-056/TR). A weighted mean of the values for each state was calculated and the probabilities were used as the weights as follows:

$$\text{predicted_value} = \sum_{i=1}^N p_i v_i$$

where:

N = number of states of variable

p_i = probability of variable being in i^{th} state

v_i = value of i^{th} state.

To calculate predicted index/score values, the term v was substituted by the presence or abundance-related score for the taxon. One of the main benefits of this approach was that the whole probability distribution was used to produce the final values. However, this method does affect the results and how the indices function. For example, the 'score' for an individual taxon might vary from sample to sample and the value tends to be fractional. Another consequence is that, because every taxon has some probability of occurring, every taxon contributes to the generation of the final sample score.

In addition to generating index scores, it was necessary to predict the number of taxa that might occur. As this value would be used alongside predictions of index/score, a

modified version of the weighted value was thought to be best. Hence, the predicted number of taxa for a sample is the sum of the probability of states associated with each taxon being present.

These methods of deriving index values and numbers of taxa are thus able to accurately reproduce assessment values based on actual data (evidence derived from a sample).

Include a report panel

The report panel is one of the three main panels in RPBBN 2.1 software, the others being charts and records. The report panel, shown in Figure 12.1, is a rich text format document that can display a list of predicted states and, where applicable, indices for each node in the RPBBN models. It can also provide a 'sample' summary, which consists of index values and other information based on the current state of all taxa. In addition to the on-screen display, a hard copy of the report can be obtained by using the 'Print Reports ...' option in 'File' menu.

RPBBN 2.1 provides several options for customizing the report through the 'Properties ...' option under 'Report' menu. These options include removing parts or the entire header, including the probability of the most likely state, and removing from the list all taxa predicted as absent (shown in Figure 12.1).

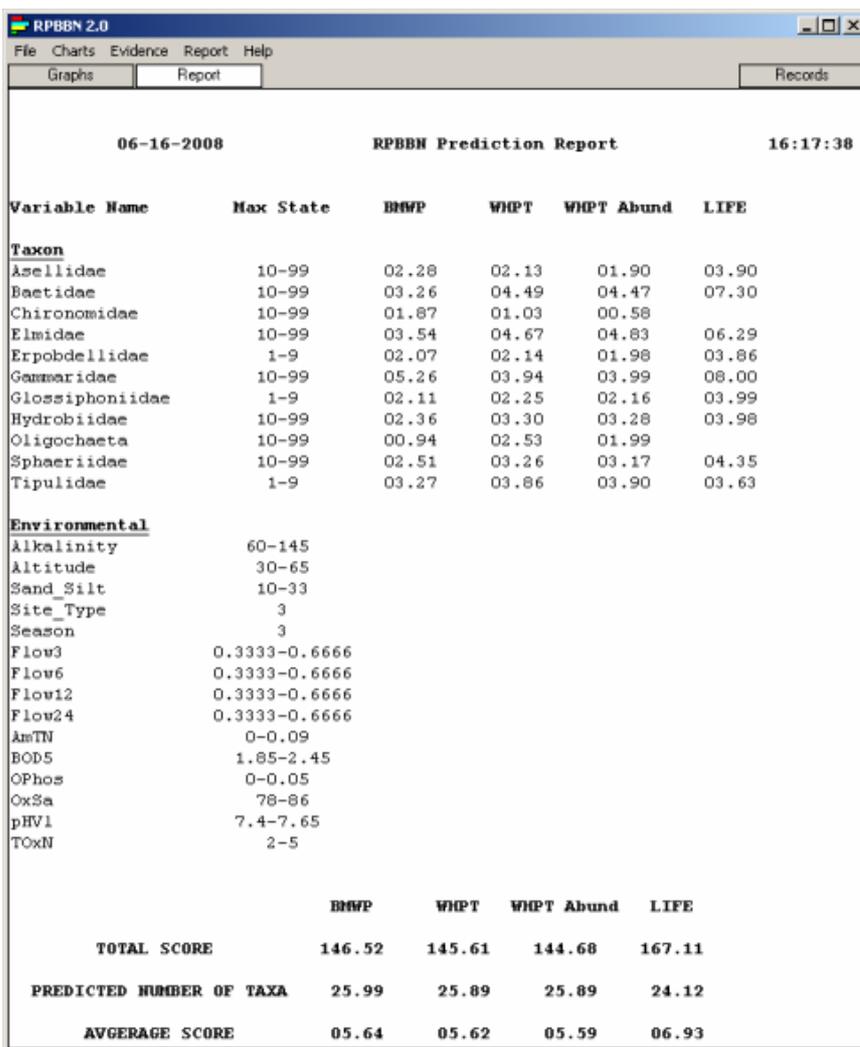


Figure 12.1 Screen-hot of report display panel showing information on predictions and index values for individual variables and summary for whole network.

Show the changes from default to new predicted values

The RPBBN 1.2 had no inbuilt method of tracking changes when evidence in the model was modified. It was left to the user to note probabilities before the change was made. In RPBBN 2.1, this is no longer necessary, as it includes a 'Store Current State' option under the 'Charts' menu. Selecting this causes the current states to be stored and they appear both graphically, as thin dark green bars displayed alongside the current probability bars and as text, as values in brackets to the left of the current probabilities (see Figure 12.2).

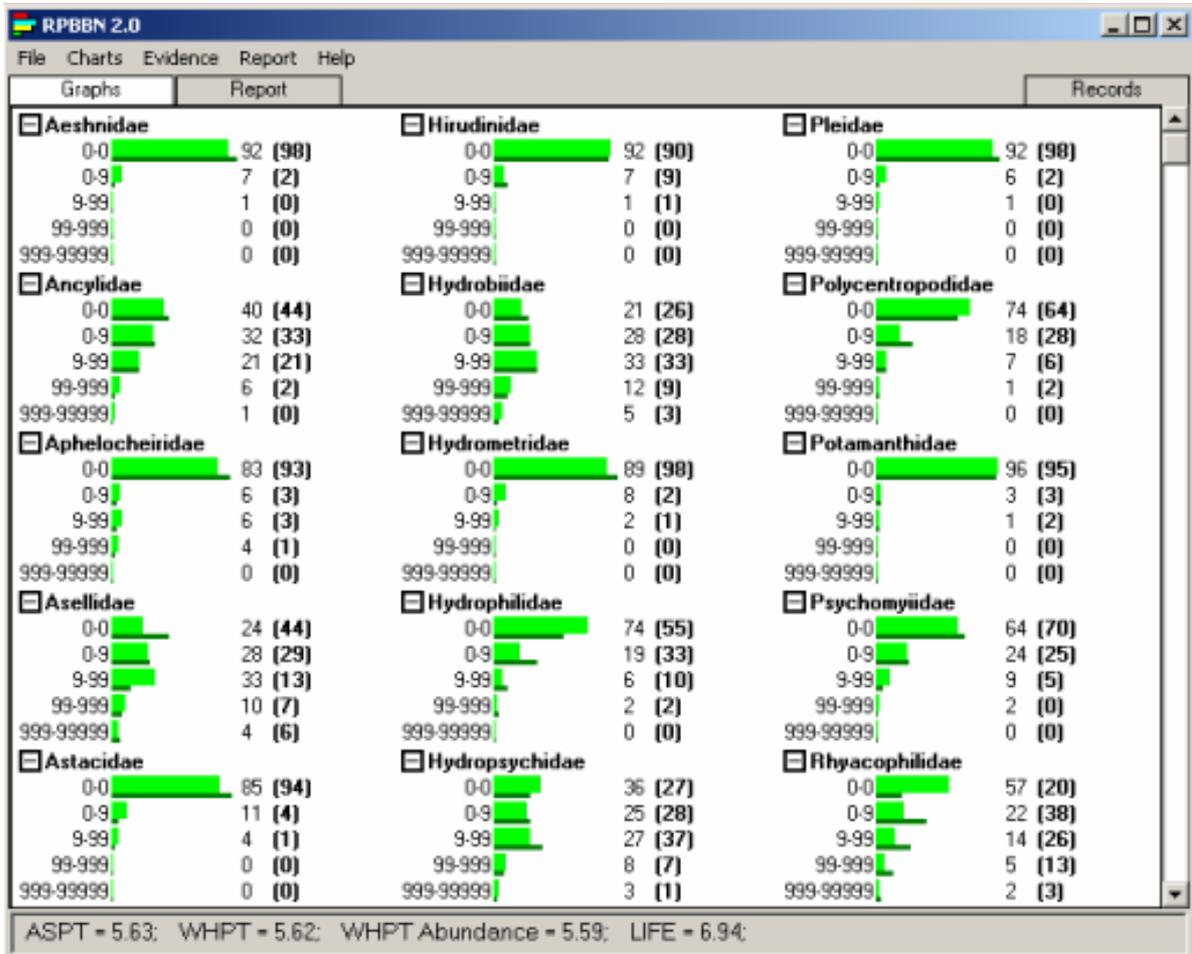


Figure 12.2 Screenshot of charts display panel with current and stored states.

Batch mode

In RPBBN 1.2, it was only possible to enter evidence manually, one sample at a time. This meant that obtaining predictions for several samples was a long process, which limited the usefulness of RPBBN as an operational tool. The batch-mode processing tool aimed to alleviate this problem.

Batch mode is designed to process multiple samples, update evidence for a user-defined set of variables and report the results. The process involves the following series of dialogues designed to collect the information necessary to run the batch process.

Stage 1 - Selecting a data source. Batch mode is currently designed only to operate external files, not its own database. The first stage involves opening an input file.

Stage 2 - Selecting the input variables. The next stage involves selecting fields in the file that will be used as evidence (see Figure 12.3). Variables can be added and removed individually or as predefined groups of variables, 'Taxon' and 'Environmental' variables can be added using the add group button.

Stage 3 - Selecting an output file. This involves selecting an output file name to which the results of the batch process will be saved.

Stage 4 - Batch processing. The final stage is actual batch processing of samples from the selected data source. The batch progress dialogue box shows the progress of the batch job, giving an estimate of the time remaining and the current record being input. The dialogue box, Figure 12.4, also allows the user to cancel the batch job. Cancelling causes the application to quit the current job after completing the current sample. The results for each sample are saved as soon as it has been processed, so even if 'Cancel' is selected, the current set of results is retained.

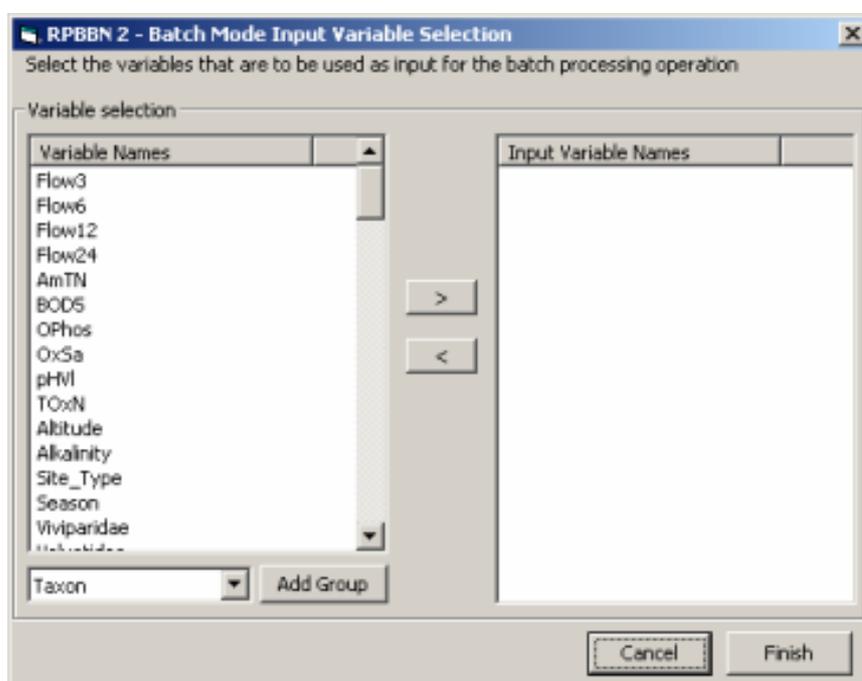


Figure 12.3 Batch mode 'Input Variable Selection' dialogue box.

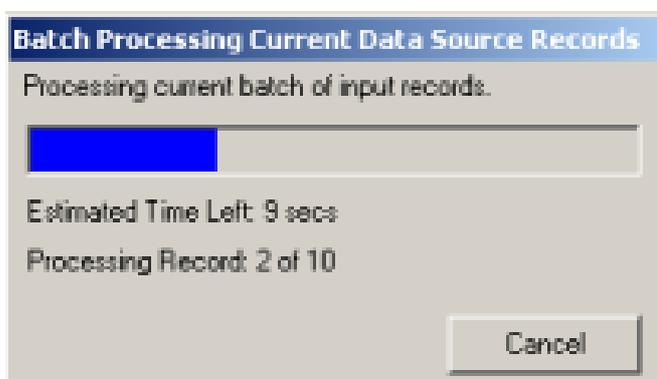


Figure 12.4 Batch progress dialogue box.

To reduce input file compatibility problems, RPBBN 2.1 accepts comma-delimited files, which can be generated easily by many databases and other applications such as Excel. In addition, the file loading mechanism uses header names to identify the relevant input information. Although this makes RPBBN 2.1 quite prescriptive in terms

of headings that have to be used in the input file, it does allow columns to be included in any order and additional unused columns to be included alongside the input data.

For future versions of RPBBN, there are plans to use input specification files to help streamline the file loading process. This would include information such as a look-up table containing the input file field names and the corresponding RPBBN variable name. This should reduce and in many cases remove the need to modify files generated by other applications before being loaded into RPBBN.

Items not undertaken in the first extension project

The first extension project was delayed because of problems with availability of data and as a result began in February instead of January 2009. Three weeks or 30 man-days had been lost so there was insufficient time to complete all work tasks and produce an updated version of RPBBN before the end of financial year deadline. After consultation, it was decided that some work items would have to be dropped so that an updated version of RPBBN could be created. However, the time lost would be made up after the project deadline, to carry out important software post-delivery tasks such as debugging, minor enhancements suggested by users and documentation.

The work items dropped had to be self-contained and not essential. Desirable upgrades that were not crucial to improving the RPBBN's usability were considered non-essential. For example, the reporting and batch mode work items were not considered for removal because they were deemed fundamental to making RPBBN 2.1 a more useful and usable tool.

In the end, the following work items were dropped:

- Include confidence of predictions based on the number of samples used to generate the probabilities.
- Allow the user to input a predicted reference value from RIVPACS so that RPBBN can produce an EQR.
- Add WFD classification so that the effect of physical and chemical changes on WFD status can be observed. In addition to requiring a field for entering predicted reference values (work item above) for the classification metrics ASPT, N-taxa and WHPT it requires a facility to specify the classification metrics (swap from ASPT to WHPT) and to change the class boundaries because the boundaries for WHPT, which is to replace ASPT from the second River Basin Management Plan in 2012 have not been set yet – maybe via a configuration page.

The first item was dropped because, although it would be informative, it was not considered essential and it required 20 man-days to complete and so was the largest work item. The second and third were dropped because they were closely related and the third required the second to function. Although the automatic calculation of EQRs and class would be extremely useful, users can calculate them outside the programme with little difficulty.

Removal of these work items reduced development time by a total of 45 man-days. This freed an extra 15 man-days that was used to make enhancements to the functionality of RPBBN 2.1 software and to implement user-requested features as described in the next section.

Additional work

Both extension projects included 'additional' unforeseen work that had to be completed to produce the final systems. The first project, however, involved much more work due to the need to fill the 45 man-days freed after the revision of work following the late start. This section covers the following six additional tasks undertaken in both extension projects:

- group templates;
- a report bar;
- manual propagation;
- documentation and help file;
- a domain manager;
- bug removal and feature enhancement.

The first four tasks were completed during the first project and the fifth task, a domain manager, during the second extension. Bug removal and feature enhancement was performed in both projects and all this work is presented together.

Group templates

One of the key issues with the RPBBN v.1 software is organising the large amounts of information. The RPBBN models have about a hundred nodes with, collectively, around five hundred states. The user interacts with the system in the following three ways: 1) displaying the states of the nodes, 2) entering evidence data and 3) reporting the predictions.

In RPBBN 1.2, these interactions were simplified by dividing the nodes into two groups, 'Environmental' and 'Biological', and introducing features that operated on these groups. The key benefits of this approach were that it helped to organise the information and enabled operations to be performed on several nodes simultaneously.

This approach was expanded in RPBBN version 2.1, to allow different 'group templates' to be used to organise the nodes. The templates themselves basically consist of a series of groups, with a simple or hierarchical structure. These groups are little more than named containers in which certain types of nodes are placed. The original concept was to allow users to construct their own templates, which would allow them to impose their own organisation on the nodes and then operate RPBBN 2.1 using these groups to speed up interaction with the model. The ability to organise the data would be particularly useful when displaying information on the screen or in a report because the user could control which nodes were displayed together and the type of information provided, that is, a full report, a summary or nothing at all. The group interactions would be useful in many different ways, but easing the burden of entering evidence would, perhaps, be one of the most beneficial. This approach also provided the opportunity to define task-orientated templates. These would allow users to set up predefined display configuration, list of evidence nodes and report configuration to tackle one task, such as prediction of organic pollutants. Then when the user needed to run one of these tasks, they would simply load the templates and the system would instantly be configured correctly.

Although the benefits of including user-defined templates in the RPBBN 2.1 were immediately apparent, upon trying to implement some of the necessary features it became clear that they would be too difficult to implement. The main problem was that

it would have a fundamental affect on the way RPBBN operates, which would require large-scale changes and managing the templates would require an additional layer of functionality, making it more complex to use. Figure 12.5 shows a screenshot of one of the new screens that would have to be added to RPBBN, to handle the templates.

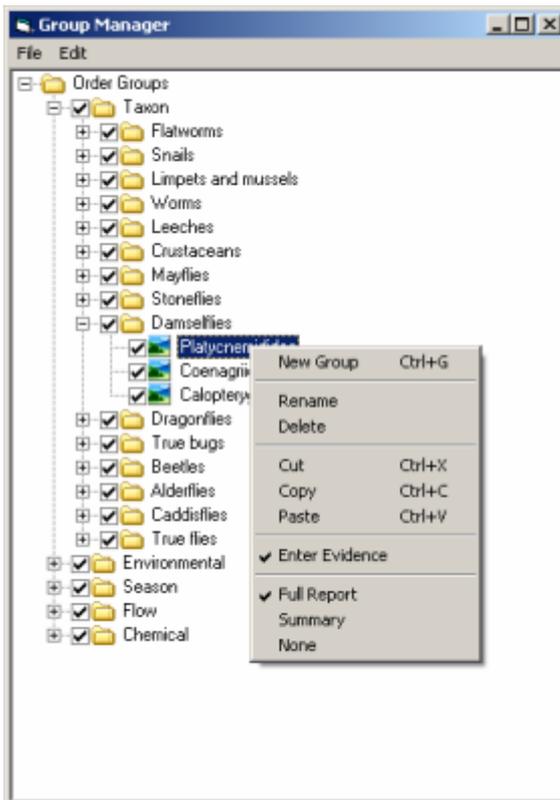


Figure 12.5 Screenshot of prototype ‘Group Manager’ application, designed to allow users to construct and modify templates.

While it would be impractical to fully implement group templates in RPBBN 2.1, it was possible to include a much more limited version. The main change to the templates was the removal of persistent state information, that is, state information saved with the template. The type of ‘state information’ that would be stored related to how variables would be displayed, reported and whether data associated with them would be used during evidence entry. The decision to remove persistent state information was taken because the storage and editing of states within the templates was the source of most of the complexity. The result of this was to reduce the role of the template to a definition of the arrangement of variables in groups.

The facility to allow users to create and edit their own templates was also abandoned. There was insufficient time to implement such a potentially complex feature. Without it, the types of templates that could be used were limited to pre-defined templates supplied with the software. These are described in Table 12.3. However, the facility for using any template for the display, reporting and evidence entry purposes was fully implemented in RPBBN 2.1. This makes the inclusion of new pre-defined templates relatively straightforward.

Table 12.3 Different types of groups in templates.

Template Name	No. of Groups	RPBBN feature in which they appear	Group Names
No Groups	None	Display – Charts arrangement	
Basic Groups	Two	Display – Charts arrangement Report – Report organisation Batch Mode – Group evidence entry	Taxonomical Environmental
Ordered Groups	Five main groups Fifteen subgroups	Display – Charts arrangement	Taxonomical Dragonflies Leeches Mayflies Crustaceans True flies Beetles Damsel flies Alderflies True bugs Worms Caddis-flies Limpets & Mussels Stoneflies Flatworms Snails Environmental Season Flow Chemical

The difficulty of writing software to introduce the added complication of templates in a user friendly way, in the limited time available, meant that their role in RPBBN 2.1 is limited. Their main application is to help manage the display, where all three templates are available to help organise the arrangement of bar charts. Only the ‘basic groups’ template is used to organise information in the report and can be used to input evidence for multiple nodes in one action.

Despite the limited way in which templates are implemented in RPBBN 2.1, this method of handling information within RPBBN has great potential and will become a key part of future releases of the software.

Report bar

The report bar was suggested in feedback given after a viewing of an early developmental version of RPBBN 2.1. During the demonstration, users commented that it was inconvenient to have to constantly flick between the chart and report panels to see how changes in the model affected predictions of biological indices. To resolve this problem, a report bar was added to the bottom of the charts panel display. This bar provides instant updates on the prediction of biological indices, as evidence in the RPBBN model is changed, see Figure 12.6.

The report bar itself was relatively easy to include in the interface. The majority of time was spent rewriting and optimising the code that calculates the biological indices and

taxon count, as the previous version, designed solely for updating the report panel, was too slow.

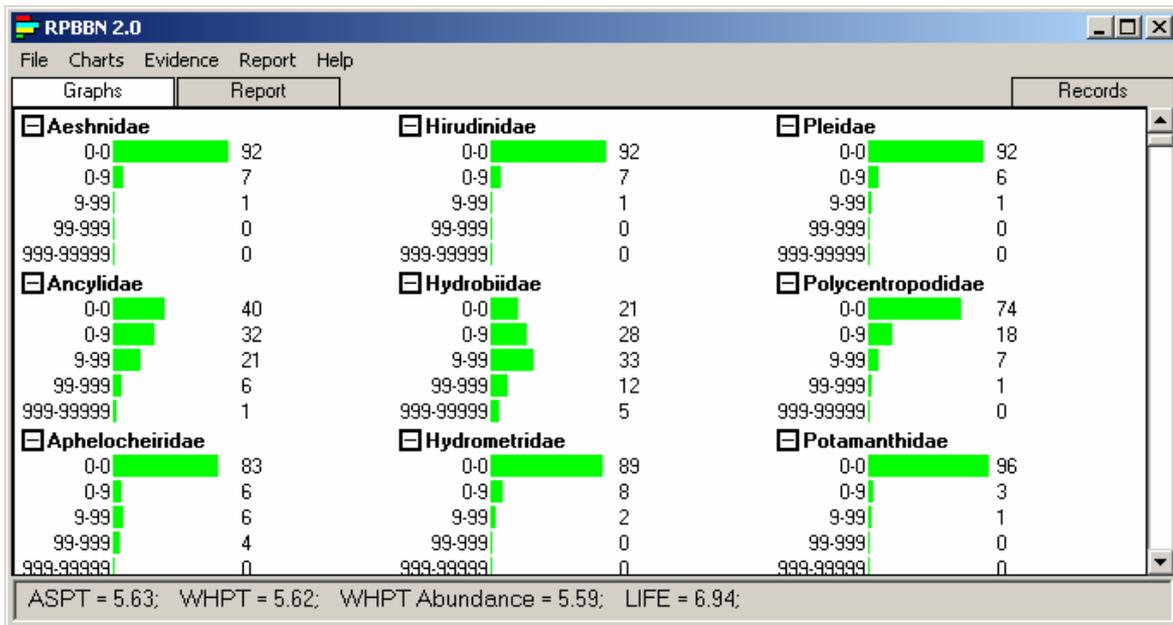


Figure 12.6 Chart panel and report bar.

Manual propagation

Propagation is the act of updating the probability of states of variables within a Bayesian Belief Network (BBN), usually as a result of evidence being entered or withdrawn. The term propagation is used because the process of updating the network starts with one variable and then propagates or ripples out to the other variables.

The most common approaches to initiating propagation available in BBN software are:

1. Automatic propagation – propagation is initiated every time evidence is updated.
2. Manual propagation – propagation is initiated only upon a user request.

Automatic propagation is usually the most convenient and 'responsive' option, with probabilities being instantly updated after any change. This tends to make comparisons and investigations easier to perform as evidence can quickly be modified and the results of these changes are available instantly. However, updating the probabilities in a BBN is not a trivial operation and potentially involves millions of calculations. Therefore, the size of the BBN model and specification of the PC on which the software is installed affects responsiveness of the automatic propagation.

In RPBBN 2.1 (beta test version) released before the end of the first extension, the automatic propagation was hard-coded. Only automatic propagation was used because the RPBBN-P1 model was sufficiently small to make updates almost instant, which negated the need to offer any alternative. The models released with RPBBN 2.1 (beta) were much larger. The RPBBN-P2 model is ten times the size of its predecessor and the RPBBN-S model is twenty-seven times larger. However, whilst the increased size of RPBBN-P2 seemed only to introduce a delay of a few tenths of a second to the duration of propagation, the much larger RPBBN-S model took approximately two and a half minutes to update on the development PC. Although two and a half minutes is not a prohibitively long time to update the model, automatic propagation means that updates occur every time a variable is changed, which makes using this RPBBN laborious.

There were two solutions to reducing propagation delays with RPBBN-S. The first was to optimize the BBN inference engine code, thereby reducing the duration of propagation operations. The second was to introduce a manual propagation option, which would enable changes to be made without propagation taking place until the user requested it. Given that the first option was likely to take more than 20 days to complete, we decided that the second option was better for this project.

'Manual Propagation' appears as an option under the 'Evidence' menu. Once selected, evidence can be modified but no update will occur until the 'Propagate' option (also found in the 'Evidence menu') is selected. When evidence has been modified and an update is required, the bars of the charts change colour to grey, to indicate the values they display no longer correspond to the evidence that has been entered, see Figure 12.7. Once the propagation option has been selected and the update performed, the bars for the predicted probabilities return to green to show that they now reflect the effect of the current evidence.

Introducing the manual propagation option took about five man-days to complete, leaving 25 man-days for remaining work.

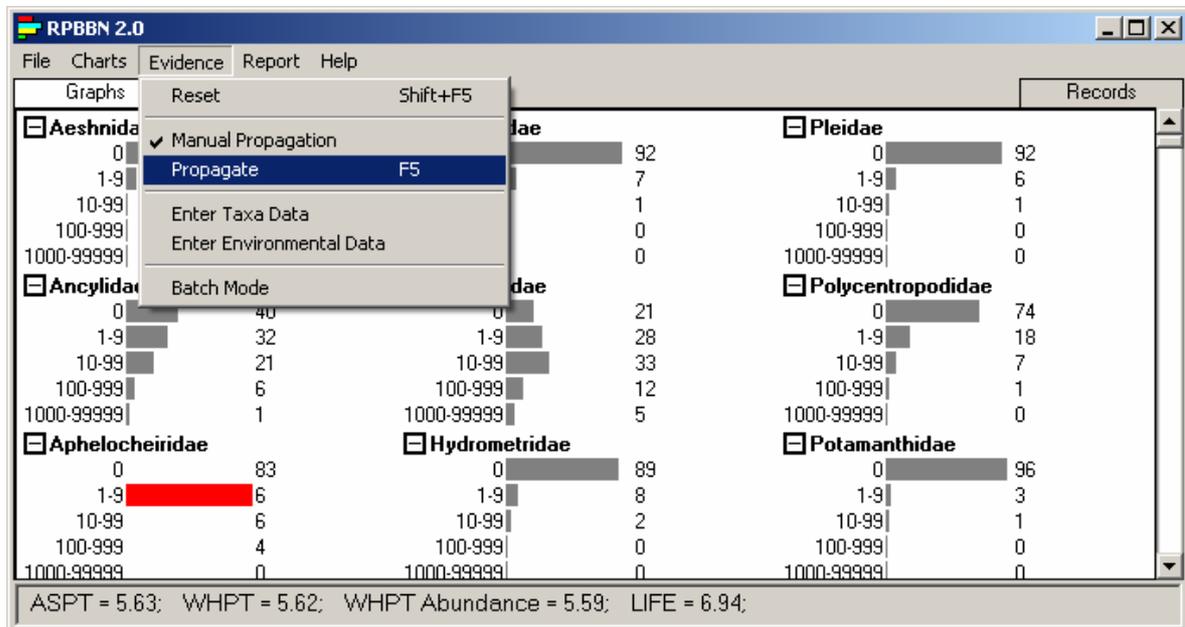


Figure 12.7 Screenshot of RPBBN with 'Manual Propagation' option selected, prior to update of model after entry of new evidence. Red bar denotes new evidence and grey bars that an update is required.

Documentation

RPBBN v.1 was a prototype and was a relatively simple piece of software with few features. As a result, the documentation was also quite simple, consisting of an eight-page manual. In updating the RPBBN software, both the interface and how the system operates underwent major changes, making it much more sophisticated. The improvements have made the RPBBN 2.1 system a much more useful and usable tool, making it attractive to a wider audience. This made good documentation in the form of a help file essential, to provide new users with adequate support when using the software for the first time.

The original project specification focused on updating existing features and introducing new ones, and documentation was limited. The time freed by the work items that were dropped provided an opportunity to resolve this by producing new help files.

Creating the help file was a laborious task that took approximately 20 man-days. The help file contains over 30 pages of information split into three sections:

Getting Started provides a general introduction to the RPBBN 2.0 interface covering its main features - charts, reports and records.

Using RPBBN describes how to use RPBBN to make predictions.

RPBBN 2.0 systematically describes each of the options in RPBBN's menus.

The help file was produced after the original deadline and is packaged with the final release version of RPBBN 2.1. This help file supersedes that provided with the beta version of RPBBN 2.1 released within the original extension, which only gave a terse description of the software and its features.

Domain manager

The first version of the RPBBN software was designed to work with the RPBBN-P1 model. Creation of the RPBBN-S model, in the first extension, broke this one-to-one relationship. However, because it was intended for a specialized audience, this problem was resolved by simply packaging both networks and a database for each in the installation software. This solution would be less practical if several models were to be packaged and released more widely. The main reason for this is the size of the RPBBN database; it occupies approximately one gigabyte, and having several different versions of the database included in the software would make both the installation package and the installation itself impractically large.

Fortunately, only a small amount of information ties an RPBBN database to an RPBBN model, with the core of the data remaining the same. Therefore, it was possible to revise the database structure to allow a set of models to function with the same database. However, this set of models would be limited to those that consist of a complete set or a subset of the nodes in the RPBBN-P2 model. These limitations are the result of specialization of RPBBN software and the information it requires to perform its task. For example, RPBBN is able to predict biological assessment values such as BMWP, but to do this the BBN model needs to contain a recognizable set of BMWP taxon nodes. Changes to the database consisted of introducing a new table called 'Domain', which contains information about the name, location and version of associated RPBBN networks, and inclusion of a DomainID field to existing Node and NodeState tables, so that data on these attributes for different Domains could be stored.

The final requirement was to introduce a way for the user to manage various versions of the RPBBN model. This took the form of the Domain Manager dialogue box, in Figure 12.8.

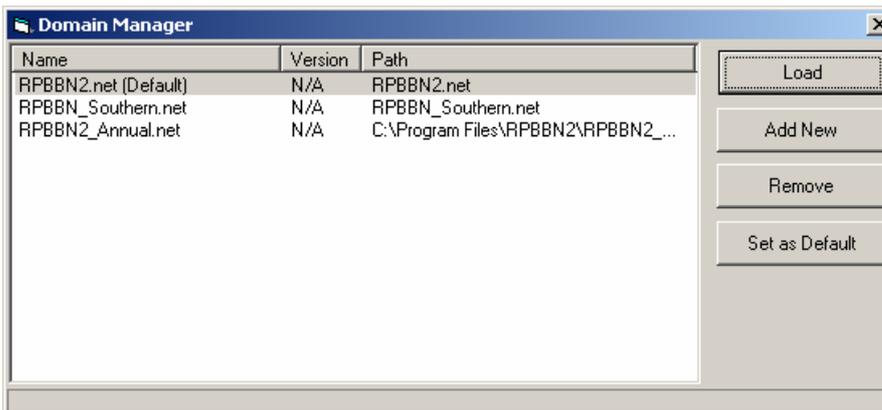


Figure 12.8 Domain manager dialogue box.

The domain manager provides a facility for changing the network being used by the RPBBN 2.2 software, an option to add or remove existing models and the option to change the network that loaded by default when starting RPBBN 2.2. The domain manager can be accessed during initialization by clicking the link on the splash screen or afterward by selecting the 'Domain Manager ...' option from the 'File' menu.

Introduction of the domain manager provided the following benefits:

- Limited growth of the installation size, by allowing several versions of the RPBBN model to be packaged with just one database.
- Ability to upgrade the model without changing the database or upgrading software.
- Ability to change models without having to reload the software.

Bugs and improvements

First extension project

During the course of the first extension project, two 'beta' versions of the RPBBN 2.1 software were produced and stand-alone installation CDs were dispatched to the Environment Agency. Following a period of development, a final release version of RPBBN 2.1 was produced. This version contained many minor improvements, particularly to batch mode, bug fixes and improved error catching.

The beta versions provided users with an opportunity to test RPBBN 2.1 in an operational setting. This 'day to day' usage quickly revealed some non-fatal errors that were difficult to detect because they did not crash the program, and some enhancements that were needed to improve how RPBBN 2.1 operated and thus increase productivity. The remaining five man-days development time were spent removing these errors and improving features.

The majority of these enhancements were made to the batch mode. The following is a list of improvements made to batch mode.

1. Improvement to file loading routines. In the version released before the original deadline, it was necessary to include headings in the CSV file for all the variables in the RPBBN model. The updated version removed this requirement, so the file only had to contain columns for which there were data.
2. The variables listed in the 'Variable Name' list box in the input selection dialogue box (see Figure 12.3), were restricted to those for which there

were data. This change was made in concert with improvement to the load routines. Removal of superfluous columns from the input files meant that it was possible to list only the variables for which data was available.

3. Improvements to the batch mode dialogue box and process. In the beta versions of RPBBN 2.1, little information was provided about the duration of the process and there was no way of cancelling it. These issues were resolved in the final version. A cancel button was added to the dialogue box to stop the process and more detailed information was provided, including the current progress through the records and an estimation of the remaining processing time (Figure 12.4).

Minor bugs were discovered in the handling of input files, printing and formatting of the on-screen report. However, none of the bugs were particularly problematic and they could be addressed relatively quickly.

Second extension project

No modifications were requested for the RPBBN software in the second extension project. Therefore, apart from the introduction of the domain manager feature, little debugging was made to the 2.1 version of RPBBN released at the end of the first extension. The only notable change was reformatting of information in the report bar to explicitly indicate the average score per taxa (ASPT) and number of taxa (NTAXa).

Changes made to the RPBBN system in the second extension focused on changes to the underlying models. As a result most work on 'bugs' and improvement related to these models.

Revision of the 'site type' sub-network

Probabilities for the 'site type' variable in the RPBBN-P2 model were derived from the environmental data and predictions of site type in the project database. However, predictions of site type associated with samples in the project database had been generated by a specially created 'site type predictor' BBN, which essentially has the same structure as the 'site type' sub-network in the RPBBN model, Figure 12.9.

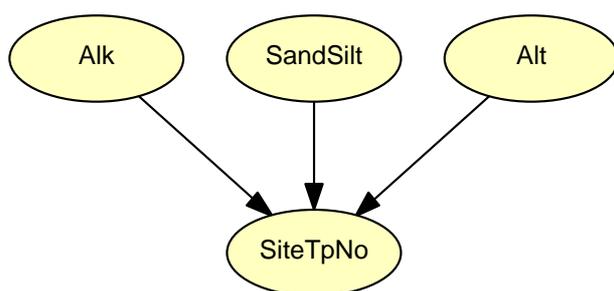


Figure 12.9 Causal network of the 'site type predictor' BBN.

Probabilities for this 'site type predictor' BBN were based on the 1995 River Survey data and original 'site type' values generated using an artificial neural network model (see Walley *et al.* 1998).

Differences in the source data, and therefore probabilities in the RPBBN and 'site type predictor' BBN, meant that the predictions of site type made by these two models differed. This occasionally manifested itself in the RPBBN by the 'site type' with the highest probability not being the one associated with a training sample. To prevent

these 'anomalous' results, the site type sub-network in the RPBBN was simply replaced with the 'site type predictor' BBN. Whilst this meant the probabilities in this part of the RPBBN model were now based on the smaller 1995 river survey dataset, 'site type' values in this set would be the product of the original neural network model, not 'secondhand' predictions generated by a BBN model.

Other minor network modifications

Two other notable modifications were made to the RPBBN models released with the beta versions of RPBBN 2.2. The first set of modifications concerned the correction of an error made when removing the three- and six-month flow condition nodes. Although these flow nodes had been removed from the RPBBN-P2 and RPBBN-S models, the 26 taxa attached to these nodes had not been reassigned a new flow parent. It was therefore necessary to reattach a flow parent to these nodes and produce updated versions of the models.

The second set of modifications was made to RPBBN-S. The updated version of the RPBBN-S model was created from scratch at the same time as the RPBBN-P2 model and the same processes were used to update it. Older models contained direct links between BOD5 and the taxa, unlike the new RPBBN-S and RPBBN-P2 models, which have 'indirect' links through total ammoniacal nitrogen and oxygen saturation. Whilst these links are likely to help the prediction of BOD5, the legitimacy of modelling a direct causal link between BOD5 and the biology is questionable. Following consultation we decided that indirect links between BOD5 and biology modelled in the 'modified' structure of the new RPBBN-S was more appropriate.

The issues discussed here only affected networks released with beta version of RPBBN 2.2, not those packaged with the final release version.

Rewrite updating procedure

RPBBN-S exposed the RPBBN software's limited ability to handle large BBN models. Manual propagation alleviated some of the problems by allowing the user to control when an update takes place. However, it has not resolved the underlying problems associated with the inference engine algorithms.

RPBBN uses its own BBN inference engine software, called dBBN. This software consists of a single dynamic link library (dll) file, the task of which is to load BBN models and perform the necessary inference/calculations to make the model function. This software was based on the HUGIN architecture (Jensen *et al.*, 1990) and used some of the algorithms suggested by Huang and Darwiche (1996) to implement the inference engine. This architecture was chosen because it was known to be quick in terms of inference speed. This speed is achieved because the algorithm performs the same set of calculations each time, making it possible to store and re-use intermediate results. The trade-off for this speed is the need for a larger working memory, as large arrays of values have to be maintained to reduce the amount of processing.

An alternative architecture for the inference engine is based on 'lazy propagation' (Madsen and Jensen 1999). The big difference between this approach and HUGIN is in the handling of the updating/propagation calculations. At each stage, 'lazy' algorithms check that all the combination calculations that are supposed to be performed are actually necessary, that is, that they cause some change. If they are unnecessary, they are avoided. This negates much of the benefit of storing intermediate results, as done by HUGIN, because there is no guarantee that these calculations will be performed again. As a result, the probability distributions in lazy algorithms are maintained in a factorized/uncombined form.

The benefit of using a 'lazy' algorithm is that it can potentially retain much of the speed of HUGIN whilst dramatically reducing memory usage. Retaining the probability distributions in factorized form and avoiding the large arrays of intermediate values required by HUGIN algorithms helps to reduce excessive memory consumption, and improves speed by reducing the computation overheads that excessive memory consumption incurs. Avoiding 'unnecessary' calculations also improves speed by reducing the overall number of calculations. The amount of 'unnecessary' calculation varies, depending on the structure of the network and the evidence that it contains. 'Lazy' propagation algorithms are naturally more processor-intensive than HUGIN, because they undertake an additional calculation-checking stage. Therefore, in the case of a network that requires limited working memory and with no 'unnecessary' calculations occurring during propagation, a 'lazy' algorithm will update more slowly than a HUGIN algorithm. A more detailed comparison of the performance of HUGIN and 'lazy' propagation algorithms is given in Madsen and Jensen (1999).

EABBN

In dBBN, the HUGIN architecture is heavily integrated with all the existing code. Ideally, the revised version of the code would include both the HUGIN and 'lazy' propagation methods, with the implementation code for these methods being separated from other components. This would involve a major change in the architecture of the existing software and an extensive rewrite. Fortunately, many elements of the two propagation methods are similar and it became apparent that a simpler solution would be to replace the parts of HUGIN algorithm that differed from their 'lazy' counterparts. From a design perspective, this solution might be the most effective but it does not constitute a major change or development of dBBN. Therefore it was decided to 'fork' development with the proposed 'lazy' propagation software being called EABBN and the name dBBN retained for further in-house development versions.

Performance

As mentioned previously, Madsen and Jensen (1999) provide a detailed comparison of the HUGIN and lazy propagation algorithms for a variety of networks. This section focuses on the changes in performance in the RPBBN model.

Implementation of the HUGIN algorithm in dBBN makes use of what Huang and Darwiche (1996) refer to as Cluster-Sepset mappings. These are arrays of mappings or pointers between probability distribution tables, which allow the corresponding values in both tables to be combined more quickly, see Figure 12.10. Cluster-Sepset mappings provide a notable improvement in performance but at the expense of a large increase in memory consumption, required for the mapping arrays. This technique was not, and could not, be used with the lazy propagation algorithm because of variations in the combinations of probability tables required with each update. Absence of these additional memory demands and storage of probability distributions in a factorized form meant that the lazy propagation algorithm was capable of loading and updating RPBBN-P2 using only half the memory of the dBBN's HUGIN algorithm.

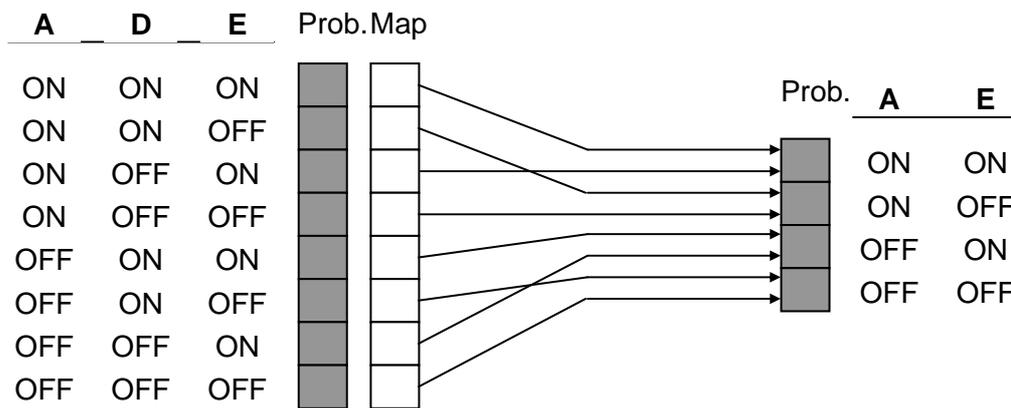


Figure 12.10 Example of mapping between two probability tables. Adapted from Huang and Darwiche (1996).

In terms of the speed, comparison of the HUGIN and lazy algorithms is a little more complicated. This is because of variation in the work required by the lazy propagation to perform an update, given the evidence entered into the model. One clear area of improvement was the initialization, which is the process of setting up tables of probabilities and beliefs and other data structures used by the algorithm. The absence of combined probability distributions and ‘mapping’ tables used in the HUGIN algorithm meant that, on average, the lazy propagation algorithm was able to decrease initialization time by at least 75 per cent (see Table 12.4).

Table 12.4 Comparison of times taken by HUGIN and lazy algorithms to initialize and perform update on RPBBN-P2, RPBBN-A and RPBBN-S models.

Network	HUGIN (seconds)	Lazy (seconds)
RPBBN-P2		
Initialize	1.297	0.344
Update	0.109	0.110
RPBBN-A		
Initialize	4.266	0.641
Update	0.406	0.281
RPBBN-S		
Initialize	244.031	1.234
Update	166.047	0.703

Table 12.4 shows a typical set of results from the HUGIN and lazy algorithms. It shows that, for the RPBBN-P2 model, the speed of update performed by the lazy algorithm is comparable with that of the HUGIN algorithm. However, given the general speed of the updates, it is unlikely that the average user would be able to notice any difference in performance between the two versions of the algorithm when used in RPBBN unless memory limitations become important. If this happened, the greater memory demands of the HUGIN algorithm would cause its performance to degrade much more rapidly than that of the lazy algorithm.

This degradation in performance when memory is limited is illustrated in a final test using the RPBBN-S model. During this test the memory consumed by the algorithm peaked at 1,050 MB (over one gigabyte), whilst the memory consumed by the lazy algorithm peaked at 200 MB. The reduced memory demands of the lazy algorithm meant that on our test PC it was able to initialize and update the model using only working memory. Table 12.4 shows that in a straightforward speed test, the lazy algorithm initialized and updated in approximately 0.5 per cent of the time of the HUGIN algorithm. Given the similarity in performance recorded during the RPBBN-P2

model tests, the dramatic difference recorded in the RPBBN-S test indicates that memory limitations can have a severe impact on the updating speed of the models.

User feedback workshop

The user feedback workshop was described in Section 10, where feedback on the new version of RPDS was summarised. The feedback on RPBBN is reported here and consisted of the following comments for improvement:

Modifications to RPBBN model

- It would help if there was a text description not just a file name in the domain manager. For example:
Two Seasons, Five Environmental Categories
Single Season (default)
Single Season, Multiple Environmental Categories.

Changes to the user interface

- If you save scores it would be handy to save indices too.
- Need to state that indices in brackets are the numbers of taxa specific to the index, and not always N-Taxa (BMWP).
- The scroll bar is not very obvious for the main window.
- Good guidance is needed on using the system and interpreting the data.
- Good guidance on how to use and more importantly how to interpret the results.

Batch mode

- Consider the option of multiplying predictions by the regression of observed versus predicted values as a short-term solution to the under-prediction of extremes.

13 Summary and conclusions

Background to the project

This project has built on the successful outcomes of previous projects, in particular Project E1-056 (Walley *et al.*, 2002). In that project, two software systems were developed for the diagnosis and prediction of river health from biological and environmental data, namely the River Pollution Diagnostic System (RPDS) and the River Pollution Bayesian Belief Network (RPBBN). Although two early objectives of this project involved using the RPDS database of matched biological and environmental data to determine chemical thresholds and potential reference sites, the main aims of the project were to enhance the software systems. The specific aims were as follows:

- Substantially extend the dataset on which the data models are based.
- Revise and test the data models on which the systems are based.
- Extend the functionality of the two systems and combine into one 'integrated system'.

The new versions of RPDS and RPBBN are known as the River Pressure Diagnostic System and the River Pressure Bayesian Belief Network to reflect the fact that they respond to a wider range of perturbations than pollution.

Summary of project outcomes

Extending the dataset (the first of the above aims) took much longer than anticipated and affected progress of the remainder of the project. In particular, the integrated system could not be developed in the remaining time available (the third aim), and the two software systems were kept separate. The outcomes of the project are summarised as follows:

- The dataset was extended substantially.
- A method was developed for estimating flow condition at the time of sampling and this was shown to have the anticipated relationship to taxa.
- MIR-max models were revised, based on the larger dataset.
- RPDS software was revised, based on the new MIR-max models and incorporating new functionality.
- The original BBN model was revised, based on the larger dataset, and further models developed.
- RPBBN software was revised, based on the new BBN models, and incorporating new functionality.

More details on each of these are provided below.

Extension of the dataset

The dataset has been extended temporally, geographically and in terms of the variables contained. Firstly, the new dataset covers the ten-year period 1995-2004 instead of the single year 1995. Secondly, it covers Scotland as well as England and

Wales. Thirdly, flow (except for Scotland), geology, land cover and land risk variables have been added to chemistry and stress as diagnostic variables.

Validation of the biological sample data was time-consuming, as was the validation of spatial coordinates of both biological and chemical samples prior to matching. However, the resulting spring and autumn datasets contain many more biological samples than those used in earlier systems.

Estimation of flow condition and impact on taxa

To reflect increasing interest in linking changes in the biological community with the physical flow of water in the river, two measures of flow were included in the new diagnostic variables. The first was the percentage impact (at 95 per cent exceedence probability) from LowFlows2000 (LowFlows Enterprise), which quantifies the extent to which the natural flow may have decreased or increased because of external influences such as abstractions and discharges. Although this may be useful, the values are based on long-term average data and may have little bearing on the condition of the river at a particular time.

To complement these values, a second measure was developed to estimate flow condition at each site at the time the sample was taken. This was based on interpolation from thirty years' monthly flow data at 121 gauged sites covering England and Wales. Values on a scale of zero (driest) to one (wettest) were determined for time periods of one, two, three, six, 12 and 24 months prior to the sample date.

The effects on the taxa were demonstrated by splitting the flow condition scale into the ranges $[0, 1/3]$, $[1/3, 2/3]$ and $[2/3, 1]$, representing 'dry', 'average' and 'wet' respectively, and deriving abundance distributions of each taxon for each flow condition from their frequencies of occurrence. This was done separately for spring and autumn, and for riffles and pools. Grouping the taxa based on the change in presence/absence probabilities from 'wet' to 'dry' conditions showed a clear pattern. Those taxa more likely to occur in wet conditions were the more sensitive taxa, as indicated by their higher revised BMWP scores. Those more likely to occur in dry conditions were the more tolerant taxa, as indicated by their lower revised BMWP scores. Furthermore, changes in the presence/absence probabilities became more marked as the time period relating to the flow condition lengthened.

Revision of MIR-max models

Preliminary MIR-max models were created with a reduced dataset consisting of samples from 1995 only, which reproduced the original MIR-max models to a high degree. Bandings used for the biological variables were kept the same as in the original model (discrete abundance categories); however, bandings used for continuous environmental variables were modified. Bands in the original model were obtained by splitting the range into equal widths, and these were replaced by more appropriate bands based on equal numbers of samples in each (as recommended in the previous project, Walley *et al.* 2002). However, this gave more prominence in the clustering to environmental variables, which tended to dominate the biological variables when ranked in terms of contribution to the overall mutual information. Following a series of tests, the set of environmental variables was reduced from eleven to five, to reach a reasonable compromise between opposing objectives of optimising the representation of environmental characteristics of a site and optimising the influence of macroinvertebrates in the final clustering. The five environmental variables adopted were chosen to cover the widest range of influences on habitat and were: alkalinity (indicative of chemical conditions), slope (flow velocity), boulders and pebbles

(substrate composition), distance from source (river size) and altitude (temperature). Further tests with calcareous geology in the place of alkalinity indicated that, although alkalinity was the preferred variable, calcareous geology would be a viable alternative.

Following testing with the reduced dataset, models were produced for the full dataset covering the years 1995-2004 for both spring and autumn. In all cases, the number of bins was kept the same as in the original models, namely 250. However, the new models were based on more than five times as much data. The original spring and autumn models contained over 6,000 samples each (an average of 24 samples in each bin), whereas the new spring model contains over 32,100 samples and the new autumn model roughly 31,500 (averages of 129 and 126 samples per bin respectively). Sample data in the original spring and autumn models was from 6,000 sites in England and Wales, whereas sample data in the new spring and autumn models was from 9,100 and 8,800 sites respectively in England, Wales and Scotland.

The new spring and autumn cluster models were ordered by MIR-max to produce hexagonal output maps of side-length 10, 15 and 20 clusters, to match the original models. Each map was rotated to align as closely as possible with the originals to permit easy comparison.

Revision of RPDS software

After appending the new diagnostic variables to the spring and autumn models, RPDS 2.0 was revised to RPDS 3.0 by streamlining operations involving database queries and including Scotland on the geographical map panel. The new spring and autumn models of RPDS 3.0 showed qualitative similarities with those of RPDS 2.0. The output maps of RPDS 3.0 for particular variables are similar to those of RPDS 2.0, demonstrating that the clustering and ordering of new models are similar to the originals despite the large increase in volume of data. The geographic locations of samples in clusters occupying similar positions in the hexagonal output map are also consistent between RPDS 2.0 and RPDS 3.0.

Preliminary evaluation of new flow variables demonstrates that the percentage impact at Q95 is negatively correlated with distance from source, as might be expected, so that the influenced flow is usually less than natural flow close to source and greater than natural flow close to the mouth. Preliminary evaluation of the flow condition variable, on the other hand, demonstrates a relationship with the taxa in terms of ASPT. The clusters containing samples taken in wetter years tend to be those with higher ASPT, while those taken in drier years tend to be those with lower ASPT.

Following meetings with potential users, the RPDS software was modified to improve its usability and incorporate additional functionality to meet requirements of the Water Framework Directive.

Revision of BBN model

As with the MIR-max models, the BBN model was derived from a much larger quantity of data. The original BBN model was based on 3,600 spring and autumn matched samples, whereas the new spring and autumn models are based on roughly 16,200 and 15,900 matched samples respectively. In addition, further changes were made: the structure of the model itself was modified; the chemical statistics were based on percentile values over the previous three years rather than mean values over the previous three months; and five states were used for the taxonomic variables instead of four.

Dependent testing of the network with these changes against the original showed major improvements in the predictions of total ammoniacal nitrogen and dissolved oxygen. Prediction of flow condition variables was poor, but this was not unexpected given their fewer connections to the taxonomic variables. Independent testing to assess the impact of each of the changes suggested that the use of percentile statistics, permitted by the larger dataset, was the greatest factor contributing to the improvements.

Two further models have been developed, one based on two-season samples, and the other with many more states for chemical variables, based on WFD and RE standards.

Revision of RPBBN software

As with the RPDS software, the RPBBN software was also modified in several ways. These modifications make the software easier to use and incorporate new functionality to meet requirements of the Water Framework Directive.

Overall conclusions

In this project, major extensions were made to the datasets on which the RPDS and RPBBN software systems are based. Biological and chemical sampling data covering the decade 1995-2004 was included, and the geographical extent increased to include Scotland as well as England and Wales. In addition, new variables such as flow, simple geology, land cover and land risk were included in RPDS 3.0. These offer enhanced diagnostic capability as well as new links with the macroinvertebrates to be explored, for example with flow condition prior to the time of sampling, which was shown to have an ecologically significant impact. This project database is useful in its own right as a basis for other research.

The MIR-max models were revised, based on the increased dataset with changes in the representation of environmental variables. RPDS 3.0 software was produced based on the new MIR-max models, and preliminary evaluation indicates good qualitative similarities to the original for variables that they have in common, and the possibility of exploring relationships with the new variables. The BBN model was revised, based on the enhanced dataset and with further changes. Dependent tests showed notable improvements in the performance of two of the chemical variables in the network. Further independent tests suggested that the largest contributing factor was the use of chemical statistics based on percentile values, permitted by the enhanced dataset.

Both RPDS and RPBBN software were revised to optimise them for operational use.

References

- Comber S, and Georges K. (2007) *Tiered approach to the assessment of metal compliance in surface waters: Extension report: Nickel*. Science Report SC050054/SR1b. Bristol, Environment Agency.
- Davy-Bowker J., Clarke R., Furse M., Davies C., Corbin T., Murphy J. and Kneebone N. (2007) *RIVPACS Pressure Data Analysis, Final Report Project WFD46*. Edinburgh: SNIFFER. Available from <http://www.sniffer.org.uk> by searching under 'WFD46'.
- Davy-Bowker J., Clarke R., Corbin T., Vincent H., Pretty J., Hawczak A., Blackburn J., Murphy J. & Jones I. (2008) *River Invertebrate Classification Tool*. Final Report Project WFD72c. Edinburgh, SNIFFER. Available from <http://www.sniffer.org.uk> by searching under 'WFD72c'.
- De Zwart D., Pemberton E., Posthuma L., Veal A., Wells C. (2008). *Understanding ecological impacts in rivers in England and Wales and identifying their possible causes: Part 1, the Effect and Probable Cause (EPC) method*. Environment Agency Science Report - SC030189/SR5. Bristol, Environment Agency.
- Fuller R.M., Smith G.M., Sanderson J.M., Hill R.A., Thomson A.G., Cox R., Brown N.J., Clarke R.T., Rothery P. and Gerard F.F. (2002) *Countryside Survey 2000 Module 7. Land Cover Map 2000 - A Guide to the Classification System*. Extract from the Draft Final Report. CEH Monkswood.
- Holzkämper A., Surridge B., Paetzold A., Kumar V., Lerner D.N., Maltby L., Wainwright J., Anderson C.W. and Harris R. (2008) A consistent framework for knowledge integration to support integrated catchment management. *Proceedings of the Fourth International Congress on Environmental Modelling and Software 7-10 July 2008*.
- Huang C. and Darwiche A. (1996). Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning* **15**(3), 225-263.
- Jensen F.V., Lauritzen S.L. and Olesen K.G. (1990). Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly* **4**, 269-282.
- Kapo K.E., Burton G.A. and Pemberton E. (2008) *Understanding ecological impacts in rivers in England and Wales and identifying their possible causes: Part 2, the GIS-based Weights of Evidence/Weighted Logistic Regression method*. Environment Agency Science Report - SC030189/SR6. Bristol, Environment Agency.
- Kumar V., Holzkämper A., Surridge B., Rockett P.I., Niranjana M. and Lerner D.N. (2008) Bayesian challenges in integrated catchment modelling. *Proceedings of the Fourth International Congress on Environmental Modelling and Software 7-10 July 2008*.
- Madsen A.L. and Jensen F.V. (1999) Lazy propagation: A junction tree inference algorithm based on lazy evaluation. *Artificial Intelligence* **113**(1), 203-245.
- Maitland P.S. (1977) *A coded checklist of animals occurring in fresh water in the British Isles*. Edinburgh, Institute of Terrestrial Ecology.

Martin R.W. and Walley W.J. (2000) *Distribution of perceived stresses in English and Welsh rivers based on the 1995 survey*. R&D Technical Report E126. Bristol, Environment Agency.

Martin R.W., O'Connor M.A. and Walley W.J. (2005) *Collection and analysis of environmental stresses influencing biological general quality assessment in 2000*. Science Report E1-114/TR. Bristol, Environment Agency.

Martin R.W. and Paisley M.F. (2005) *Distribution of perceived stresses in English and Welsh rivers in 2000-2003: the quality of the dataset and a comparison with the 1995 dataset*. R&D Technical Report EMC(05)08\TR. Bristol, Environment Agency.

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

O'Connor M.A. (2004) *Development of a pattern recognition and data visualisation system for the diagnosis of river health from biological and environmental data*. PhD. Thesis, Staffordshire University.

O'Connor M.A., Paisley M.F., Trigg D.J., and Walley W.J. (2005) *Further development and testing of artificial intelligence systems for the classification and diagnosis of river quality based on biological and environmental data*. R&D Technical Report E1-056/TR2. Bristol, Environment Agency.

Paisley, M.F., Trigg D.J. and Walley W.J. (2007) *Revision and testing of BMWP scores. Final report for Scottish and Northern Ireland Forum for Environmental Research (SNIFFER) Project WFD72a*. Edinburgh, SNIFFER. Available from <http://www.sniffer.org.uk> by searching under 'WFD72A'.

REFCOND (2003) *Rivers and Lakes – Typology, Reference Conditions and Classification Systems. Common Implementation Strategy for the Water Framework Directive (2000/60/EC) Guidance Document No 10*. Luxembourg, Office for Official Publications of the European Communities.

Trigg D.J. (2004) *Development of an expert system using plausible reasoning for the diagnosis and prognosis of river quality*. PhD. Thesis, Staffordshire University.

Walley W. J. and O'Connor M. A., (2001) Unsupervised pattern recognition for the interpretation of ecological data. *Ecological Modelling* **146**, 219-230.

Walley W.J., O'Connor M.A., Trigg D.J. and Martin R.W. (2002) *Diagnosing and predicting river health from biological survey data using pattern recognition and plausible reasoning*. Environment Agency Technical Report E1-056/TR2. Water Research Centre.

Walley W.J., Fontama V.N. and Martin R.W. (1998) *Applications of artificial intelligence in river quality surveys*. R&D Technical Report E52. Publication No. TH-01/98-B-BASA. Bristol, Environment Agency.

Young A.R., Grew R. and Holmes M.G.R. (2003) Low Flows 2000: a national water resources assessment and decision support tool. *Water Science and Technology* **48**(10), 119-126.

Glossary of terms

AMNI	CIES code for ammoniacal nitrogen non-ionised, mg/l
AMTN	CIES code for total ammoniacal nitrogen, mg/l
ANC	Acid-neutralising capacity
ASPT	Average BMWP-score per taxon, a biotic index of organic pollution
B4W	Biology for Windows, predecessor of BIOSYS
BBN	Bayesian Belief Network.
BIOSYS	Biological Information System, the Environment Agency's biological database
BMWP	Biological Monitoring Working Party. BMWP-score is a numerical index of river invertebrate quality based on the sum of individual values for each family based on their sensitivity to organic pollution.
BOD	Biochemical oxygen demand. BOD5 is 5-day BOD.
CAMS	Catchment Abstraction Management Strategy
CEH	Centre for Hydrology and Ecology
CIES	Centre for Intelligent Environmental Systems (Faculty of Computing, Engineering and Technology, Staffordshire University).
CGI	Common Gateway Interface
CSF	Catchment sensitive farming
CSV	Coma separated value, a file format
dBBN	Dynamic link library (dll) file for RPBBN for its inference engine, based on HUGIN
DO	Dissolved oxygen, mg/l.
d/s	Downstream
EA	Environment Agency
EABBN	Environment Agency BBN – an alternative to dBBN, this inference engine is based on lazy propagation
EQI	Ecological quality index (observed/RIVPACS expected)
EQR	Ecological quality ratio (observed/WFD reference value)
GQA	General Quality Assessment
GIS	Geographical information system
Hex10	Main RPDS display with a hexagon of 10 locations for bins per side
ID	Identifier
LCM2000	Land Cover Map 2000
LIFE	Lotic-invertebrate Index for Flow Evaluation
MEM	Macro-Ecological Model
MI	Mutual Information.
MI-max	MI Maximisation – a clustering algorithm.
MINTA	Minimum of Number of Taxa and ASPT
MIR-max	MI and Regression maximisation
ms-PAF	Multiple substance potentially affected fraction of species
NBN	National Biodiversity Network
NTaxa	Number of BMWP-scoring taxa, N-taxa
NVZ	Nitrate vulnerable zone
OXDS	CIES code for dissolved oxygen, mg/l
OXSA	CIES code for oxygen percentage saturation.
PHOS	CIES code for phosphorus as phosphate, mg/l.
PHVL	CIES code for pH value
PISCES	Environment Agency's Pressure Information Supporting Classification Elements for the Water Framework Directive database.
Q95	Flow exceeded 95 per cent of the time
r	Correlation coefficient (Pearson)
RE	River Ecosystem chemical classification

REFCOND	Reference Condition – a WFD Common Implementation Strategy Working Group
RHS	River Habitat Survey
RICT	River Invertebrate Classification Tool, software that implements RIVPACS IV.
R.I.P.	River Intercalibration Project
RIVPACS	RIVER Prediction And Classification System (Moss <i>et al.</i> , 1987). RIVPACS III and RIVPACS IV are successive versions of RIVPACS.
R-max	Regression maximisation
RPBBN	River Pressure Bayesian Belief Network, formerly River Pollution Bayesian Belief Network (Software developed by CIES).
RPBBN1	RPBBN software version 1, implementing RPBBN model RPBBN-P1
RPBBN2	RPBBN software version 2, implementing RPBBN models RPBBN-A, RPBBN-P2 and RPBBN-S
RPBBN-A	RPBBN-Annual, RPBBN model based on combined seasonal (spring/autumn) biological index values used to derive WFD quality classifications
RPBBN-P1	RPBBN-Project 1, RPBBN model produced in previous project
RPBBN-P2	RPBBN - Project 2, RPBBN model developed in this project with five states for chemical variables
RPBBN-S	RPBBN – Southern, RPBBN model with more than five states for chemical variables based on WFD and RE standards
RPDS	River Pressure Diagnostic System, formerly River Pressure Diagnostic System (Software developed by CIES).
r_s	Rank correlation coefficient (Spearman).
SEPA	Scottish Environment Protection Agency
SSD	Species sensitivity distribution
SUSS	CIES code for suspended solids, mg/l
TEMP	CIES code for temperature, °C
TOXN	CIES code for total oxidised nitrogen, mg/l
u/s	Upstream
VBA	Visual Basic for Applications
WFD	Water Framework Directive
WHPT	Walley Hawkes Paisley Trigg index, a revision of ASPT using revised taxonomic sensitivity values

List of Appendices

Appendix A

Investigation into the potential use of toxicity data 148

Appendix B

Impact of flow condition on the occurrence of taxa 170

Appendix C

Proposed data specification 174

Appendix D

MIR-max User Guide 187

Appendix E

PISCES codes for sector, activity and pressure 209

Appendix F

Stress categories and associated activity, source and pressure codes 218

Appendix G

BBN Creator User Guide 224

APPENDIX A

Investigation into the potential use of toxicity data

Dr Veronique Adriaenssens
Environment Agency of England and Wales

Investigation into the potential use of toxicity data

Introduction and objectives

This work package is a feasibility study reviewing the possibility of using toxicity data as part of the Bayesian Belief Network (BBN) models. This study focuses on the toxic effects of pesticides on macroinvertebrate populations in rivers. A review of the possible use of pesticides as an input parameter in the BBN was needed because the current input variables mainly focus on nutrients, organic substances, pH and parameters that describe the river type. Because of the lack of pesticide survey data from river stretches where we have biological monitoring data, we need to look for alternative approaches that can use knowledge instead of survey data or can work in a combined data-knowledge way to model pesticide impacts on biota. This approach could be part of the current BBN network or separate from it.

This feasibility study is structured into four sections. The first section is a literature review on existing impact-response models for pesticides and macroinvertebrates in rivers. The second section is a review of studies within the UK related to pesticide impacts on macroinvertebrates. In the third section different model approaches are reviewed, and the conclusions of the feasibility study and recommendations for implementation are given in section four.

Literature review

This review covers different types of approaches for analysing the effects of pesticides on macroinvertebrate communities: field-based research, microcosm-mesocosm approaches, laboratory studies and knowledge-based approaches. The focus of this review is on evaluating each approach and its possible implementation in the current BBN model structure. A few examples are given to illustrate the general ideas and outcomes of each approach. Several of these and other approaches that could be used to develop a biological indicator of pesticide contamination are also reviewed in Schriever *et al.* (2008).

Field-based approaches

An example of a field-based approach is the study by Friberg *et al.* (2003). This study investigated the effects of sediment-bound pesticides on macroinvertebrate communities in rivers. Some of the outcomes are summarised below:

- Oligochaeta and leeches increased with pesticide exposure, as found in other studies.
- A drop in the number of amphipod *Gammarus pulex* is consistent with previous findings.
- A rise in number of Tanypodinae with increasing pesticide concentrations is only partly consistent with previous findings.
- The dipteran family Chironomidae appears to respond differently to pesticide exposure.
- In this study, no single species or taxonomic group within Insecta showed a clear negative relationship with pesticide concentration.

- Both leeches and Tanypodinae are predators and their increase might reflect the increase in prey, as the total number of macroinvertebrates rose with greater pesticide concentrations.
- It is difficult to assess the direct toxicity of all sediment-bound pesticides because studies on pesticide toxicity in sediments are scarce.
- Results confirm what has been found in mesocosm experiments and investigations involving only a few streams.
- It is not possible to separate other impacts from those of pesticides in this study.

A number of other field study approaches have been published (such as Schulz and Liess, 1999; Probst *et al.*, 2005).

Experimental studies investigating the impacts of a specific substance introduced into the field under controlled conditions have also been reported (such as Liess and Von Der Ohe, 2005).

Evaluation of field-based approaches

Field studies often provide us with valuable information because they take into account the effects of other environmental parameters on the toxic effects of a certain pesticide. However, because of the multitude of influencing parameters at any site, caused by the difficulty of finding pristine conditions, it is not entirely clear how to separate out the pesticide effects on the community of interest. The outcomes of such studies can often only be described in a semi-quantitative way, which can make it difficult to integrate them into models. To discriminate between real causal effects and statistical correlation, large quantities of biotic and abiotic data are needed and this is often not available, mainly because of the cost of measuring pesticides and macroinvertebrates.

Microcosm-mesocosm studies

Microcosms and mesocosms are model ecosystems, that is, experimental systems that mimic parts of natural ecosystems. The use of microcosms or mesocosms provides a bridge between the laboratory and the field, in terms of providing the opportunity to perform ecosystem-level research in replicable test systems.

As an example: a study by Schulz *et al.* (2002) describes the results of a combined microcosm and a field approach to evaluate the aquatic toxicity of azinphosmethyl to stream macroinvertebrate communities in South-Africa. The results are summarised as follows:

- Reduced invertebrate density, attributed mainly to various insect taxa, such as *Demoreptus* sp., *Castanophlebia* sp., Simuliidae and Chironomidae.
- In contrast, *Aeshna* sp., *Dugesia* sp., Ceratopogonidae and *Cheumatopsyche* sp. were unaffected.
- Field surveys: comparable results with microcosm.
- Microcosm studies employing a field-relevant design could be linked to field studies.

Evaluation of microcosm/mesocosm studies

There is considerable scope for using microcosm-mesocosm studies to gain insight into impact-effect relationships, as they can be linked more easily to the results of field studies than laboratory ones. However, they can be expensive and labour-intensive. Currently, we do not have enough data to specifically define the cause-effect relationships.

Laboratory studies

There are numerous examples of laboratory studies on macroinvertebrate species, and the main results are compiled in AQUIRE (Aquatic Toxicity Information Retrieval) Ecotoxicology Database from US EPA (www.epa.gov/ecotox). The database gives lethal concentration (LC₅₀), effective concentration (EC₅₀), no observed effect concentration (NOEC) and lowest observed effect concentration (LOEC) values (mean, standard deviation and median) for macroinvertebrate taxa at different taxonomic levels. Laboratory results are difficult to extrapolate to the field because of the limited number of environmental factors taken into account. However, a vast amount of laboratory toxicology data is available, although not always at the taxonomic level required, but one could try to amalgamate data to obtain more insight into the toxic effects (for example by species sensitivity distributions, Section 2.3.4).

Knowledge-based approaches

Results from field, laboratory or mesocosm/microcosm studies can be integrated in knowledge-based systems. These systems are often developed to support decisions in environmental management.

SPEcies At Risk list (SPEAR)

In Liess and Von Der Ohe (2005), species were grouped according to their sensitivity to pesticides and life-cycle traits known to influence recovery from toxicant stress. The data file with sensitivity values is available at <http://www.ufz.de> and is called the SPEAR list. The sensitivity values defined are based on a literature review. Current research projects in the UFZ Centre for Environmental Research in Germany show interesting approaches to system ecotoxicology. A British version of SPEAR has been produced (Beketov *et al.*, 2008) and is currently being evaluated. The results of this evaluation are not expected until mid 2010.

Classification of macroinvertebrates according to their relative sensitivities to toxic substances

A classification of macroinvertebrates according to their specific relative sensitivities to toxic substances is given in Wogran and Liess (2001), using the order as basic taxonomic level. The data evaluated for this purpose, drawn from the literature (via AQUIRE from US EPA), comprised LC₅₀ and EC₅₀ values for the exposure of macroinvertebrate species to many substances. Their relative sensitivities are calculated by comparison with toxicity data for the standard test species *Daphnia magna*, for which a large database is available. Only organic and metal toxicants were differentiated in this investigation. Von Der Ohe and Liess (2004) presented a similar study but LC₅₀ values were used in contrast to sub-lethal endpoints as used by Wogran and Liess (2001).

PERPEST

A number of publications have been published describing the expert model PERPEST (Van den Brink *et al.*, 2002, 2006; Van Nes *et al.*, 2003). The Wageningen University and Alterra have developed a case-based reasoning method to predict pesticide effects on freshwater ecosystems. This method is named Prediction of the Ecological Risks of PESTicides (PERPEST) and is incorporated into a user-friendly interface. A literature review resulted in a database containing the effects of 22 herbicides and 24 insecticides. In total 104 experiments were evaluated, resulting in 421 cases. The PERPEST model searches for analogous situations in the database based on

environmental fate characteristics of the compound, exposure concentration and type of ecosystem to be evaluated. A prediction is provided by using weighted averages of the effects reported in the most relevant literature references (cases are weighted based on the 'toxic unit', 'molecule group' and 'substance'). PERPEST results in the prediction showing the probability of 'no', 'slight' or 'clear' effects on the various grouped endpoints. Further information can be found at www.perpest.alterra.nl.

LIMPACT

LIMPACT is another expert system of interest (Neumann *et al.*, 2003 a,b; Neumann and Baumeister, 2005), developed by the Technical University of Braunschweig and the University of Wurzburg. This knowledge system estimates pesticide contamination in small streams within agricultural catchments using benthic macroinvertebrates as biological indicators. The system considers the abundance of 39 macroinvertebrate taxa during four time frames within a year. In the knowledge base, selected taxa are identified as a positive or negative indicators. Based on the scores of taxa and presence of (at least five) taxa from the list, it gives a diagnosis in one of four classes. The four diagnoses 'not detected', 'low', 'moderate' and 'high' pesticide contamination represent a calculated annual toxic sum without any specification of the chemical agents. The required input parameters of LIMPACT are abundance data for any of the 39 macroinvertebrate taxa in the stream. Apart from abundance data, LIMPACT evaluates nine basic water quality and morphological parameters such as stream size or conductivity of the water, to characterise the stream. Further information can be found at <http://www.limpact.de>.

Evaluation of knowledge-based systems

Although these systems are recognised as of great value, one has to take into account limitations mainly caused by using the outcomes of studies with different endpoints and different conditions to obtain a final result. However, for environmental management, they are of particular interest as these knowledge-based systems are able to make predictions for the component or community of interest and often integrate the best knowledge available. Some restrictions of the approaches are discussed below.

In the knowledge-based approaches, there is often one general rule that applies to different types of ecosystems (such as SPEAR) or the study is only applicable to a certain type of stream, for example, LIMPACT is only designed to estimate the pesticide contamination of small lowland headwater streams within an agricultural area. However, some authors have argued that threshold values and direct effects observed for the same compound are similar in different aquatic ecosystems (Brock *et al.* 2000a; Van Wijngaarden *et al.* 2005).

Some studies (such as LIMPACT) do not distinguish between different groups of pesticides. A practical way of specifying the different types of toxic effects caused by a group of pesticides is describing them by their toxic mode of action. De Zwart (2003) produced a list assigning each pesticide to a different 'mode of action' group.

Current knowledge-based systems are not specifically developed to assess the ecological risk of a mixture of pesticides. It would therefore be a great improvement if systems could estimate the overall ecological risks associated with measured concentrations of different pesticides. If one wants to add the effects of different pesticide concentrations, one possibility is to use 'toxic units' (TU). TU can be calculated by different methods, most of which calculate TU values based on the acute (48-hour) LC₅₀ of *Daphnia magna* (such as PERPEST).

The taxonomic resolution of studies is often of concern. None of the knowledge-based systems considered work at the same taxonomic level, although there might be

opportunities to, given that the information compiled often originates from the species level. Knowledge-based systems are either based on the species level (such as LIMPACT, limited number of species) or the order level (such as PERPEST distinguishing between macrocrustaceans, microcrustaceans and insects). BMWP (Biological Monitoring Working Party) families are used in the Environment Agency's Artificial Intelligence (AI) models.

It is questionable whether sufficient empirical data is available for predictions to be made, although the uncertainty linked to predictions is expressed when using the knowledge bases. The current shortage of empirical data is reflected in the 95 per cent confidence intervals of predictions, which are usually quite large when probabilities around 50 per cent are predicted.

It is difficult to combine information from laboratory tests and field studies as they represent different conditions for the communities considered. However, microcosm/mesocosm studies employing a field-relevant design could be linked successfully to field studies.

In all knowledge-based systems, recovery of species is not included as a parameter of concern. However, data from the sensitivity list produced by Liess and Von Der Ohe (2005) can be used for that purpose.

Conclusions and future approaches

When reviewing the different approaches, there is a trend in favour of microcosm/mesocosm studies or field-based (experimental) approaches over laboratory-based studies when a prediction of the effect of a pesticide on a biological community is required. They show toxic effects on riverine macroinvertebrate communities more realistically. There are, however, issues about the reproducibility of the results and being able to factor out the pesticide effects in streams where a range of pressures interact. Research in this area is growing, which is clear from the range of new European approaches, some of them even integrating biological interactions (for example, INTERACT - Improving EU risk assessment of toxicants for aquatic communities by considering competition on the population and community level). Liess *et al.* (2005) summarise the findings of a European workshop on the effects of pesticides in the field and give a good overview of research in this area. The paper focuses on opportunities and limitations of field studies (including microcosms/mesocosms) and implications for regulatory risk assessment and management. To conclude, there is much valuable data and information available, but this doesn't always lead to more insight into clear cause-effect relationships.

Overview of UK data and studies on pesticide impacts on macroinvertebrates

Environment Agency pesticide monitoring network

The Environment Agency monitors pesticides at a number of sites in England and Wales. However, this data is limited in the following way:

- Monitoring is driven by statutory requirements to report against standards for certain active ingredients. Many of these are no longer approved, such as DDT.
- Most of the measured concentrations are 'less than the limit of detection'.
- The number of sampling sites is limited.

- Pesticides are often monitored at a different site and time than the macroinvertebrate community.
- Generally, pesticides are only measured by spot samples from the water phase rather than from sediment samples.
- The monitoring programme is not determined centrally and so there is bias between regions.

The limitations of this data for determining cause-effect relationships between the pesticide pressure and biological communities were highlighted in Crane *et al.* (2003).

Other sources of data are the Environmental Change Network, Countryside Surveys and other studies, some of which are discussed in the next section.

Crane et al. (2003). Assessing the impact of agricultural pesticides in the aquatic environment. A scoping study.

Environment Agency data were analysed as part of this study to assess the impact of agricultural pesticides on the aquatic environment. The conclusions of the study were:

- Neither biological nor chemical data collected by the Environment Agency were adequate to establish whether pesticides are causing adverse effects in UK surface waters.
- The location and frequency of sampling is insufficient to detect effects.
- There are insufficient links between samples taken for pesticide analysis and biological sampling.
- The biological monitoring network does not cover all potentially affected water bodies and may under-represent small headwater streams that are, potentially, most at risk.
- Any change in the invertebrate assemblage because of pesticides is indistinguishable from other changes because of other environmental gradients.
- Results of the analyses, in combination with the literature review, suggested that the following invertebrate families could be used to discriminate between sanitary and other types of pollution: Gammaridae, Asellidae, Coenagriidae and Baetidae.
- The data did not provide a firm basis for the conclusions made, according to the authors (under-representation of headwater streams, BMWP ranking bears large uncertainty).

Humpheryes, I. & Bennet, B. (Unpublished MS). The impact of pesticides in the Teise catchment leading to the development of a pesticide index. (Contact = Ian Humpheryes, Environment Agency).

The main result from this study is a table showing the rank order sensitivity of macroinvertebrate BMWP families to pesticides. The ranking of sensitivity to the specific group of pesticides investigated is different to the sensitivity to organic pollution as defined by the BMWP score system. In a further stage of this project, a pesticide index was developed using this ranked list of taxa. To use the index, the sample that is under investigation is run through RIVPACS (River Invertebrate Prediction and Classification System) to produce a predicted list of taxa in decreasing probability of capture. Provided a taxon is on the pesticide index list and was predicted to be in the sample but was actually absent, the probability of capture is entered against that taxon on the pesticide index database.

A summary of investigations of sheep dip pollution in Southwest Wales 2002-2004 by Graham Rutt. Proposal for the assessment of the impact of sheep dip pesticides on

watercourses in Wales in 2005 by Jerry Griffiths. (Contact = Jerry Griffiths, Environment Agency)

Extensive use of sheep dip pesticides in the catchment area led to a reduction in the abundance or elimination of stoneflies and mayflies whilst other insect species remained in reasonable abundance. This is highly characteristic of pollution by pyrethroid sheep dip.

C.D. Brown, L. Maltby, J. Biggs, P. Van den Brink, M. Liess. WEBFRAM2: Web-integrated tools to support higher-tier aquatic risk assessment.

WEBFRAM is a suite of projects funded by Defra to develop a web-integrated model framework for the assessment of risks to non-target organisms from pesticides. The models include explicit descriptions of variability and uncertainty and provide a toolbox to support higher-tier risk assessment with the context of European Directive 91/414/EC concerning the placing of plant protection products on the market. A database of life-history characteristics has been developed for 150 representative species and this will be linked to a model to predict within-site recovery of impacted populations.

ADAS (2000). The efficacy of no-spray buffer zones in protecting field boundary watercourses from pesticide spray drift. Report on Project PS0417 to the Ministry of Agriculture, Fisheries and Food, London, UK.

In this study, invertebrate assemblages were monitored at the stream site, using Surber samplers and drift nets, so that effects on naturally occurring populations could be compared to effects on caged organisms. Application of chlorpyrifos to the edge of the stream on two occasions was associated with reduced acetylcholinesterase and feeding activity in caged *Gammarus pulex*. However, there was no strong evidence for immediate or long-term effects on the abundance of naturally occurring *G. pulex* populations, or on the overall structure of macroinvertebrate assemblages.

The Ponds Conservation Trust: Policy and Research, c/o Oxford Brookes University. Aquatic ecosystems in the UK agricultural landscape.

This project analysed national freshwater datasets to create a characterisation of aquatic habitats in the UK agricultural landscape. Regional field data were collected to support and test the findings. Desk studies were undertaken to review the main factors that determine the exposure of, and risk posed to, aquatic species and habitats in agricultural areas.

Brown et al. (2006) Assessing the impact of agricultural pesticides in the environment.

This report covers the initial stage of a project to determine whether pesticides used in the approved manner cause adverse effects in the aquatic environment. National risk mapping and landscape analysis were used to identify the high risk crop-pesticide scenarios and high risk locations. The highest risk scenario was spray drift from orchard crops. However, on visiting these locations to identify monitoring sites it was discovered that these sites were also affected by other stressors (such as saline intrusions, canalisation, nutrients). The project was stopped before monitoring began as the authors were not certain that any impacts detected could be attributed to pesticide exposure with confidence.

Conclusions

An overview of studies to unravel the effects of pesticides on macroinvertebrate communities in the UK is given. The main conclusions from this (limited) set of studies are:

- Datasets currently available are not sufficient for detailed analysis and modelling.
- Although a few reports state that any change in the invertebrate assemblage caused by pesticides is difficult to distinguish from changes caused by other environmental gradients, it seems to be possible to detect pesticide stress in a river based on a ranking procedure or a selection of macroinvertebrate families. However, the approaches need to be supported by research and need to be tested in different types of catchments (ongoing in Environment Agency, Ian Humphreys).
- Some recent projects, such as WEBFRAM might offer opportunities to develop a knowledge-base, bringing data and research conclusions together.

Possible approaches for modelling the impact-effect: feasibility study

A. General Model Approach

If pesticide concentration data is available along with macroinvertebrate data monitored at the same site, the following model structure (Figure A1) could be used as part of the overall BBN network to model the toxic effect of pesticides on macroinvertebrate communities.

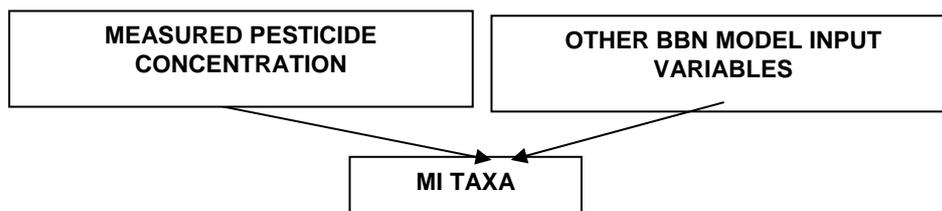


Figure A1 BBN model structure modelling effect of pesticides on macroinvertebrate communities in a river stretch.

Because of the lack of pesticide monitoring data, other opportunities for integrating knowledge and/or data that can be used to structure an impact-effect model for pesticides need to be analysed. This can be done by (a) estimating pesticide concentrations based on alternative input variables, or (b) using the knowledge from studies to model the impact-effect relation between macroinvertebrates and pesticides, taking into account the effect of other pressures acting on the biota.

Estimating pesticide concentrations from other (alternative) input variables

The model structure required to account for the effect of pesticides on macroinvertebrate communities in rivers is shown in Figure A2.

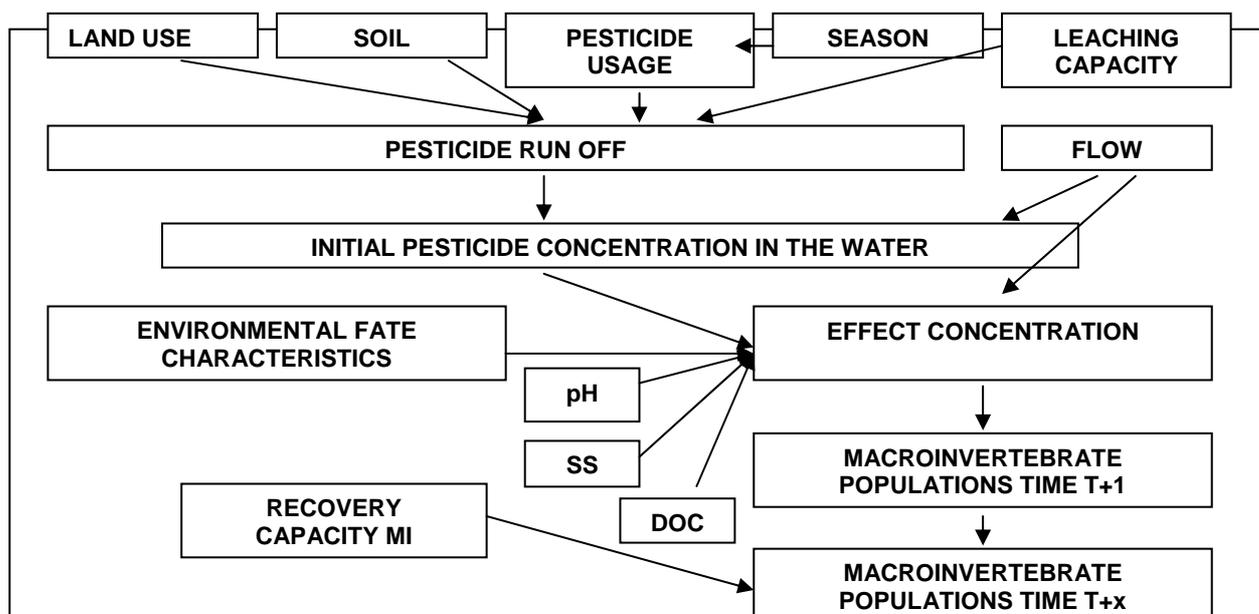


Figure A2 Model structure required to estimate toxic effects of pesticides on macroinvertebrate communities in a river stretch.

If pesticide concentrations are not available, alternative variables could be used to estimate pesticide concentrations in the river. These variables, as given in Figure A2 are land use, soil, pesticide usage and leaching capacity. To account for the effects of other environmental variables that determine the final effect of a pesticide on the macroinvertebrate community, data for parameters such as pH, flow suspended solids and dissolved organic carbon are required. The environmental fate characteristics of the pesticide have to be known as well. Potential data or information sources available for the input variable into the model (from model in Figure A2) are:

- CORINE (Coordination of information on the environment) land cover map (EC programme), though this is not very refined towards specific crops.
- Agricultural census and land cover maps (such as Brown *et al.*, 2006).
- Soil maps, available from British Geological Survey.
- Pesticide usage statistics are available from 1990 onwards from pesticide usage surveys commissioned by the Advisory Committee on Pesticides. Data is collected by Pesticide Usage Survey Teams at the Food and Environment Research Agency and the Scottish Agricultural Science Agency (<http://pusstats.csl.gov.uk/>). Other pesticide usage data are available commercially.
- The Groundwater Ubiquity Score (GUS, Gustafson, 1989) can be used to provide an estimate of the leachability of a pesticide active ingredient, or can be used to classify pesticides into various categories.
- DOC and SS: these parameters are indirect predictors of the percentage oxygen saturation that is part of the current BBN structure.
- Time series: a pesticide measurement is only a snapshot in time and the survival of macroinvertebrates will mainly be determined by severity (concentration) and length of exposure to the pesticide. This is determined mainly by flow velocity, fate of the pesticide and recovery potential of the taxon itself. Hence, an integrated time series of pesticide concentrations would be useful, but is not yet possible.
- Season, flow: these parameters are available in the current BBN structure.
- Fate of the pesticide: a pesticide's fate is described by how and where it enters the environment, how long it lasts, and where it goes. A measure of how long a pesticide lasts in an environment is given in the Environment Agency pesticides handbook. The other components of environmental fate

are much more complex, but suspended solids, DOC, pH and flow partly determine the fate of a pesticide in a river.

- Recovery capacity of the macroinvertebrate taxa: these parameters are listed by Liess and Von Der Ohe (2005) and are in the public domain.

Conclusions

The parameters required such as season, land use, soil, pesticide usage, leaching capacity, season, flow and pH (as given in Figure A2) could be included in the current BBN structure, but there is an issue about (a) data quality, (b) data resolution and (c) lack of data. The temporal and spatial aspects of the prediction would also need to be reviewed. There is potential in using this approach as most of the data are available, but a thorough review of all the parameters involved is needed. The BBN can be trained to derive the relation between different pesticide components (which could be expressed in terms of toxic units) and the macroinvertebrate communities.

An ongoing Environment Agency project to develop a decision support tool is a way of modelling pesticide concentrations in rivers without using the BBN structure. It incorporates aspects of CatchIS (www.catchis.com, a collaboration between Cranfield University and ADAS). Within this there are two surface water models: SWATCATCH (Hollis and Brown, 1996) - predictions of concentrations at catchment outlets and SWAT (Brown and Hollis, 1996) - predictions of concentrations at the field edge. Both models predict pesticide concentrations based on diffuse agricultural inputs. Point source inputs of pesticides and inputs from non-agricultural uses of pesticides are not considered. SWATCATCH was also the model used in the Environment Agency's previous tool, Prediction of Pesticide Pollution In the Environment (POPPIE). Further information can be obtained from Neil Preedy (Environment Agency Science, Geosystems) and Anthony Williamson (Environment Agency).

B. Species sensitivity distributions

General

Because the approach using species sensitivity distributions (SSD) is a vast and complex area of research, this section starts with some references to key publications in this field.

What are species sensitive distributions?

Newman, M.C., Ownby, D.R., Mezin, C.Z. Posell, D.C. Christensen, T.R.L., Lerber, S.B and Anderson, B.A. (1999). Applying species-sensitivity distributions in ecological risk assessment: assumptions of distribution type and sufficient number of species. *Environmental Toxicology and Chemistry*, 19(2), 508-515.

Van den Brink, P.J., Brown, C.D. and Dubus, I.G. (2006). Using the expert model PERPEST to translate measured and predicted pesticide exposure data into ecological risks. *Ecological Modelling*, 191, 106-117.

Species sensitivity distributions are used to calculate concentrations at which a specified proportion of species will be affected, referred to as HC_p , the hazardous concentration for $p\%$ of the species. It is a popular method to extrapolate from the species to the community level. A statistical distribution is estimated from a sample of toxicity data (LC/EC₅₀ or NOEC values) and visualised as a cumulative distribution function. Although there is growing evidence that HC_5 are indeed protective for the aquatic ecosystem, the approach makes a number of assumptions.

Concerns with this method can be summarised as follows:

- There is a bias towards mortality data despite the plausibility of non-lethal effects.
- Laboratory-based tests do not represent field conditions.
- There is no evaluation of function – only diversity loss is evaluated.
- There is no specific importance to keystone species or on the influence of species interactions.
- Distributions might not be log-normal, as is assumed by this method.
- The assumption is that the data are from a random sample of species. In practice, the species for which data are available are determined in non-random ways and are likely to be highly non-representative of the population (as some species are more widely cultivated for laboratory use than others).
- These studies often rely on data from species not native or indigenous to the country where the toxicology data are used.

Some illustrations of applications of SSD in UK studies:

P. Whitehouse, M. Crane, E. Grist, A. O'Hagan and N. Sorokin. (2004) *Derivation and expression of water quality standards. Opportunities and constraints in adopting risk-based approaches in EQS setting*. Environment Agency R&D Technical Report (P2-157/TR). Bristol, Environment Agency.

Maltby, L., Blake, N., Brock, T.C.M. and Vanden Brink, P. (2005). Insecticide species sensitivity distributions: importance of test species selection and relevance to aquatic ecosystems. *Environmental Toxicology and Chemistry*, 24, 2, 379-388.

Maltby *et al.* (2005) provide an overview of possible risks of using SSD:

- Test species are not representative of the ecosystem: SSD should select data for those species that occur in the ecosystem under consideration.
- Test species in lab (single species test) can be difficult to extrapolate to species in an ecosystem (multispecies).
- A limited range of species may be tested, and this may mainly include test species recommended by OECD, USEPA and other international organisations.

Maltby *et al.* (2005) advise on which taxonomic groups to consider. One of their conclusions is that insecticides are more toxic to arthropods than to vertebrates or non-arthropod invertebrates such as Mollusca, Annelida and Platyhelminthes.

A few conclusions based on case studies evaluated by Maltby *et al.* (2005) explained:

- SSDs for two insecticides were compared for freshwater arthropods, and no evidence of a difference among or within compounds was found (comparison of acute lab SSD and mesocosm SSD): similar or related species do not have different sensitivities under field or laboratory conditions.
- Acute toxicity data for freshwater arthropods from different geographical regions and different freshwater habitats may be combined with a single SSD (data from freshwater and saltwater can be combined but one must be aware of the effect of differences in taxonomic composition, especially for Crustacea).

- There is insufficient chronic toxicity data for most chemicals to generate appropriate sensitivity distributions and so it is difficult to make any general conclusions.

Case study: Developing SSD for macroinvertebrates at family level

In this study, the possibility of defining species sensitivity based on laboratory ecotoxicology data by means of the USEPA AQUIRE database was analysed.

Data was abstracted from the AQUIRE database for the following pesticides: carbendazim, chlorpyrifos, cypermethrin and diazinon. Features of each pesticide are explained below:

Carbendazim

Carbendazim is a systemic benzimidazole fungicide with a range of agricultural, horticultural and home/garden uses. Leaching is transitional between high and low (GUS 2.14). Although highly toxic to aquatic organisms, it is unlikely to cause problems during normal agricultural practice because of low bioavailability.

Chlorpyrifos

A broad-spectrum organophosphate insecticide used on a wide variety of horticultural and arable crops, but mainly on wheat, grassland and apples. It has very low leaching potential (GUS 0.37). As the pesticide adheres to sediments and suspended organic matter, concentrations rapidly decline. Volatilisation is probably the primary route of loss of chlorpyrifos from water. It is very toxic to aquatic invertebrates.

Cypermethrin

Cypermethrin is a synthetic pyrethroid insecticide used to control many insect pests. Usage is mainly on wheat, barley and oilseed rape and (previously) as a sheep dip. Leaching is transitional between high and low (GUS 1.97). Cypermethrin is stable to hydrolysis and photodegradation. It is very toxic to aquatic invertebrates.

Diazinon

Diazinon is a broad-spectrum organophosphate insecticide. It is the main component of sheep dip. Low application of diazinon on UK farmland is obvious from the pesticide usage results as it is mainly used for sheep dip which is not reviewed. The composition of sheep dip is as follows: 65 per cent diazinon, 27 per cent cypermethrin, eight per cent flumethrin.

The criteria used (and associated references) for selection of these four pesticides are:

- Pesticide usage data from these pesticides were available from the Central Science Laboratory: <http://pusstats.csl.gov.uk/index.cfm>.
- Toxicology data for these pesticides were available from the EPA database. <http://www.epa.gov/ecotox/>. Available toxicology data of macroinvertebrates affected by pesticides (chosen from list of selected pesticides Science Pesticide Project).
- These pesticides were part of the Water Framework Directive (WFD) list of pesticides registered in the UK and listed according to their harm to algae/invertebrates/fish.
- Diazinon and cypermethrin are used in sheep dip. There is no good record of sheep dip use (it is not taken into account for the programme of pesticide usage, but seems to be a major problem as decline in macroinvertebrate diversity and abundance is obvious in areas affected by sheep dip pollution). Sheep dip used to consist of 65% diazinon, 27% cypermethrin, 8% flumethrin. Recent data suggest flumethrin is no longer approved and the marketing authorisation for cypermethrin was suspended in February

2006, although farmers could still use and buy existing stocks, so the only approved active agent for sheep dip is diazinon at the moment. However, at the time of the study, the three components were suggested as being of importance).

The US AQUIRE database provides NOEC, LOEC, EC₅₀ and LC₅₀ concentrations based on results from field studies, laboratory experiments or mesocosm/microcosms.

The objective was to develop a species sensitivity distribution based on NOEC, LOEC, EC₅₀ or LC₅₀ values, as the different endpoints for developing a SSD cannot be combined. Two problems were encountered: firstly, most of the taxa do not have results for all four endpoints and secondly, for most of the macroinvertebrate families common in the UK, there are only a few species and these may not be fully representative of the family. It may be necessary to consider the family response as relatively homogenous in response to pesticides (similar physiology and sensitivity because of similar life cycle and life stages). The revised coded checklist of freshwater animals in the British Isles (Mike Furse, Centre for Ecology and Hydrology) was used to select UK taxa from the AQUIRE database.

The development of SSDs based on expert knowledge is possible and is partly illustrated in O'Hagan *et al.* (2005). Other knowledge sources could provide valuable input for SSDs.

Conclusions

As different endpoints are involved in studies, only some species have enough data to create SSDs. The extrapolation of species-level data to the family level is rather difficult as the family data are only represented by a few (test) species, which might not include the full range of responses towards pesticide toxicity. The data do not include any knowledge about recovery capabilities of macroinvertebrate taxa although this information is available from Liess and Von Der Ohe (2004). As most of the data from the AQUIRE database originated from laboratory studies, we are uncertain about extrapolation to the field, although the BBN model could account for influences of the main environmental parameters at a river stretch.

C. The msPAF approach

General

The Dutch National Institute for Public Health, RIVM, recently developed an approach called msPAF: Multiple Substance Potentially Affected Fraction of species. It differs in a number of technical respects from the BBN although it has similar aims. Rather than the empirical approach used in the BBN, msPAF builds on ecotoxicological principles to identify probable causes of local impacts from chemicals and other factors at discreet locations. These are presented visually as pie charts denoting the size of impact and likely contributing causes.

The focus of much of the work is on toxic chemicals. Species sensitivity distributions are used in conjunction with a database of ecotoxicological data to predict the proportion of species that would be affected (PAF) by a given chemical concentration. This approach is now well established in ecotoxicology, and is gaining increasing regulatory acceptance for chemical risk assessment and the setting of environmental standards. The method integrates effects of multiple stressors by summing the proportion of species predicted to be affected from each individual stressor. Thus, a single PAF can be estimated for a single location that can be regarded as a 'risk ruler' for the level of risk resulting from multiple pressures. It assumes that the stressors

interact in an additive fashion (concentration addition for chemicals with the same mode of action and response addition for chemicals with different modes of action).

The degree to which impacts can be explained by physical, water quality and toxic chemicals is estimated by generalised linear modelling (GLM) in which msPAF predictions are used to determine the contribution from toxic chemicals. GLM also gives an estimate of unexplained differences between expected and observed biology, that is, residual impacts that the selected factors cannot explain.

Further references to the msPAF technique are given below:

De Zwart (2003) *Ecological effects of pesticide use in the Netherlands. Modelled and observed effects in field ditch*. RIVM report 50000203/2003

Leo Posthuma and Dick De Zwart. *Diagnostic tools for monitoring data - the power of ms-PAF*. Environment Agency catchment pressures workshop.

Dick De Zwart. *Eco-epidemiology, biodiversity and toxic risk. A case study*.

SSD and msPAF methodology

The following paragraph gives a brief overview of the msPAF method. The text is based on De Zwart (2003). The illustrations given are based on toxic exposure of a field ditch.

The calculated exposure of the field ditch is converted to the estimate of risk by applying SSD and theory on mixture toxicity. The risk is expressed in terms of the fraction of species that is expected to be exposed to a concentration or a mixture exceeding levels where effects are considered negligible. Figure A3 shows an exemplar cumulative probability distribution of species sensitivity fitted to observed chronic toxicity values (NOEC).

EC₅₀ (acute median effect concentration) and chronic NOEC (no observed effect concentration) were log₁₀ transformed before calculating the average log (toxicity) over major taxonomic groups and the associated standard deviation.

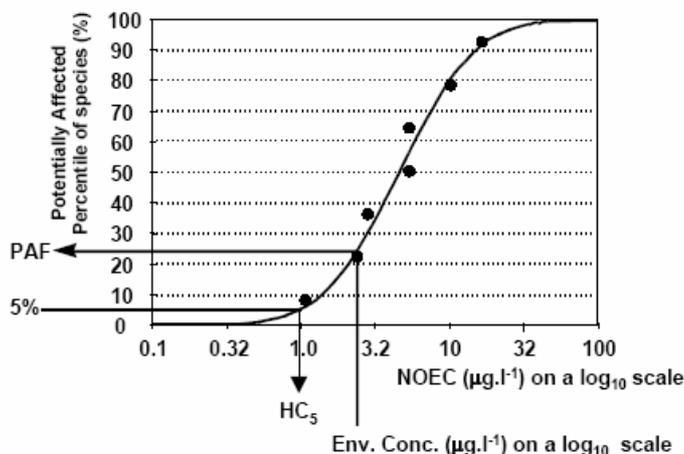


Figure A3 Exemplar cumulative probability distribution of species sensitivity fitted (curve) to observed chronic toxicity values (NOEC; dots). Arrows indicate the inference of a Potentially Affected Fraction of species (PAF-value) and HC₅ (from De Zwart, 2003).

Eighteen different major taxonomic groups were recognised in the study by De Zwart (2003) as given in Table A1.

Table A1 Taxonomic groups represented in the toxicity data (De Zwart, 2003).

Taxonomic group	Taxonomic Group
Insects (larval stage)	Amphibians
Liverworts and Ferns	Annelids
Reed and grasses	Mites and spiders
Molluscs	Bacteria
Nematoda	Arrow worms
Fish	Hydroids
Flatworms	Crustaceans
Protozoa	Cyanobacteria
Rotatoria	Algae

The toxic risk per pesticide ingredient is calculated first. Then the toxic risk per gradient and per major taxon was averaged over the major taxonomic groups. For ingredients with the same toxic mode of action (TMoA), concentration additivity is assumed. The weekly calculated concentrations per ingredient are transformed to hazard units per taxonomic group. The weekly combined toxic risk per TMoA and per major taxonomic group is then calculated. Response addition per major taxonomic group is calculated where it is assumed that the species are uncorrelated in their sensitivity for the different toxicants. Finally, the average ms-PAF taxonomic group is calculated, assuming equal weight of major taxonomical groups.

Validating msPAF (based on de Zwart, 2003)

From de Zwart (2003): The risk scale (PAF) is dimensionless, but based on the sensitivity of species under laboratory conditions. In view of these facts, the association between risk and changes in biodiversity is not obvious. However, if the calculated overall toxic risk of pesticide exposure to aquatic species, expressed as the proportion of species expected to suffer effects from exposure, is considerable and properly scaled, this should be reflected in species composition in the field. Pesticide toxicity is not the only environmental condition governing species composition. A plethora of physico-chemical and habitat characteristics, as well as biological interactions, all determine the type of community to be expected. The observed species composition in the field, in terms of the number and abundance of species, may be related directly to the predicted toxic risk of pesticide exposure. This will be easy to determine when the driving force of pesticide toxicity has a major influence over other driving forces. However, in view of the absence of extreme exposure levels and the expected relevance of other driving forces, this approach was considered unlikely to yield sufficient explanatory power. To be able to isolate the slight effects of pesticide exposure from a dataset on measured biodiversity, other driving forces have to be taken into account.

In the study of De Zwart (2003), environmental predictors were taken into account by multiple linear regression (GLM), to allow for the environmental impact of other factors than pesticides.

Effects of modelled risk: comparison with field data (based on de Zwart (2003)): There seems to be a weak relationship between predicted mixture risk values and species composition.

Approaches to follow according to de Zwart (2003): (a) attribution of a tolerance score to individual species and (b) relate species composition to a reference community.

Conclusions

A meeting was organised on 27 February 2006 to discuss the ms-PAF approach and its link to the AI project. Attendees from the Environment Agency were Paul Whitehouse (PW), Claire Wells (CW) and Veronique Adriaenssens (VA). VA presented the outcomes of the feasibility study (including pesticides in BBN) and discussed opportunities of msPAF approach and its link with BBN with PW and CW.

Paul Whitehouse summarised the 'what next' phase in the *Identifying Catchment Pressures* workshop (Cribbs Lodge, 12 December 2005) taking into account the outcomes of the workshop as well as the meeting on 27 February 2006 as follows:

Further development is required before useful methods can be deployed. Two opportunities for development are:

1. The influence of toxic chemicals on biology within the BBN is under development, but limited by the small quantity of field data. The msPAF approach provides a mechanistic way of attributing impacts of such stressors, arguably making more effective use of toxicological data to avoid relying solely on field data. It would be sensible to examine whether the msPAF could be incorporated as a 'module' within the BBN (with the BBN possibly replacing the GLM regression step). There is an opportunity to bring both approaches together in a way that is mutually beneficial.
2. The Environment Agency possesses substantial biology and chemical datasets collected for GQA monitoring. These datasets underpin the development of the BBN but they could also be used to trial the msPAF approach in one or more UK catchments, in a similar way to that described for Ohio. For example, a rural catchment where land use and habitat factors might dominate impacts could be selected alongside an industrial catchment in which toxic impacts might assume greater importance. Data on present/absent species will be important given the insensitivity of classical biodiversity indices to toxic chemicals. Coincident sampling for biology and chemistry will also be important.

We propose a meeting between the Environment Agency (Water Management, National Data Unit, Science), RIVM and University of Stafford to determine whether the ideas above can (and should) be pursued. The discussion would identify what is technically achievable with the data resources at our disposal and current scientific understanding, what the realistic benefits might be, the timescale over which these benefits could be delivered, and what a suitable mechanism for collaboration might be.

This initiated a scoping project to trial this technique on a dataset collected by the Environment Agency that is reported in De Zwart *et al.* (2008). Because of the lack of measured pesticide data, modelled data from the POPPIE database was used instead. The findings suggest that the loss of species from some sites could be attributable to pesticide exposure. It is hoped that the calculated msPAF values that represent total pesticide toxicity could be incorporated into the BBN as an input parameter in a future project.

Final conclusions and recommendations of feasibility study

Conclusions

The following issues affect the feasibility of using pesticides in the BBN model structure:

- The almost complete lack of pesticide concentration data.
- The almost complete lack of field data on effects on macroinvertebrate communities.
- The many limitations involved in using laboratory toxicology data to develop family or species distributions, although some of these may be overcome by the msPAF approach (De Zwart *et al.*, 2008).
- Estimating pesticide concentration in the river from other input parameters is feasible but needs more research. The main issue is the spatial and temporal resolution that is different for all the input parameters in the potential model. The approach based on the CatchIS model offers opportunities (Brown and Hollis, 1996).

Recommendations

A combination of data and knowledge could be used to create a BBN that gives us shifts in abundance levels for BMWP families, based on field data, toxicology data and knowledge from studies (see literature review).

Field data on pesticide concentrations can be requested from the Environment Agency's data centre at Twerton. These data would be of value as part of the BBN network when combined with toxicology data and information from the literature and the various knowledge bases mentioned in this study. However, the lack of targeting and consistency in the monitoring programme is a drawback in using these data in the BBN modelling network.

Toxicology data and knowledge from different studies are mainly at higher taxonomic levels than the macroinvertebrate family level. Further discussion is needed on whether the taxonomic levels used by Maltby *et al.* (2005) might be appropriate (Mollusca, Annelida, Platyhelminthes, Insecta), but the Environment Agency would certainly be interested in family-level impacts as all tools developed so far have been to BMWP family level. We do not have enough knowledge to assume the genera or species within one family have a different sensitivity towards the impact of pesticides. Moving to a higher resolution would incur a higher cost but if different species react in different ways to toxic components we need more specific information at the species level. This domain remains open for further research.

As we do not have information to show the effect of each pesticide component on all macroinvertebrate families, we can group pesticides based on their toxic mode of action or select the 'priority' pesticides as was done in this study. The toxic effect can be calculated by means of species sensitivity distribution (SSD) approach given by the msPAF approach.

Integrating the msPAF approach in the BBN structure (as an alternative to GLM modelling by RIVM) would be a good way forward. This would offer us the ability to model all the main environmental impacts on macroinvertebrate communities (at the family level) at a site and what the expected 'natural community' at a certain site would look like (based on the reference condition). The combined BBN-msPAF approach

would use a valuable source of Environment Agency data and integrate a state-of-the-art approach of modelling the pesticide impact on macroinvertebrate families. One limitation so far is the use of modelled pesticide data, as these are catchment-scale predictions.

References

ADAS (2000). *The efficacy of no-spray buffer zones in protecting field boundary watercourses from pesticide spray drift*. Report on Project PS0417 to the Ministry of Agriculture, Fisheries and Food, London, UK.

Beketov, M.A., Foit, K., Biggs, J.P., Sacchi, A., Schäfer, R.B., Schriever, C.A., Liess and M. (2008). *Freshwater biological indicators of pesticide contamination – an adaptation of the SPEAR approach for the UK*. Environment Agency Science Report SC030189/SR4. Bristol, Environment Agency.

Brown, C.D. and Hollis, J.M. (1996). SWAT: A semi-empirical model to predict concentrations of pesticides entering surface waters from agricultural land. *Pesticide Science* 47, 41-50.

Crane M., Wells C, Pemberton E. and Croxford A. (2003). *Assessing the impact of agricultural pesticides in the aquatic environment: a scooping study*. Environment Agency Report for Project 12545. Bristol, Environment Agency.

De Zwart, D. (2003). *Ecological effects of pesticide use in the Netherlands. Modelled and observed effects in the field ditch*. RIVM, report 500002003/2003, the Netherlands.

De Zwart D., Pemberton E., Posthuma L., Veal A. and Wells C. (2008). *Understanding ecological impacts in rivers in England and Wales and identifying their possible causes: Part 1, the Effect and Probable Cause (EPC) method*. Environment Agency Science Report - SC030189/SR5. Bristol, Environment Agency.

Environment Agency (2001). *Collation and Analysis of UK Pesticide Monitoring Information (2001). Monitoring of pesticides in the environment*. ADAS.

Brown C., Beulke S., Biggs J., Holmes C., Maltby L, van Beinum W., Williams R and Yallop M. (2006). *Assessing the impact of agricultural pesticides in the environment (phase II)*. Environment Agency Science Report SC030189/SR1. Bristol, Environment Agency.

Friberg, N., Linstrom, M., Konvang, B. and Larsen, S.E. (2003). Macroinvertebrate/sediment relationships along a pesticide gradient in Danish streams. *Hydrobiologia*, 494, 103-110.

Gustafson (1989) Groundwater Ubiquity Score: a simple method for assessing pesticide leachability. *Environmental Toxicology and Chemistry*, 8, 339-357.

Hollis, J.M. and Brown, C.D. (1996). A catchment-scale model for pesticides in surface waters. In: *The Environmental Fate of Xenobiotics. Proceedings X Symposium on Pesticide Chemistry*, (Editors A.A.M. Del Re, E. Capri, S.P. Evans & M. Trevisan. Universita Cattolica del Sacro Curo, Piacenza. Italy). pp 378-379.

Liess, M. and Von Der Ohe, P.C. (2005). Analyzing effects of pesticides on invertebrate communities in streams. *Environmental Toxicology and Chemistry*, 24(4), 954-965.

- Maltby, L., Blake, N., Brock, T.C.M. and Van den Brink, P. (2005). Insecticide species sensitivity distributions: importance of test species selection and relevance to aquatic ecosystems. *Environmental Toxicology and Chemistry*, 24, 2, 379-388.
- Neumann, M. and Baumeister, J. (2005). A rule-based versus model-based implementation of the knowledge system LIMPACT and its significance for maintaining and discovery of ecological knowledge. In: Lek, S., Scardi, M., Verdonschot, P., Jorgensen, S.E. & Park, Y.S. (Editors) *Modelling community structure in freshwater ecosystems*. Springer-Verlag, Berlin.
- Neumann, M., Liess, M. Schulz, R. (2003a). An expert system to estimate the pesticide contamination of small streams using benthic macroinvertebrates as bioindicators. I. The database of LIMPACT. *Ecological Indicators*, 2, 379-389.
- Neumann, M., Liess, M. Schulz, R. (2003b). An expert system to estimate the pesticide contamination of small streams using benthic macroinvertebrates as bioindicators. II. The knowledge base of LIMPACT. *Ecological Indicators*, 2, 391-401.
- Newman, M.C., Ownby, D.R., Mezin, C.Z. Posell, D.C. Christensen, T.R.L., Lerber, S.B and Anderson, B.A. (1999). Applying species-sensitivity distributions in ecological risk assessment: assumptions of distribution type and sufficient number of species. *Environmental Toxicology and Chemistry*, 19(2), 508-515.
- O'Hagan, A., Crane, M., Grist, E. and Whitehouse, P. (2005). *Estimating species sensitivity distributions with the aid of expert judgements*. Research Report N0. 556/05. Department of Probability and Statistics, University of Sheffield.
- Probst, M., Berenzen, N.m Lentzen-Godding, A., Schulz, R. and Liess, M. (2005). Linking land use variables and invertebrate taxon richness in small and medium-sized agricultural streams on a landscape level. *Ecotoxicology and Environmental Safety*, 60, 140-146.
- Schriever, C.A., Callaghan, A., Biggs, J.P. and Liess, M. (2008). *Freshwater biological indicators of pesticide contamination*. Environment Agency Science Report SC030189/SR3. Bristol, Environment Agency.
- Schulz, R. and Liess, M. (1999). A field study of the effects of agriculturally derived insecticide input on stream macroinvertebrate dynamics. *Aquatic Toxicology*, 46, 155-176.
- Schulz, R., Thiere, G and Dabrowski, M. (2002). A combined microcosm and field approach to evaluate the aquatic toxicity of azinphosmethyl to stream communities. *Environmental Toxicology and Chemistry*, 21(10), 2172-2178.
- Van den Brink, P.J., Brown, C.D. and Dubus, I.G. (2006). Using the expert model PERPEST to translate measured and predicted pesticide exposure data into ecological risks. *Ecological Modelling*, 191, 106-117.
- Van den Brink, P.J., Roelsma, J. Van Nes, E.H., Scheffer, M. and Brock, T.C.M. (2002). PERPEST model, a case-based reasoning approach to predict ecological risks of pesticides. *Environmental Toxicology and Chemistry*, 21(11), 2500-2506.
- Van Nes, E.H. and Van den Brink, P.J. (2003). PERPEST Version 1.0, Manual and Technical Description. A model that predicts the ecological risks of pesticides in freshwater ecosystems. Alterra report 787. Wageningen, the Netherlands.

Von Der Ohe, P.C. and Liess, M. (2004). Relative sensitivity distribution of aquatic invertebrates to organic and metal compounds. *Environmental Toxicology and Chemistry*, 23(1), 150-156.

Wogran, J. and Liess, M. (2001). Rank ordering of macroinvertebrate species sensitivity to toxic compounds by comparison with that of *Daphnia magna*. *Bull. Environ. Contam. Toxicol.*, 67, 360-367.

Abbreviations

AI:	Artificial Intelligence
AQUIRE:	Aquatic Toxicity Information Retrieval
BBN:	Bayesian Belief Network
BMWP:	Biological Monitoring Working Party
CatchIS:	Catchment Information System
DOC:	Dissolved Organic Carbon
EC ₅₀ :	This term represents the concentration of a compound where 50% of its maximal effect is observed.
GLM:	Generalised Linear Modelling
GUS:	Groundwater Ubiquity Score
HC _p :	Hazardous concentration for <i>p</i> % of the species.
LC ₅₀ :	This term represents the concentration where there is 50% mortality of the test population.
LOEC:	Lowest Observed Effect Concentration
msPAF:	Multiple Substance – Potentially Affected Fraction of species
NOEC:	No Observed Effect Concentration
PAF:	Potentially Affected Fraction of species
PERPEST:	Prediction of the Ecological Risks of PESTicides
RIVPACS:	River Invertebrate Prediction and Classification System
SPEAR:	SPEcies at Risk
SS:	Suspended Solids
SSD:	Species Sensitivity Distribution
TmoA:	Toxic Mode of Action
US EPA:	United States Environmental Protection Agency
WFD:	Water Framework Directive

APPENDIX B

Impact of flow condition on the occurrence of taxa

Table B1 Taxa for which probability of absence was generally greater in ‘wet’ conditions than ‘dry’, for riffle sites in autumn.

Taxon	Score	Change in Absence Prob (W to D) Prior to Sample Date (Months)				
		1	3	6	12	24
Sialidae	4.3	0.05	0.08	0.10	0.17	0.12
Glossiphoniidae	3.2	-0.02	0.02	0.05	0.18	0.07
Coenagriidae	3.5	0.02	0.03	0.04	0.08	0.07
Psychomyiidae	5.9	0.01	0.03	0.03	0.06	0.07
Dytiscidae	4.7	0.16	0.18	0.19	0.13	0.06
Haliplidae	3.6	0.07	0.10	0.11	0.13	0.05
Planorbidae	3.1	0.00	0.02	0.03	0.15	0.04
Corixidae	3.8	0.04	0.06	0.06	0.08	0.04
Hydrobiidae	4.2	0.05	0.07	0.06	0.07	0.04
Erpobdellidae	3.1	-0.04	-0.03	-0.01	0.07	0.03
Sphaeriidae_Pea_mussels	3.9	-0.03	-0.01	-0.01	0.09	0.03
Asellidae	2.8	-0.09	-0.08	-0.07	0.09	0.03
Lymnaeidae	3.3	0.09	0.13	0.13	0.11	0.01
Physidae	2.4	0.00	0.01	0.01	0.06	0.01
Caenidae	6.5	0.07	0.09	0.12	0.09	0.00
Valvatidae	3.2	-0.02	0.00	0.01	0.06	-0.01
Dendrocoelidae	3.0	-0.03	-0.02	-0.01	0.05	-0.03
Leptoceridae	6.7	0.08	0.11	0.11	0.06	-0.03

Table B2 Taxa for which probability of absence was generally less in ‘wet’ conditions than ‘dry’, for riffle sites in autumn.

Taxon	Score	Change in Absence Prob (W to D) Prior to Sample Date (Months)				
		1	3	6	12	24
Rhyacophilidae	8.2	-0.05	-0.10	-0.12	-0.27	-0.25
Goeridae	8.8	-0.07	-0.06	-0.07	-0.16	-0.24
Simuliidae	5.8	-0.08	-0.12	-0.13	-0.19	-0.19
Ephemerellidae	8.2	0.02	-0.06	-0.06	-0.07	-0.17
Elmidae	6.6	0.01	0.00	0.00	-0.11	-0.17
Sericostomatidae	9.1	0.07	0.07	0.04	-0.13	-0.16
Heptageniidae	9.7	0.03	-0.01	-0.04	-0.21	-0.16
Limnephilidae	6.2	-0.13	-0.12	-0.13	-0.18	-0.15
Planariidae	5.0	-0.04	-0.04	-0.06	-0.11	-0.14
Hydrophilidae	7.4	0.06	0.04	0.03	-0.14	-0.13
Baetidae	5.5	-0.04	-0.07	-0.08	-0.13	-0.12
Leptophlebiidae	8.8	0.01	0.00	-0.01	-0.10	-0.12
Leuctridae	10.0	0.06	0.00	-0.02	-0.19	-0.12
Lepidostomatidae	10.1	0.05	0.03	0.01	-0.13	-0.11
Piscicolidae	5.2	-0.02	-0.01	-0.01	-0.04	-0.11
Hydroptilidae	6.2	0.04	0.02	0.03	-0.02	-0.10
Ancylidae	5.8	0.03	0.02	0.00	-0.04	-0.10
Polycentropodidae	8.1	0.08	0.09	0.10	-0.05	-0.09
Hydropsychidae	6.6	0.02	0.00	0.00	-0.08	-0.09
Gyrinidae	8.2	0.11	0.13	0.13	-0.02	-0.08
Ephemeridae	8.4	0.01	0.03	0.03	0.02	-0.08
Odontoceridae	11.0	-0.02	-0.03	-0.03	-0.07	-0.08
Tipulidae	5.9	0.01	0.00	-0.01	-0.08	-0.07
Gammaridae	4.5	-0.05	-0.04	-0.04	-0.01	-0.06
Perlodidae	10.8	0.05	0.02	0.01	-0.11	-0.06
Nemouridae	9.3	0.01	0.00	-0.01	-0.12	-0.04

Table B3 Taxa for which probability of absence was generally greater in ‘wet’ conditions than ‘dry’, for pool sites in spring.

Taxon	Score	Change in Absence Prob (W to D) Prior to Sample Date (Months)				
		1	3	6	12	24
Coenagriidae	3.5	0.03	0.11	0.13	0.17	0.23
Haliplidae	3.6	0.00	0.08	0.09	0.13	0.16
Sialidae	4.3	0.07	0.07	0.09	0.13	0.15
Dytiscidae	4.7	0.06	0.11	0.10	0.10	0.14
Planorbidae	3.1	-0.02	0.06	0.08	0.13	0.14
Lymnaeidae	3.3	0.01	0.06	0.06	0.10	0.13
Physidae	2.4	0.00	0.02	0.05	0.08	0.13
Notonectidae	3.4	0.00	0.04	0.07	0.10	0.12
Corixidae	3.8	0.01	0.05	0.09	0.10	0.12
Valvatidae	3.2	-0.01	0.05	0.07	0.09	0.12
Glossiphoniidae	3.2	-0.02	0.00	0.00	0.02	0.06
Gerridae	5.2	0.00	0.03	0.02	0.04	0.06
Corophiidae	5.8	0.02	0.03	0.03	0.06	0.02
Hydrophilidae	7.4	0.08	0.08	0.06	0.06	0.00

Table B4 Taxa for which probability of absence was generally less in ‘wet’ conditions than ‘dry’, for pool sites in spring.

Taxon	Score	Change in Absence Prob (W to D) Prior to Sample Date (Months)				
		1	3	6	12	24
Simuliidae	5.8	-0.09	-0.15	-0.14	-0.15	-0.16
Hydroptilidae	6.2	-0.06	-0.05	-0.06	-0.12	-0.16
Leptophlebiidae	8.8	0.00	-0.09	-0.11	-0.15	-0.14
Calopterygidae	6.0	-0.07	-0.11	-0.10	-0.14	-0.14
Hydropsychidae	6.6	-0.03	-0.11	-0.11	-0.14	-0.14
Elmidae	6.6	0.00	-0.06	-0.07	-0.12	-0.12
Baetidae	5.5	-0.08	-0.10	-0.07	-0.13	-0.12
Psychomyiidae	5.9	-0.04	-0.08	-0.06	-0.11	-0.10
Ephemerellidae	8.2	-0.03	-0.05	-0.04	-0.08	-0.10
Ephemeridae	8.4	0.01	-0.05	-0.06	-0.09	-0.09
Polycentropodidae	8.1	-0.03	-0.05	-0.03	-0.08	-0.07
Tipulidae	5.9	-0.01	-0.07	-0.07	-0.10	-0.06
Nemouridae	9.3	-0.02	-0.07	-0.07	-0.08	-0.06
Lepidostomatidae	10.1	0.00	-0.02	-0.03	-0.05	-0.05
Heptageniidae	9.7	-0.02	-0.04	-0.04	-0.06	-0.05
Sericostomatidae	9.1	0.02	-0.01	-0.02	-0.06	-0.05
Leuctridae	10.0	-0.01	-0.02	-0.03	-0.06	-0.05
Ancylidae	5.8	-0.02	-0.04	-0.03	-0.05	-0.02
Gammaridae	4.5	-0.03	-0.07	-0.05	-0.05	0.00
Gyrinidae	8.2	-0.02	-0.05	-0.05	-0.05	-0.03
Limnephilidae	6.2	-0.06	-0.12	-0.09	-0.10	-0.03

Table B5 Taxa for which probability of absence was generally greater in 'wet' conditions than 'dry', for pool sites in autumn.

		Change in Absence Prob (W to D) Prior to Sample Date (Months)				
Taxon	Score	1	3	6	12	24
Coenagriidae	3.5	0.07	0.09	0.10	0.17	0.18
Haliplidae	3.6	0.07	0.09	0.10	0.16	0.13
Corixidae	3.8	0.06	0.06	0.06	0.10	0.12
Sialidae	4.3	0.08	0.09	0.08	0.14	0.11
Planorbidae	3.1	0.00	0.00	0.02	0.09	0.10
Lymnaeidae	3.3	0.06	0.06	0.05	0.09	0.10
Valvatidae	3.2	0.03	0.02	0.03	0.06	0.09
Notonectidae	3.4	0.04	0.04	0.05	0.09	0.08
Physidae	2.4	0.00	0.01	0.01	0.06	0.07
Dytiscidae	4.7	0.10	0.08	0.09	0.10	0.06
Hydrometridae	4.3	0.04	0.04	0.05	0.06	0.05

Table B6 Taxa for which probability of absence was generally less in 'wet' conditions than 'dry', for pool sites in autumn.

		Change in Absence Prob (W to D) Prior to Sample Date (Months)				
Taxon	Score	1	3	6	12	24
Limnephilidae	6.2	0.13	-0.16	-0.17	-0.21	-0.21
Elmidae	6.6	0.03	-0.04	-0.05	-0.14	-0.21
Simuliidae	5.8	0.14	-0.16	-0.16	-0.21	-0.17
Calopterygidae	6.0	0.02	-0.03	-0.02	-0.10	-0.15
Ephemeridae	8.4	0.01	-0.03	-0.03	-0.08	-0.12
Piscicolidae	5.2	0.04	-0.04	-0.06	-0.10	-0.11
Hydropsychidae	6.6	0.03	-0.05	-0.03	-0.09	-0.11
Sericostomatidae	9.1	0.00	-0.01	-0.02	-0.06	-0.10
Polycentropodidae	8.1	0.03	-0.04	-0.02	-0.06	-0.10
Psychomyiidae	5.9	0.03	-0.06	-0.06	-0.08	-0.09
Rhyacophilidae	8.2	0.00	-0.01	-0.02	-0.06	-0.08
Leptoceridae	6.7	0.01	-0.01	-0.01	-0.04	-0.08
Ephemerellidae	8.2	0.01	-0.02	-0.02	-0.04	-0.08
Leptophlebiidae	8.8	0.02	-0.03	-0.03	-0.05	-0.07
Ancylidae	5.8	0.00	-0.03	-0.02	-0.05	-0.07
Tipulidae	5.9	0.04	-0.05	-0.04	-0.07	-0.07
Goeridae	8.8	0.00	-0.01	-0.01	-0.05	-0.07
Nemouridae	9.3	0.02	-0.03	-0.02	-0.05	-0.06
Gyrinidae	8.2	0.00	-0.01	0.00	-0.04	-0.06
Hydrophilidae	7.4	0.01	0.00	0.00	-0.04	-0.05
Heptageniidae	9.7	0.01	-0.02	-0.03	-0.06	-0.05
Baetidae	5.5	0.04	-0.08	-0.08	-0.10	-0.05
Gammaridae	4.5	0.05	-0.06	-0.06	-0.04	-0.04
Planariidae	5.0	0.03	-0.05	-0.07	-0.05	-0.03

Appendix C

Proposed data specification

Introduction

This document contains information related to the definition of proposed specification for an input file for software that handles 'water quality' information, in particular RIVPACS/RICT, RPDS and RPBBN. The document is broken up into the following four sections.

- i) A brief introduction to XML.
- ii) A review of current input files including content and format.
- iii) A discussion of the requirements used to define the proposed specification.
- iv) A description of the proposed format.

The purpose of this document is to obtain comments on and suggestions for the improvement to the proposed specification, from those that will be involved in producing or consuming files based on it. It is only through feedback from interested parties that the specification can be modified to better suit their needs.

Extensible Mark-up Language (XML)

This section provides a brief introduction to XML. Those readers already familiar with XML might wish to skip this section. XML can be defined as follows:

XML is a meta-language (a language for describing other languages) for the design of mark-up languages capable of representing different types of data and documents.

and mark-up languages as:

Mark-up languages provide a method of incorporating additional information (non-textual) into plain text files.

The rest of this section attempts to expand on these definitions and give some background to XML, such as what it does, why it is useful and how it works. This is done in the following stages.

- i) The definition of 'plain text files' and a brief introduction to character encoding, along with the reasons why it is so important in the world of computing.
- ii) The concept of mark-up is discussed, as is the importance of publication of, and conformance to, mark-up language definitions.
- iii) Approaches to the definition of a mark-up language are discussed and XML is introduced as a method of defining a mark-up language.

Plain text files

The contents of a plain text file (created in notepad or vi), in the memory of a computer or on a hard disk consist of a series of electrical or magnetic units that correspond to a one or a zero (a bit), the same as all other files. The bits contained in the file only become letters through the application of a character code, for example ASCII (American Standard Code for Information Interchange), which is a code for converting groups of eight bits (a byte) into letters, for example 01000001 is an 'A' and 01100001

is an 'a'. The ASCII and Unicode¹² character encoding standards are particularly important because they are globally recognised, accepted and implemented, which means that computers worldwide have the ability to interpret files in this format. Therefore a file written as text can be sent anywhere in the world as a stream of bytes and the original text can then be accurately reproduced by the receiving computer. This ability makes plain text files the ideal platform for conveying information that can be understood by humans in an 'electronic' format anywhere in the world.

Mark-up and mark-up languages

A problem with character codes is that they only define characters and therefore have no intrinsic method of conveying style or meaning in the text other than that contained in the words themselves. For example, there is no way to define differences in font for sections of text, as the data is simply a stream of characters¹³.

One method of overcoming this problem is to use 'mark-up tags'; these are characters or words that can be distinguished in some way from the actual text itself and contain information on content of the document. It is important to point out that, even though the tags are part of the file, they are not meant to be part of the content, that is, the text. To illustrate this, consider the following piece of plain text:

```
<italic>Hello World</italic>.
```

In the text '<italic>' and '</italic>' are the tags; in this case, the tags are identified by being enclosed in right-angled brackets and they indicate the start and end of the text with which the information on style is associated. However, the tags are not meant to be included in the content of the document when the text is displayed by the application that interprets the mark-up. The application should identify the tags, understand what they represent, strip them from the text and then apply the information they convey. So in our example the text when displayed by an application that interprets the mark-up should appear as just '*Hello World*'.

The ability to 'understand' what constitutes a tag and what information it conveys are extremely important, otherwise the tag may be displayed as a part of text of the document or identified but ignored. It is essential then to produce a full definition of a mark-up language so that others can understand it and develop software capable of interpreting it. In producing the definition, it is imperative that the descriptions are clear, understandable and as unambiguous as possible, to prevent misinterpretation. The final aspect to the success of a mark-up language is in its publication. The more freely and easily available the definition is, the greater the chance of uptake and use by others.

Definition of a mark-up language and XML

The traditional approach to creating and promoting a mark-up language is to produce a full and complete definition, usually accompanied by the developer's applications which promote the advantages of the language, and then embark on a mass-marketing campaign to get the language recognised and accepted as a *de facto* standard. The problem here is that this is no easy undertaking. The definitions of languages such as postscript, portable document format (pdf) and TeX are lengthy technical documents that describe every aspect from scratch, which is essential to remove ambiguity. As a result, they often require a large investment of time and effort for the creator of the language to produce the language and for the software developer to read and digest it.

¹² Unicode is a more recently produced character coding standard that uses multiple bytes, which allows the encoding of different character sets such as Arabic and Chinese.

¹³ The font used to display text files is defined by the settings in the application, not in the text file itself.

Therefore, whilst this type of approach may be viable for broad and potentially lucrative computer application markets like word processing, where the investment of time and effort on the behalf of the creator and software developer may be justified, it tends not to be in smaller, more specific, application areas.

For example, consider a purchase-order system for a manufacturer. To increase speed and efficiency and cut costs, the company wants to integrate with their suppliers and be able to send purchase order documents containing product name, product code, quantity and price electronically. For example, the values (text) that need to be sent for a purchase order may be represented textually as 'test tube 213 15.00 75'. The problem with sending data in this format is identifying which value corresponds to which attribute. The use of mark-up offers a solution to this problem as it allows information, in this case the names of the attributes, to be embedded in the text, the list of values. An example of a simple mark-up solution might be to place each value on a separate line and put the attribute name at the start, such as 'price 15.00', where 'price' is the tag. As with other mark-up languages files, the software that reads it will strip the tags, their purpose being solely to identify the database field into which the value must be copied.

Ensuring that a mark-up language definition is robust is an extremely involved task. Consider the mark-up solution of using a separate line for each attribute and then putting the name and value just proposed. It would fail immediately because it would produce the line 'product name test tube' in which the tag and text could not be clearly distinguished. This is a rather simplistic example, but the fact that the text and information are stored in the same file and use the same character set will always lead to problems with ambiguity, unless the mark-up solution has been well designed. As a result, in the majority of cases, like that of the manufacturer, the time and effort required to implement a mark-up language as a method of exchanging data makes it impractical. In this scenario, consideration also needs to be given to the role of the supplier, who may have fifty customers all wanting to use their own mark-up languages. Developing and maintaining software to handle all fifty of these languages would be a huge undertaking.

A solution to these problems is to use a mark-up meta-language, a language to describe mark-up languages. A meta-language provides a documented method for describing the tags in your mark-up language and how they should be structured in the document. The following are the two main benefits that using a meta-language in the definition of a mark-up language offer both the creator and consumer.

- i) The meta-language usually contains a predefined method for implementing tags. This removes the need for the creator to include this often complex and technical information in their definition. It also allows the consumer to employ a generic parser, designed to the meta-language specification, which is able to extract text and tags from a document automatically, leaving them free to focus on developing a system that recognises the meaning of tags and responds to them.
- ii) The meta-language provides a structured and clearly defined method for specifying a mark-up language. This should result in the definition being clear, well-structured, easy to understand and less ambiguous. In addition, the creator should benefit from a more clearly defined approach to creating a specification and for the end user, if they are conversant with the meta-language, it should reduce the time required to get to grips with the 'new' mark-up language.

As stated in the initial definition, XML is a meta-language for mark-up languages. The popularity of XML is probably due to its origins. It is based on Standard Generalized

Mark-up Language, the international standard for defining descriptions of the structure of different types of electronic document, a meta-language itself and the basis for probably the most famous mark-up language, the Hyper Text Mark-up Language (HTML), which is the language of the *World Wide Web* (WWW). As a result, the formatting of XML is familiar to millions of web developers worldwide, which has certainly been a factor in its uptake. Another factor is that the body that maintains the XML specification is the World Wide Web Consortium (W3C), a widely known and respected body that is also responsible for maintaining HTML. Whatever the reason is, it is reasonable to suggest that XML is currently the *de facto* method of defining mark-up languages for the exchange of information between organisations.

Summary

Globally accepted standards for character encoding, such as ASCII and Unicode, have paved the way for reliable transmission of text in an electronic format between computers. However, these codes do not include any intrinsic methods of defining information or meta-data about the text itself. 'Mark-up' is a method for including such information and involves embedding clearly differentiable, information carrying sections of text, called tags, into the main body of text. For a mark-up scheme to work successfully it needs to be clearly defined, so that those using the files understand what constitutes a tag and the information it conveys. The definition of a robust mark-up language tends to be both complex and technical, making them difficult to develop and learn. Meta-languages offer a method of easing these problems through standardisation. Standardisation enables the transfer of existing knowledge and tools between mark-up languages produced using a specific meta-language. This reduces the cost and effort associated with creating and using a 'new' mark-up language. XML is currently the *de facto* mark-up meta-language and is being increasingly adopted by business and government organisations to simplify the process of information exchange between bodies whose data are held in different formats.

Current data input files

It is hoped that, ultimately, the input file format defined in this document might become an accepted standard for the exchange of river quality monitoring data. The short-term goals for this format are to be accepted as the standard format for input files for the RIVPACS, RPDS and RPBBN applications. To assess how well this goal has been achieved and to provide some background to the design and development process, this section describes and discusses the content and structure of the input files currently used by these systems.

Content

Although the purpose and outputs of RIVPACS, RPDS and RPBBN systems differ, they all operate in the same domain. As a result, in the places where systems share the same type of input data, the variables they use tend to be the same. The types of input variables can be loosely categorised as environmental, biological or chemical. However, even when the systems share the same input variables, the uses they are put to often differ. In the three systems, variables tend to be input for one of two reasons, either to be used by the model as part of the categorisation/prediction process or for information/comparison purposes, depending on the underlying model¹⁴. Table

¹⁴ Inputs to RIVPACS also include instruction and quality band limit files. The purpose of these input files is to supply the functional parameters to process within the system. Given that the focus for developing this specification was the representation of water quality sample data, these files and the parameters were not included.

gives an outline of the systems and the types of inputs they use, where the inputs are split into the model input and information categories.

Table C2 lists all 173 variables that act as inputs for the three systems and which of the systems use them. It is clear that there is a great deal of similarity in the biological and environmental inputs used by the systems. The chemical and stress variables are mainly used by RPDS for information purposes, the only exceptions being total ammoniacal nitrogen, percentage oxygen saturation, phosphate, pH and total oxidised nitrogen, which are used by the RPBBN system. RIVPACS uses seven variables that the other systems do not. These include five BMWP families¹⁵, velocity category and sample bias. The variables sample ID, site ID, site name, region and watercourse are interesting inclusions because they convey descriptive information rather than data on quantifiable attributes. Strictly speaking, these variables are not necessary because they are not used by the models in making classifications or predictions, nor are they of use in direct comparisons. The information they provide is vital though, helping to identify specific samples and set the data into context.

¹⁵ The reason for the absence of these five taxa from the RPDS and RPBBN was lack of data. These taxa occurred in few or no records and as a result, were excluded from the construction of the models on which these systems are based.

Table C1 Outline of RIVPACS, RPDS and RPBBN systems and their inputs.

System	Type	Model Input	Info. Input	Description
RIVPACS	Classifier	Environmental	Biological	RIVPACS's primary purpose is to predict a macroinvertebrate community based on the environmental characteristics of a site. Biological information is used to compare predicted and actual communities.
RPDS	Classifier	Environmental Biological	Environmental Biological Chemical Stress	RPDS's primary purpose is to diagnose potential pressures affecting a site based on its environmental and macroinvertebrate characteristics. The main output is a report detailing the prediction of these pressures. However, one of the system's key features is the ability to display and compare information from the model and input samples on screen. To achieve this, the system is capable of inputting data on any of the variables that the model contains.
RPBBN	Reasoning	Environmental Biological Chemical	Environmental Biological Chemical	RPBBN's primary purpose is to predict the state of all the variables in the model based on whatever input information is available. This means that any number, type and combination of variables can be used as input variables. The inputs can also be used for information purposes, permitting the comparison of actual and predicted values.

Table C1 List of all input variables and which of the three systems uses them.

(V = RIVPACS, D = RPDS and B = RPBBN)

Environmental & Site Data		Asellidae	V D B	Hydroptilidae	V D B	Nickel (total)	D		
		Corophiidae	V D B	Philopotamidae	V D B	Nitrite	D		
SampleID	V D B	Gammaridae (incl. Crangonyctidae & Niphargidae)	V D B	Psychomyiidae (incl. Ecnomidae)	V D B	Nitrate	D		
Date	D					Oxygen (dissolved)	D		
Season	V D B			Polycentropodidae	V D B	Oxygen (saturation)	D B		
SiteID	D	Siphonuridae	V D B	Hydropsychidae	V D B	Lead (dissolved)	D		
Region	D B	Baetidae	V D B	Phyrganeidae	V D B	Lead (total)	D		
Watercourse	D B	Heptageniidae	V D B	Brachycentridae	V D B	Phosphate	D B		
SiteName	D	Leptophlebiidae	V D B	Lepidostomatidae	V D B	pH value	D B		
X	V D B	Potamanthidae	V D B	Limnephilidae	V D B	Suspended Solids	D		
Y	V D B	Ephemeridae	V D B	Goeridae	V D B	Temperature	D		
Altitude	V D B	Ephemerellidae	V D B	Beraeidae	V D B	TON	D B		
Slope	V D B	Caenidae	V D B	Sericostomatidae	V D B	Zinc (dissolved)	D		
Discharge	V D	Taeniopterygidae	V D B	Odontoceridae	V D B	Zinc (total)	D		
Velocity Category	V	Nemouridae	V D B	Molannidae	V D B	Stresses Sample Data			
Distance from Source	V D	Leuctridae	V D B	Leptoceridae	V D B				
Width	V D	Capniidae	V D B	Tipulidae	V D B	Agri-industry	D		
Depth	V D	Perlodidae	V D B	Simuliidae	V D B	Artificial bank	D		
Alkalinity	V D B	Perlidae	V D B	Chironomidae	V D B	Bank erosion	D		
Hardness	V D	Chloroperlidae	V D B	Biological Sample Details		Channel at site	D		
Calcium (dissolved)	V D	Platycnemididae	V D B			Sample Bias	V	Construction	D
Conductivity	V D	Coenagrionidae	V D B	Number of Families	D	Eroded material	D		
Boulders	V D	Lestidae	V	BMWP score	D	Eutrophication	D		
Pebbles	V D	Calopterygidae	V D B	ASPT	D	Farming	D		
Sand	V D B	Gomphidae	V	RIVPACS GQA class	D	Impoundments	D		
Silt	V D B	Cordulegastridae	V D B	NN GQA class	D	Industrial	D		
SiteType	D B	Aeshnidae	V D B	BOD (mean)	D	Land use	D		
Biological Sample Data (Macroinvertebrate)		Corduliidae	V	Chemical Sample Data		Leachate	D		
		Libellulidae	V D B			Ammonia (mean)	D	Mine	D
		Mesovelidae	V			DO (mean)	D	No flow	D
Planariidae (incl. Dugesidae)	V D B	Hydrometridae	V D B	Alkalinity(2)	D	Oils	D		
Dendrocoelidae	V D B	Gerridae	V D B	Ammoniacal nitrogen (total)	D B	Pesticides	D		
Neritidae	V D B	Nepidae	V D B	Ammoniacal nitrogen (non-ionised)	D	Reclamation	D		
Viviparidae	V D B	Naucoridae	V D B	BOD	D	Run-off	D		
Valvatidae	V D B	Aphelocheiridae	V D B	Calcium (total)	D	Salinity	D		
Hydrobiidae (incl. Bithyniidae)	V D B	Notonectidae	V D B	Cadmium (dissolved)	D	Sediment	D		
Physidae	V D B	Pleidae	V	Cadmium (total)	D	Sewage treatment works	D		
Physidae	V D B	Corixidae	V D B	Chloride	D	Waste	D		
Lymnaeidae	V D B	Halplidae	V D B	Chromium (dissolved)	D	Water treatment works	D		
Planorbidae	V D B	Hygrobiidae	V	Chromium (total)	D	Other	D		
Ancylidae (incl. Acroloxidae)	V D B	Dytiscidae (incl. Noteridae)	V D B	Copper (dissolved)	D	No perceived stress	D		
Unionidae	V D B	Gyrinidae	V D B	Copper (total)	D				
Sphaeriidae	V D B	Hydrophilidae (incl. Hydraenidae)	V D B	Iron (dissolved)	D				
Oligochaeta	V D B	Scirtidae	V D B	Iron (total)	D				
Piscicolidae	V D B	Dryopidae	V D B	Magnesium (dissolved)	D				
Glossiphoniidae	V D B	Elmidae	V D B	Magnesium (total)	D				
Hirudinidae	V D B	Sialidae	V D B	Nickel (dissolved)	D				
Erpobdellidae	V D B	Rhyacophilidae (incl. Glossosomatidae)	V D B						
Astacidae	V D B								

Format

RIVPACS accepts inputs in two different types of file formats, text and dbf database files. Optionally, the system requires up to three different input files containing environmental, biological and sample error data.

The dbf files must conform to a predefined data table format, which dictates field names, data types and field widths. The text input files use a fixed format, that is a line defines an entity and within that line, variables have predefined locations, for example the characters 1-20 are allocated to the variables sample code.

In the environmental and sample error files, variables are assigned specific locations within each line. However, in the biological data file, each line contains multiple taxonomic codes, which identify the species in the sample and multiple lines can be associated with one sample, by the use of an initial identifier. These may only appear to be slight variations in format but they permit the number of variables associated with one sample to vary and addition, removal or modification to the biological variables used by the system without requiring a change to the file format. This flexibility comes at the cost of increased sophistication in the software that reads the files because the system has to be able to identify the sample and taxonomic code before copying the value to the appropriate place. It may already be obvious that the biological file employs a mark-up system, the taxonomic codes being the tags identifying the appropriate variable in the software to which the value indicating presence or the supplied abundance value must be assigned.

RPDS and RPBBN systems use only one type of input file, in the text format in which each line corresponds to an individual record and values in the record are delimited by commas. The order in which the variables must be placed in the line is predefined. Using a system of delimitation to define variables is more flexible than fixed positioning, in that it allows variation in the length of fields. The drawbacks are that it requires every character in the line to be searched to find the delimiters and problems can occur when the delimiting character appears as part of a value. However, even though the use of delimitation adds some flexibility to the file format, the strict ordering of variables in each line makes the file format, as a whole, rigid.

Discussion

An overriding feature of data used by the three systems, shown in Table C2, is the amount of variation in the 173 variables. The following is a list of some ways in which they differ.

- i) In types of data, if the data is continuous and discrete, ordinal or nominal.
- ii) In the ranges of values, their respective maximum and minimum values or states of a discrete variable.
- iii) In the types of values, whether values are a scientific measure, a count, an estimate or a calculated value, like mean or percentile.
- iv) In the methods by which they are defined, that is, does the variable have a 'definitive' definition like chemical compounds or is the definition that of a particular authority, like macroinvertebrate families.
- v) In the method by which they are measured, for example milligrams per litre, micrograms per litre, centimetres metres, numbers alive, colour, and so on.
- vi) In the methods used for identification and quantification, a few examples of methods are: filtration, gas chromatography, kick sample, altimeter or GIS query.

It is worth noting that some of these variations may occur in between variables of the 'same type'. For example, for 'distance from source' for a river, the definitions of the location of the source, the method of measurement and units used may all vary, yet despite these variations, this group of values may all be considered to be the 'same' variable.

In addition to the potential for variation in the variables themselves, there is potential variation in the structure of data and definition of what constitutes a complete and valid record. In the case of the three systems, the most notable variation is in the composition of variables that make up a record. The combination of a diverse range of variables from different sources and types of sampling programmes raises the question of how closely matched the samples must be, in terms of date and location, for them to be considered as part of one unified record. The definition of 'matching' criteria is ultimately discretionary and thus also a source of variation.

All the potential variation in water quality sample data is an unavoidable consequence of the diversity of the data. However, as long as variations in the data are properly documented, they should not present a problem. If they are not completely documented, they can be a potential source of ambiguity, misinterpretation and error. RIVPACS, RPDS and RPBBN have definitions for the variables they use; the problem is that this information is contained in the file specifications rather than the data files themselves. The name, type, measurement and any other data associated with a variable reside in the specification and are conferred to a value by virtue of its location in a data file. The main problems with associating information with a value in this way are that the file is practically unusable without a copy of the correct file specification, and errors in interpreting the specification or use of the wrong specification may lead to the information for one variable being incorrectly ascribed to the value of another.

Requirements for a new format

The existing file formats used by RIVPACS, RPDS and RPBBN all have one objective and that is to enable new data to be input into their respective systems. They fulfil this aim efficiently with only the data that is needed by the system included in the files. For a new file format, it is anticipated that the objectives would differ and it would have the capacity to permit information on additional, perhaps as yet undefined, variables to be included. The purpose of this would be both to 'future proof' the file format against further development of RIVPACS, RPDS and RPBBN systems and permit the format to be used by other developers who might use different suites of variables.

The remit of this file format requires some discussion. However, looking at the problem in general terms, the range of potential formats lies between a tightly defined or prescribed format, in which each variable is named specifically in the format, and a completely generic solution, in which variables can be specified on an *ad hoc* basis in the data itself.

The file format defined in this document is a generic solution, and below are the three main reasons for that decision.

1. In the author's view, a generic design offers a longer-term solution, should simplify the task of sharing data and provides an opportunity to increase the amount of information on variables in the data, thus decreasing the likelihood of incorrect properties being inferred on values.
2. A generic solution would need to be developed from scratch and therefore requires more time and effort to develop. In contrast, a prescribed format would be relatively straightforward to develop as it would involve combining

existing file formats and translating them into XML. Therefore, if the decision to pursue a more generic format were taken, at least some of the work required would already have been undertaken.

3. The development of a generic file format would facilitate a discussion of requirements and specifics of the new file format by demonstrating a possible generic solution. As mentioned previously, a prescribed format would basically be a reproduction of current formats and so has, to a large extent, already been defined in the section on current file formats. In producing a generic solution, the hope is to demonstrate an alternative solution and introduce some of the 'generic concepts', to provide a common frame of reference for further discussion.

Costs and benefits of adopting a generic format

The costs and benefits of a generic solution can be reduced to a discussion on flexibility and extensibility versus speed, storage requirements and the ease of processing. At the simplest level, as solutions become more generic there is an increase in the amount and complexity of information stored in the file. The benefits of this are the ability to store and clearly define a wide range of variables, which is a product of the greater capacity to describe them. The costs are increased demands for storing, reading and understanding information in the file. Thus the generic solution proposed in this document has the advantage of enabling a variety of variables to be included and described in the file format and, as a result, will allow developers to modify their software without requiring changes to the input file format. The drawback will be the need for more sophisticated 'parsing' algorithms to deal with the potential for variation in files. To clarify this last point, in the current software values are easy to extract and process because of the rigid formatting. In a generic solution the variables and their associated attributes are defined in the file itself and may differ from file to file, so software that uses the generic file format not only has to be able to read the information associated with the variables but also check whether they match the data requirements of the system. In terms of demands on hardware, all the extra information in the files inevitably leads to increases in storage and processing requirements.

Requirements for a generic format

For a generic solution to be successful it must be able to encapsulate all potential variation in the data that it has to hold. The extent of potential variation in the water quality sample data files has already been largely addressed. Six types of variation - type of data, values, range of values, variable definition, measurement and method of identification - were highlighted and any proposed generic file format should at least be able to handle these types. In addition to the variation in individual variables, the specification must also be able to handle variation in the time and type of samples that can be combined into a single record.

Proposed XML format

Figure C1 shows the entities and relationships that would be defined in the proposed generic solution. In the diagram, the entities have been broken down into three groups: core entities, variable information entities and sample information entities.

Core entities consist of the 'record', 'result set' and 'result' entities. The 'record' entity is the equivalent of the complete 'samples' that are used by the systems as the basis of analysis, that is the set of biological, environmental and chemical variables fed into the system as a single entity. As discussed already, 'samples' used by the systems are

often a combination of data from different sources. To enable the details of these different ‘sample’ sources to be recorded in the files, the ‘result set’ sub-component was introduced. The ‘result set’ entity’s primary purpose is to act as a container for the result entities, which hold individual sample values. In addition, it contains a link to the sample information group of entities, in which specific details of the sample(s) or data query from which the data was obtained can be stored. At the lowest level of this hierarchy is the ‘result’ entity, which contains the individual sample values and has a link to the variable information entities, which hold information about the value itself, what it represents and how it was obtained.

Variable information entities comprise six entities, each designed to encapsulate the six types of data variation identified at the start of the section.

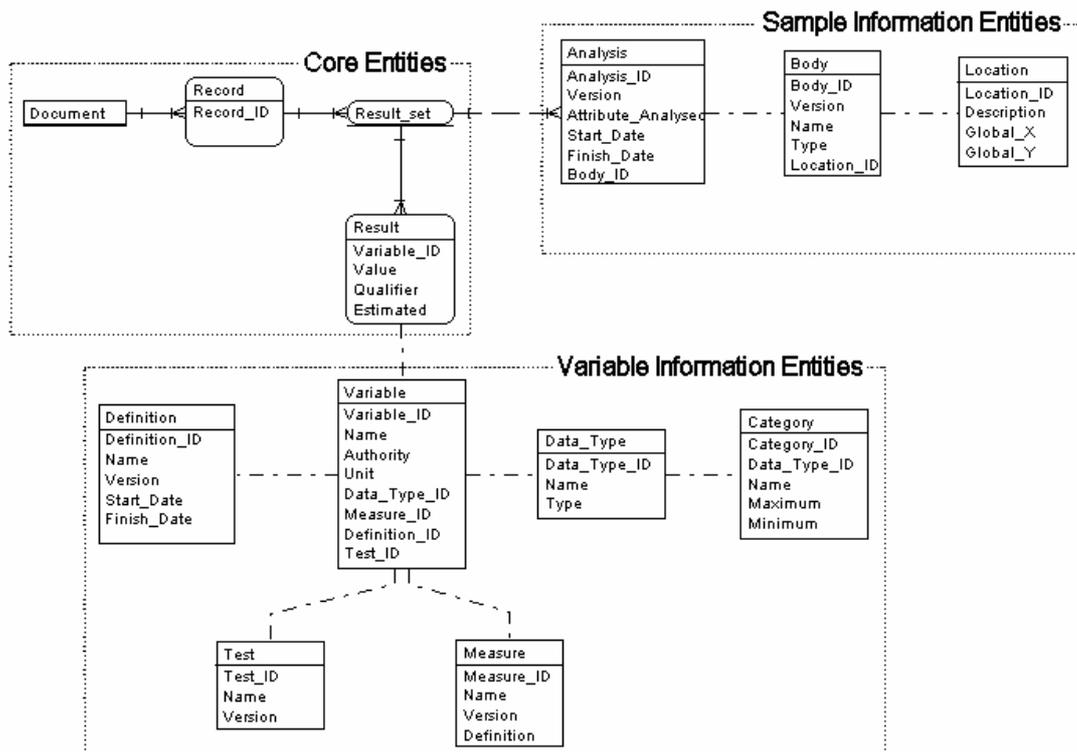


Figure C1 Outline of proposed XML specification for river quality data files.

Sample information entities are designed to handle information on the source of the value(s) recorded. The term ‘sample’ is perhaps not ideal since the source may be a database query. However, it does help encapsulate the concept that a particular set of values were obtained in the same process. The term sample is avoided in the specification itself in favour of ‘analysis’. Whether this term needs to be changed for another, perhaps more meaningful term, is open to discussion. In any case, the analysis entity attempts to capture the information about the process to analyse a particular aspect of an object, whether it be a river or a database. This is linked to the ‘body’ entity and is designed to hold information on the object under analysis, which in turn has a location entity associated with it. These objects appear to be generic enough to capture two different sources of data and in the process make the solution more straightforward. They also appear capable of storing additional information about particular parameters which may vary between data sources. This information plays a vital role in deciding whether the result sets that comprise a record entity are sufficiently similar. However, in striving to produce a generic solution, important source-specific information may have been missed out. So another issue to consider is whether the analysis entity should be split into different entities for different sources of data.

In the proposed generic solution, shown in Figure C1, the information entities are linked to their associated core entity through a key. The association of information via a key was used because it allows the information to be re-used rather than embedding it in the core entity. This decentralisation of information has its drawbacks because the information required to define an element fully is no longer stored in one place. However, the potential savings in storage and processing from the reduction in repetition was judged to be more beneficial.

The apparent complexity of this specification may be an issue, for there is no doubt that it is more sophisticated than the current file formats. However, implementation of the specification does not necessarily have to include all entities or information they contain. In reality, entities such as data type, category and test may be excluded more often than not and in some cases all the information entities may be left out except perhaps the variable entity. However in many situations, especially when sharing data, omitting the information entities may lead to ambiguity in what values represent. There would be circumstances, for example creating files for input into a specific application, where the ability to exclude unnecessary information is justifiable. The range of tools and supporting technologies available to developers and users of XML should ease the complexity of implementing the new specification.

Discussion

The specification outlined in this document is a first attempt at developing a general XML based file format for the distribution of river sample data. There is plenty of scope for refinement. The following issues concern the specification and its design:

- Are the design requirements sufficient or is something missing?
- Is the specification too generic or too prescriptive?
- Is the design too complex or too difficult to implement?
- Are any elements missing or the structuring of the entities or relationships wrong?

On a more general note, are there any issues that have not been addressed in this document or are there areas that need more research or explanation? For example the role the electronic-Government Interoperability Framework (e-GIF) might play in the development of this specification.

The more feedback that can be elicited, the better the final specification should be.

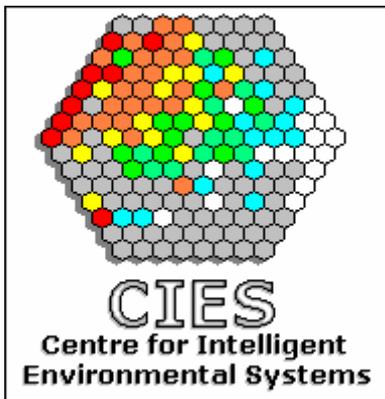
Appendix D

MIR-max User Guide

MIR-max

Data Analysis and Visualisation System
Version 0.2 (prototype)

User Guide



Mark A. O'Connor & William J. Walley
Centre for Intelligent Environmental Systems
School of Computing
Staffordshire University

February 2002 (Draft)

MIR-max data analysis and visualisation system

Version 0.1 (prototype)

User Guide

Contents

1.	Introduction	
1.1.	Background	D-4
1.2.	Pattern recognition and clustering	D-4
1.3.	Data visualisation	D-7
2.	Getting Started	
2.1.	Installing MIR-max	D-9
2.2.	Running MIR-max	D-9
3.	MIR-max Interface	
3.1.	Indicators	D-11
3.2.	Base map	D-13
3.3.	MI-max clustering	D-14
3.4.	R-max ordering	D-14
3.5.	Main data viewer	D-15
3.5.1.	Template panel	D-17
3.5.2.	Report panel	D-18
3.5.3.	Archive samples panel	D-18
3.5.4.	Input samples panel	D-18
3.5.5.	Other functions	D-19
4.	File Formats	
4.1.	General	D-19
4.2.	Data files	D-19
4.3.	Indicator files	D-19
4.4.	Base map files	D-20
4.5.	Clustered data files	D-20
4.6.	Cluster files	D-20
4.7.	Mapping files	D-20
4.8.	Configuration files	D-20
5.	Acknowledgements and contact details	D-21

1. Introduction

Background

MIR-max (Mutual Information and Regression maximisation) is a method for clustering data and arranging the clusters in an output space for user-friendly data analysis and visualisation. MIR-max was originally used as part of a research project on river pollution in England and Wales, which resulted in the development of a software package, RPDS (River Pollution Diagnostic System, later known as the River Pressure Diagnostic System). The pattern recognition and data visualisation techniques used by RPDS were not specific to river water pollution. It was clear that RPDS could be developed into a more generic program that would allow users to take advantage of the same techniques for their own data: MIR-max is the result. This document describes the free prototype version (0.2), which has a number of restrictions and is intended for evaluation purposes only; a fully functional version (1.0) will be available shortly, and it is hoped that a more complete pattern recognition and visualisation package (including other techniques alongside MIR-max) will be produced in the near future.

Pattern recognition and clustering

Experts use two complementary mental processes when interpreting data: plausible reasoning based upon scientific knowledge, and pattern recognition based upon experience of past cases. MIR-max simulates a pattern recognition approach to data interpretation. Essentially, a pattern recognition process separates a set of samples into a number of 'clusters' such that similar samples are clustered together. Each cluster can then be regarded as being representative of a particular 'class' of sample. When a new sample is presented to the system, it attempts to find the best-matching cluster. It then bases its analysis of the new sample on previous knowledge of other samples in the same cluster. The following example illustrates the ideas of pattern recognition and clustering, and is based on biological monitoring of river water quality – the application for which MIR-max was first developed. Consider river water samples containing abundance level data for 12 creatures (taxa). Such a sample can be represented by a bar chart (Figure D1.1) to provide a visual 'profile' or 'pattern'.

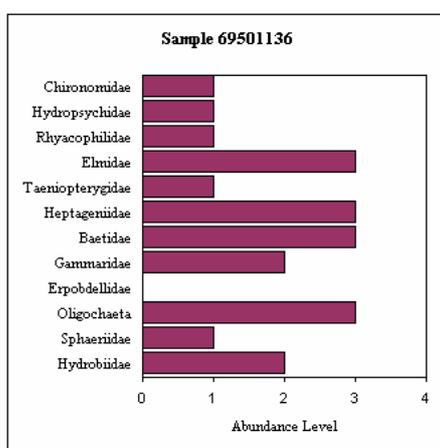


Figure D1.1 Biological sample represented as a bar chart.

Assume that 18 such samples are collected from various sites. Each of the samples can be represented as bar charts for easier visualisation (Figure D1.2); the bars represent the same taxa as in Figure D1.1, although all labels have been removed so as to focus on the pattern.

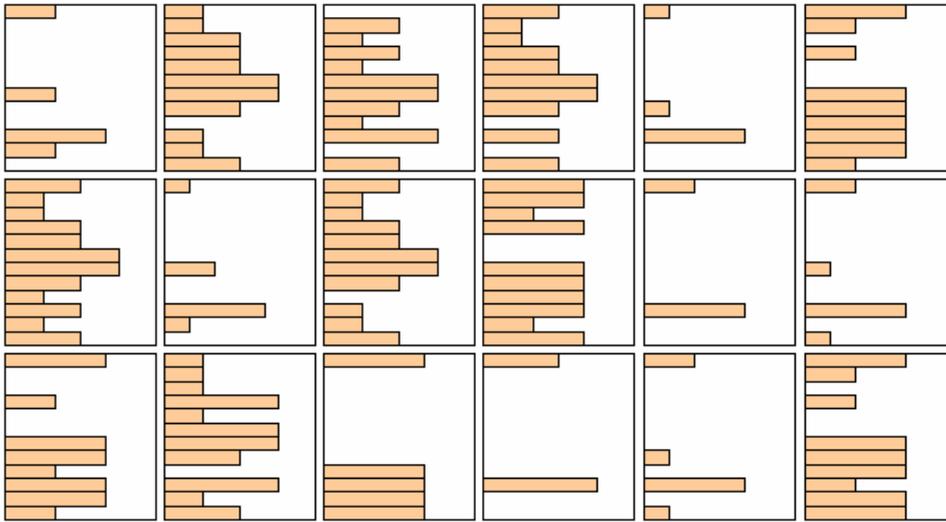


Figure D1.2 Eighteen biological samples represented as bar charts.

The sites can now be clustered according to which represent the most similar patterns. For this example, three clusters will be used. The patterns from Figure D1.2 are all compared, and are grouped into three clusters (Figure D1.3).

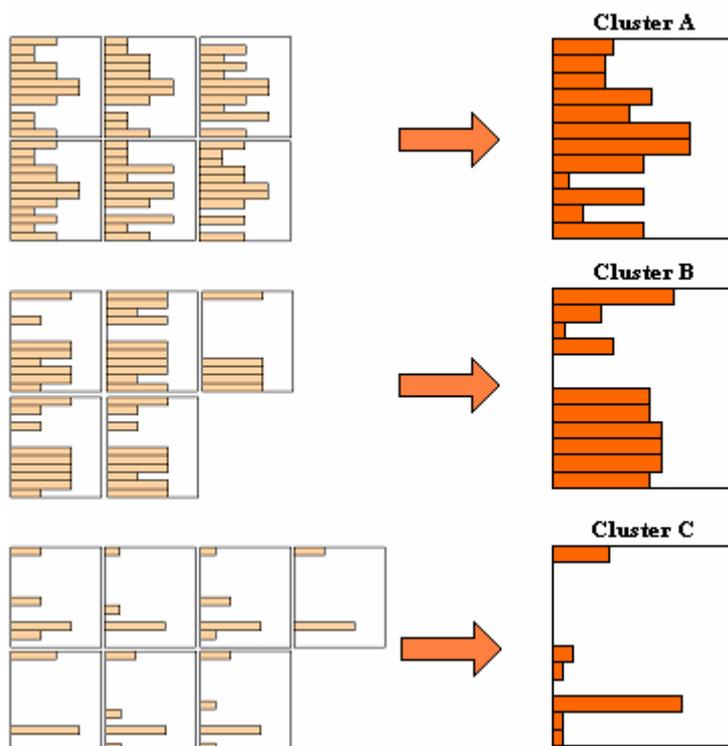


Figure D1.3 Samples grouped into three clusters (A, B, and C).

The algorithm used to perform the clustering – that is, the method by which sample patterns are compared and grouped together – is MI-max (Mutual Information maximisation).

The clusters produced are labelled in Figure D1.3 as A, B, and C, and for each cluster a pattern is produced representing the average of samples contained within that cluster. This pattern can then be regarded as an exemplar for all samples of type A, B, or C; it is referred to as the cluster ‘template’. When a new sample is collected, it can be compared with the templates and assigned to one of the clusters A, B, or C (Figure D1.4).

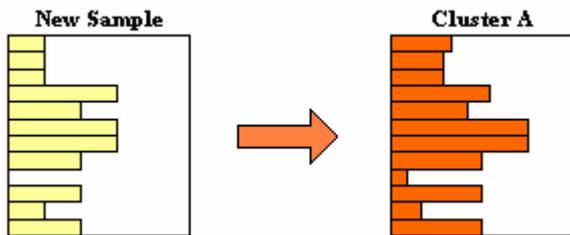


Figure D1.4 New sample matched with template for cluster A.

Now that the new sample has been classified as belonging to a particular cluster; previous knowledge of the samples that were initially used to construct the cluster can be used to draw inferences about the state of health of the river from which the new sample was taken. In this example, cluster A was formed from six of the original 18 samples. Although only the 12 taxa were used in cluster formation, other data may be available for some or all of the original six samples. So, the cluster template can be extended to provide this additional knowledge (Figure D1.5; the values of the additional variables, shown as blue bars, have been suitably rescaled so as to be directly comparable).

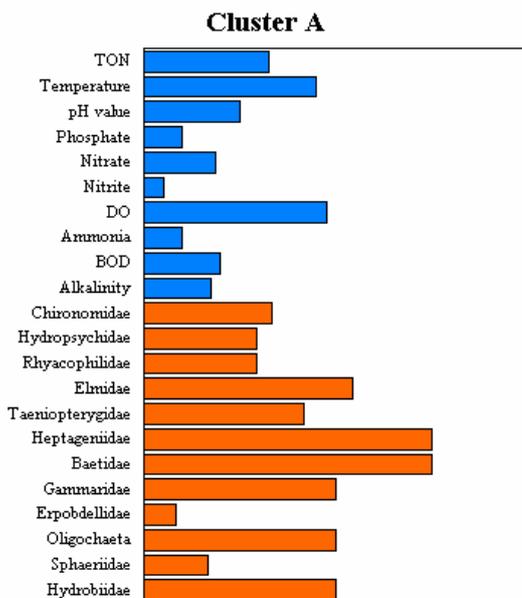


Figure D1.5 Extended cluster template.

Now, although only the biological data is known for the new input sample, predictions of the likely values for the chemical data associated with the new sample are possible using the exemplar values for cluster A, based on the known values for the samples that originally formed the cluster.

Data visualisation

MIR-max uses a cluster 'map' to visualise the clusters. The clusters are each represented by a coloured circle, and are arranged on the screen so that the most similar clusters are located close together, whilst those that represent very different conditions are located far apart. The following example (again based on biological river water quality) illustrates the construction of a cluster map.

Consider 12 clusters, represented (as in Section 1.2. above) by templates and labelled A – L (Figure D1.6).

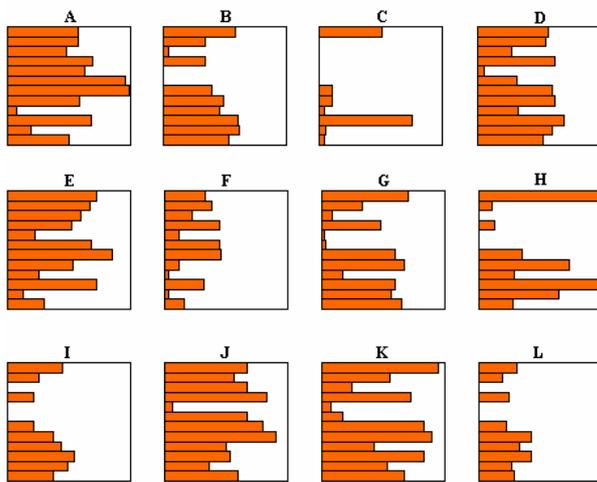


Figure D1.6 Cluster templates.

The 12 clusters are arranged in an output space consisting of 19 possible locations that form a hexagonal 'map', so that the most similar clusters are positioned close together (Figure D1.7). The 'ordering' process is performed by an algorithm R-max (Regression maximisation). This algorithm together with the clustering algorithm MI-max forms the basis of MIR-max data analysis.

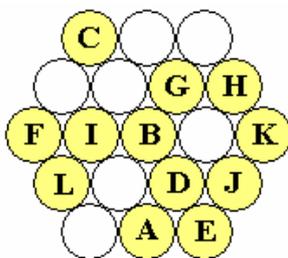


Figure D1.7 Clusters arranged in a hexagonal output space.

The resultant map (Figure D1.7) gives a general impression of the similarities between clusters; for example, clusters I and L, represented by similar templates, are positioned close together on the map, but are relatively far away from cluster J, which is represented by a very different template. The degree to which the map truly represents

the relationships (the ‘distances’) between the clusters can be enhanced by using a greater number of possible output locations. However, this also affects the utility of the map: depending on the user’s intention, a more or less ‘tightly packed’ map may be desirable.

Now that the clusters have been arranged, different aspects of the cluster templates can be viewed using ‘feature maps’ (Figure D1.8).

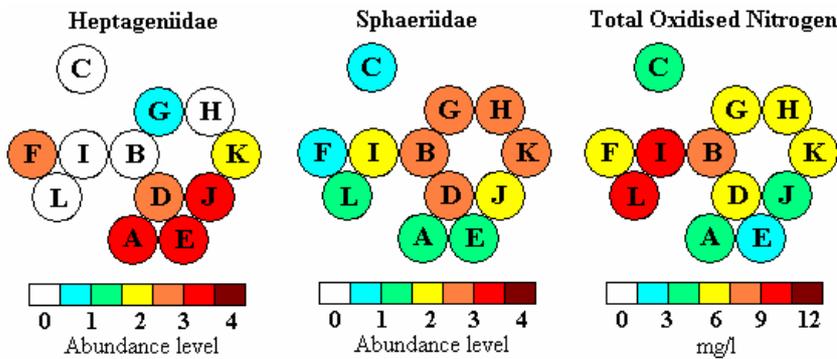


Figure D1.8 Feature maps.

The different features may be those used in creating the clusters (in this case, the abundance levels of the 12 taxa) or any from the ‘extended template’ (for example, in Figure D1.8, total oxidised nitrogen (TON) is pictured). The maps are colour-coded and can be directly compared; for example, in clusters I and L, where abundance levels of the mayfly Heptageniidae are low, the TON is relatively high.

With a small number of clusters, features and input samples (as in the above example) the visual power of the feature maps is perhaps not immediately evident. In RPDS, 250 clusters are used, formed from 6,039 samples. The resultant maps are consequently more informative. In Figure D1.9, for example, the inverse relationship between abundance of Heptageniidae and TON is clear.

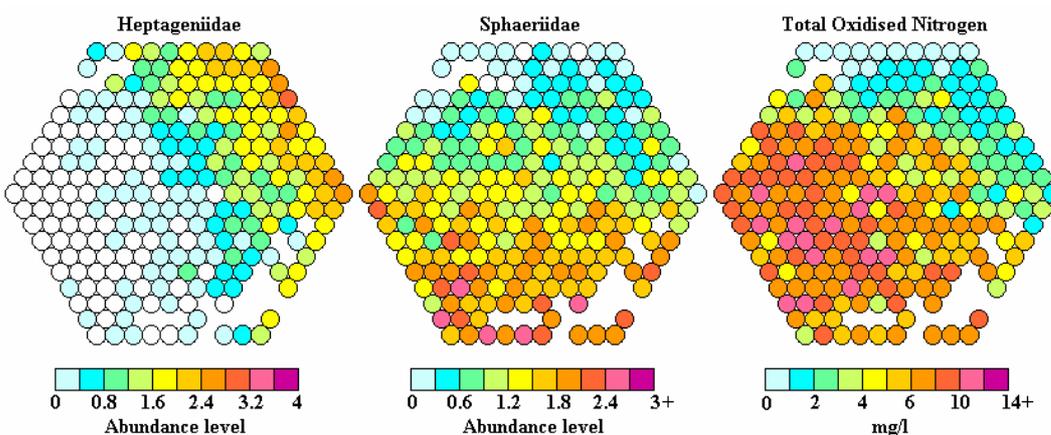


Figure D1.9 RPDS feature maps.

So, there are two main methods that form the basis of data visualisation in MIR-max:

- Feature maps, which show information for a chosen feature across all the clusters.
- Bar charts (referred to as ‘templates’), which show information for a range of features within a chosen cluster.

2. Getting Started

2.1. Installing MIR-max

MIR-max is designed for use with 32-bit Windows platforms (Windows 95, 98, NT, and 2000). To install MIR-max, run the setup.exe program provided on the CD or floppy disk, and follow the on-screen instructions.

2.2. Running MIR-max

Once installation is complete, you can run the main program, MIR-max.exe. You will initially be presented with a disclaimer (because this version of the software, 0.1, is a prototype). Press 'OK' to accept the conditions and start MIR-max.

The 'File and Number Formats' screen (Figure D2.1) will be displayed. MIR-max stores information in a number of files, which should all be of the same format. The usual (and recommended) option is to use comma-delimited files; that is, each set of data is stored in a file as a series of plain text lines, with each part of any line of data separated from the others by a comma (the 'delimiter').

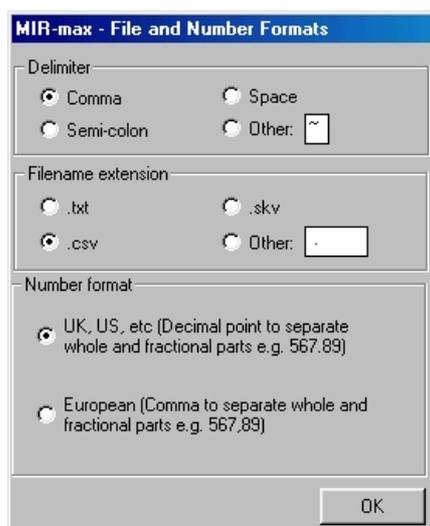


Figure D2.1 Options for file and number formats.

For example, a comma-delimited file storing data about regular polyhedra may appear, when viewed by a standard text editor, as:

```
Name,Faces,Vertices,Edges  
Tetrahedron,4,4,6  
Cube,6,8,12  
Octahedron,8,6,12  
Dodecahedron,12,20,30  
Icosahedron,20,12,30
```

You can choose any single character as the 'delimiter'; ensure that the chosen delimiter will not appear in any of the data items you wish to store (for example if you are storing names like 'Mark O'Connor', do not choose a space or an apostrophe as the delimiter).

Because the files are stored as plain text, they can be checked and edited manually using a standard text editor such as NotePad. It is often easier, however, to use a spreadsheet. Microsoft Excel automatically recognises comma-delimited files with the filename extension .csv for viewing and editing, and so this is the recommended option. The polyhedra data above would appear in Excel as shown in Figure D2.2.

	A	B	C	D	E
1	Name	Faces	Vertices	Edges	
2	Tetrahedron	4	4	6	
3	Cube	6	8	12	
4	Octahedron	8	6	12	
5	Dodecahedron	12	20	30	
6	Icosahedron	20	12	30	
7					
8					

Figure D2.2 File with .csv extension viewed in Excel.

Other standard filename extensions include .txt for plain text files and .skv for semicolon-delimited files, or you can specify any other three-letter extension if you wish.

You can also choose to store numbers using a decimal point or a comma to separate the whole and fractional parts. (If you choose to use a comma, ensure that the files are not stored as comma-delimited – in countries where the convention is to use commas rather than points, the semicolon-delimited format .skv is commonly used instead of the comma-delimited .csv).

When you have selected the required file and number formats, click 'OK'. The MIR-max main menu (Figure D2.3) will be displayed.

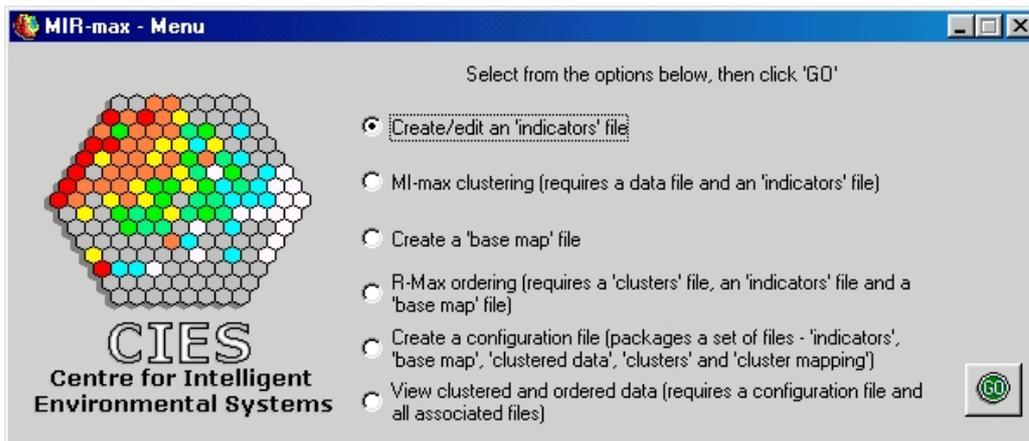


Figure D2.3 MIR-max main menu.

There are six options, as shown in Figure D2.3. Each is described in Section 3 below.

3. MIR-max Interface

3.1. Indicators

In order to analyse data samples, MIR-Max needs to know about the variables recorded in each sample. The variables are referred to as 'indicators', and MIR-Max will store knowledge of these in an 'indicators file'. An indicators file can be created from the main data file, which stores a list of samples together with the associated values for each indicator. The format for these data files is given in Section 4.2.

To create an indicators file, choose the option *Create/edit an indicators file* from the main menu (Figure D2.3). You will be presented with a further choice – to edit an existing file, or create a new file from a data file. Choose the option *Use delimited data file to construct indicators file*. A standard Windows dialogue box then enables you to select the appropriate data file. You will be asked to input the number of discrete categories for MIR-Max to use in data analysis – MIR-max is based on discrete data, so any continuous data needs to be converted into discrete categories first. You should not exceed 12 categories. After inputting the number of categories, MIR-max will read the chosen file, and display the indicator definition screen (Figure D3.1)

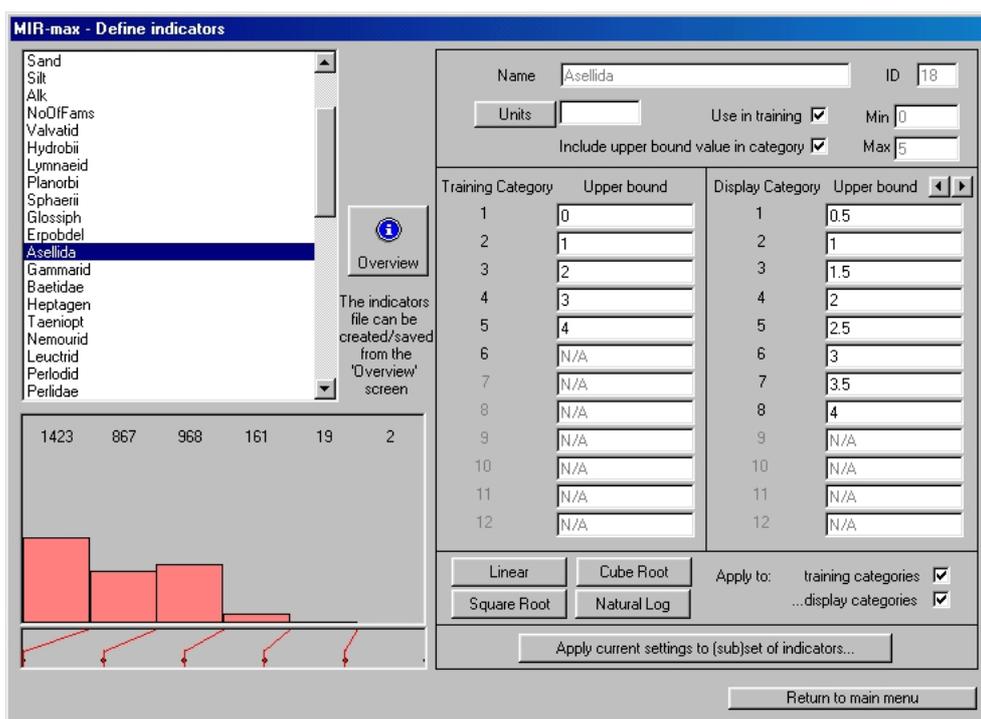


Figure D3.1 Indicator definition screen.

Each indicator is listed by name (taken from the first line of the data file) in the scrollable area at the top left of the screen. To view and edit the information for an indicator, select by clicking the indicator name in the list. The indicator name, a unique ID number (allocated by MIR-max), and the minimum and maximum values recorded for the indicator in the data file will all be displayed at the top right of the screen. These cannot be changed. Also at the top right is a tick-box *Use in training* - by default, all indicators in the data file are used in the training of MIR-max models (unless there is data missing); uncheck the box if you do not wish the selected indicator to be used in the training process. A small button *Units* is also available – click this button to input the type of units used when recording the data.

Two lists at the right of the screen show the 'category bounds'. These figures determine the boundaries for converting 'raw' data into discrete categories. The figure shown is the upper bound for each category – all data with a recorded value up to the figure shown for each category (and greater than the previous category bound) will be regarded by MIR-max as belonging to that category. The number defining the bound can be included or excluded from the category using the check-box *Include upper bound value in category* at the top right of the screen. There are two sets of bounds, one for training categories and one for display categories. The number of training categories is fixed (as that number previously defined by the user), but a different set of categories may be used for display purposes. To increase or decrease the number of display categories, use the left (decrease) and right (increase) arrows at the top of the list.

By default, the category bounds are set to equal intervals between the minimum and maximum recorded values. A variety of alternative methods can be chosen (at the bottom right of the screen) to define the category bounds, such as using a division based on the square root or natural logarithm of the data. Another alternative is to set the bounds manually, by typing in the required values and selecting the *Set category bounds manually* option. If you decide to change the bounds again, click the *Update* button to confirm the changes.

At the bottom left of the screen is a graph showing the number of data samples from the input file that belong to each of the training categories. Below the graph, the value of the selected indicator for each individual data sample is displayed as a red dot on a panel (scaled between the minimum and maximum values), and red lines show where the bounds occur. (These graphs are not available when editing an existing indicators file, only when creating an indicators file from a data file.)

Often, you may want to define the same values for a large number of indicators. To avoid having to edit each indicator individually, first edit just one of them and click *Apply current settings to (sub)set of indicators...* for the 'multi-update' screen (Figure D3.2).

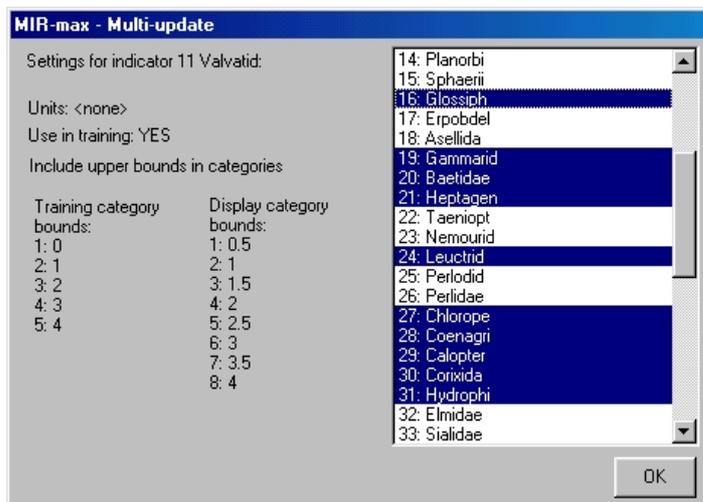


Figure D3.2 Indicator multi-update screen.

The settings for the indicator you have just edited appear at the left of the screen, and a list of all other indicators appears at the right. Select and highlight all the indicators you wish to apply these settings to. (Clicking the indicator name in the list selects the indicator; to select more than one indicator, hold down the keyboard *Ctrl* key when making each new selection.) Click *OK* to close the multi-update screen and return to the indicator definition screen.

When you have completed editing for each of the indicators, click *Overview* to check the values. The indicator overview screen (Figure D3.3) shows a summary of the boundary values for each indicator.

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
Stype	<= 1	<= 2	<= 3	<= 4	<= 5	> 5
GQA	Not used in training					
Width	Not used in training					
Depth	Not used in training					
Boulders	Not used in training					
Pebbles	Not used in training					
Sand	Not used in training					
Silt	Not used in training					
Alk	Not used in training					
NoOfams	<= 7	<= 14	<= 21	<= 28	<= 35	> 35
Valvalid	<= 0	<= 1	<= 2	<= 3	<= 4	> 4
Hydrobi	<= 0	<= 1	<= 2	<= 3	<= 4	> 4
Lymnaeid	<= 0	<= 1	<= 2	<= 3	<= 4	> 4
Planorb	<= 0	<= 1	<= 2	<= 3	<= 4	> 4
Sphaeri	<= 0	<= 1	<= 2	<= 3	<= 4	> 4
Glossiph	<= 0	<= 1	<= 2	<= 3	<= 4	> 4
Eropedel	<= 0	<= 1	<= 2	<= 3	<= 4	> 4
Asellida	<= 0	<= 1	<= 2	<= 3	<= 4	> 4
Gammarid	<= 0	<= 1	<= 2	<= 3	<= 4	> 4

Figure D3.3 Indicator overview screen.

The boundary values for training or display can be viewed by clicking the appropriate tab at the top of the screen. If you wish to accept the values shown, click *Create file* to save the information as an indicators file: a standard Windows dialogue box enables you to choose a file name and location. If you wish to change some of the information, click *Edit individual indicators* to continue editing.

Click *Return to main menu* to exit the indicator definition screen (or indicator overview screen) – you will be prompted to confirm this choice, as a reminder to save any changes.

3.2. Base map

Part of MIR-max data visualisation is based on 'feature maps' – an organisation of data clusters in a fixed output space. To define this output space (the 'base map'), choose the option *Create a 'base map' file* from the main menu (Figure D2.1). The base map definition screen (Figure D3.4) will appear. Select the topology and shape of the required output map, and enter the dimensions in the adjacent boxes. Click *Preview* to see what the chosen map looks like and how many 'nodes' (or separate output locations) there are.

Figure D3.4 Base map definition screen.

When you have chosen the map you would like to use, click *Create file* to store the map as a file. A standard Windows dialogue box allows you to choose the file name and location.

Click *Return to main menu* to exit the base map definition screen.

3.3. MI-max clustering

To analyse the data and create clusters using MI-Max, select the option *MI-Max clustering* from the main menu (Figure D2.1). The MI-Max clustering screen appears (Figure D3.5). Enter the number of clusters you would like to use, then click *Go*.

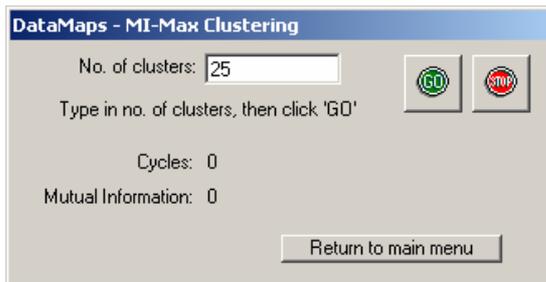


Figure D3.5 MI-max clustering screen.

You will then be asked to select the indicators file and data file, using standard Windows dialogue boxes. The files are read by MIR-max; this initialisation may take a few seconds, after which you will be asked to input a 'threshold' value that determines how many training cycles MI-max should perform. MI-max clustering may take a long time, especially if there are a lot of indicators, clusters or data samples. The MI-max screen keeps the user informed of how many training cycles have passed, and the current mutual information value (which it is attempting to maximise). At any stage in the training, the user can choose to end the process by clicking *Stop*.

When training has completed (or has been halted by the user), two files will automatically be created. These will be in the same location as the original data file, and take the same name but with 'CLUSTERED' and 'CLUSTERS' added, together with the number of clusters used and an index to avoid overwriting any previously created files. For example, clustering data from C:\Temp\MyData.csv into 25 clusters will produce the two files C:\Temp\MyData_CLUSTERED_25_1.csv and C:\Temp\MyData_CLUSTERS_25_1.csv. These are, respectively, the 'clustered data file' and the 'clusters file', and are required for viewing the clusters that have been created.

Click *Return to main menu* to exit the MI-max clustering screen.

3.4. R-max ordering

To arrange a set of data clusters in an output space, select the option *R-Max ordering* from the main menu (Figure D2.1). The R-Max ordering screen appears (Figure D3.6).

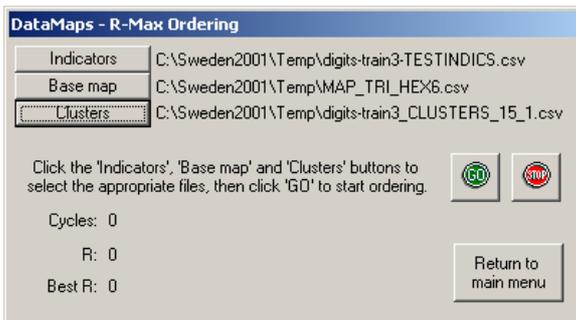


Figure D3.6 R-max ordering screen.

Click the *Indicators*, *Base map* and *Clusters* buttons to enter the file names and locations of the files, using standard Windows dialogue boxes, then click *Go* to start the R-max ordering process. You will be asked to input the maximum number of cycles for which R-max should run, and to define/select a file to store the output (using a standard Windows dialogue box). This output file is referred to as the mapping file, and simply specifies a mapping between the nodes in the chosen output map and clusters from the clusters file; if you are working with files based on the original main data file C:\Temp\MyData.csv (for example), you should choose a related ‘meaningful’ filename such as C:\Temp\MyData_MAPPING_1.csv. When a mapping file has been chosen, the R-max process will begin. R-max ordering may take a long time to complete, particularly if there are a large number of indicators or clusters, or if the output map is large. The user is kept informed of how many cycles have passed, and the current value of the regression coefficient R (which R-Max attempts to maximise) as well as the highest value of R so far achieved. Whenever a higher value of R is achieved, a file is created (or updated if the file already exists) to store the current mapping between clusters and base map locations. The R-Max process can be halted at any time by clicking *Stop*.

Click *Return to main menu* to exit the R-Max ordering screen.

3.5. Main data viewer

The results of MIR-Max clustering and ordering can be viewed by selecting *View clustered and ordered data* from the main menu (Figure D2.1). This option requires a ‘configuration file’ – a plain text file that specifies the locations of the various files required by MIR-max (see Section 4.8). Configuration files can be created using the *Create a configuration file* option, which displays the configuration file creation screen (Figure D3.7).

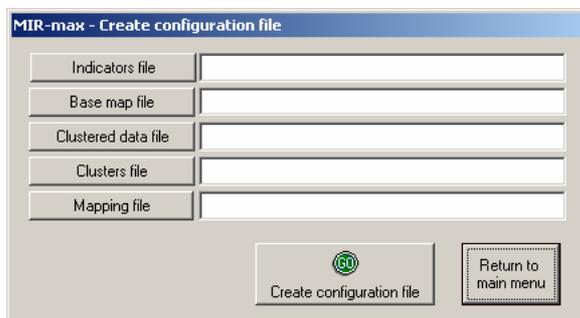


Figure D3.7 Configuration file creation screen.

Type in the names of the various files (including the full pathname), or click the file buttons to select using standard Windows dialogue boxes. When all the information has been correctly filled in, click *Create configuration file* to create the file. Again, a standard Windows dialogue box enables you to specify the file name and location.

On selecting *View clustered and ordered data* from the main menu, you are asked to choose the appropriate configuration file. MIR-max then reads the files and initialises the system. This may take a short while; a series of bars keep the user informed of progress. When initialisation is complete, the main data viewer is displayed (Figure D3.8).

On the left of the main viewer is the output map. Each cluster is represented as a coloured circle showing the average value of the currently selected indicator for that cluster. If the circle is grey, there is not enough information for that cluster to provide an average value. By default, the first indicator is displayed; to change to a different indicator, use the drop-down list below the map.

At the right of the screen is a set of six tabbed panels: *About MIR-max*, *Template*, *Report*, *Print*, *Archive samples* and *Input samples*. By default, the *About MIR-max* panel is displayed initially. This gives information about MIR-max, repeating the original disclaimer and providing a reference to the CIES website. (In future versions, this panel is intended to provide interactive help and information functions for the user.) Each panel is selected by clicking its title 'tab'.

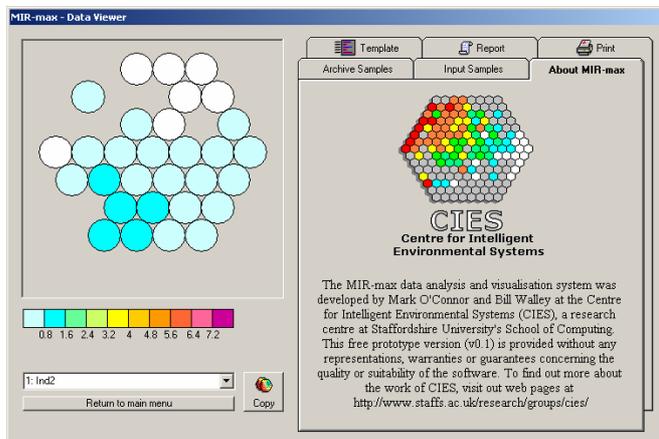


Figure D3.8 Main data viewer.

3.5.1. Template panel

The *Template* panel displays a bar chart or 'template', each indicator represented by a separate bar. Initially this will be blank, because no cluster has been selected. To select a cluster, click the coloured circle on the output map at the left of the screen. The cluster is highlighted by a blue circle and becomes the 'current cluster'. The current cluster can be changed at any time by clicking on the output map. The average values of each indicator for samples in the current cluster are displayed as blue bars on the *Template* panel, scaled between the minimum and maximum recorded indicator values (Figure D3.9).

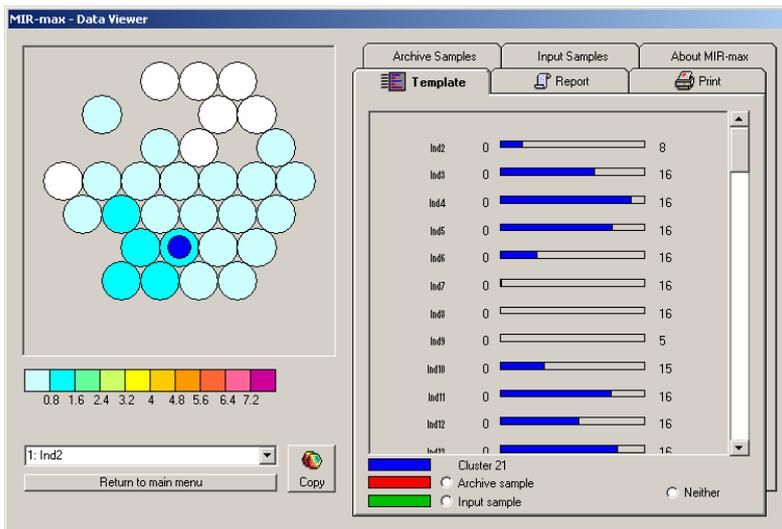


Figure D3.9 Highlighted cluster, and *Template* panel.

If an archive or input sample has been selected, the actual values recorded for that sample can also be displayed on the *Template* panel for comparison with cluster averages (Figure D3.10). Archive samples use a red bar, and input samples use a green bar. The key at the bottom of the *Template* screen shows which cluster, archive sample and input sample are currently chosen.

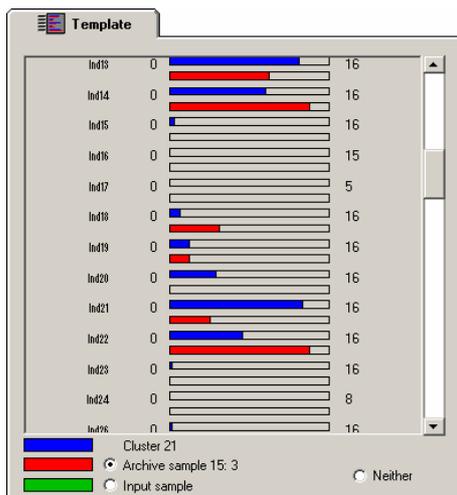


Figure D3.10 *Template* panel with archive sample selected.

3.5.2. Report panel

The *Report* panel (Figure D3.11) provides a summary of information about the current cluster. The unique cluster ID number and coordinates of the cluster on the output map are given, together with the total number of samples in the cluster. This information is followed by a list of average, minimum and maximum values of each of the indicators for the current cluster, and number of samples in the cluster for which information was available for that indicator. Below this information is another list, which gives the ID numbers and names of all samples in the cluster.

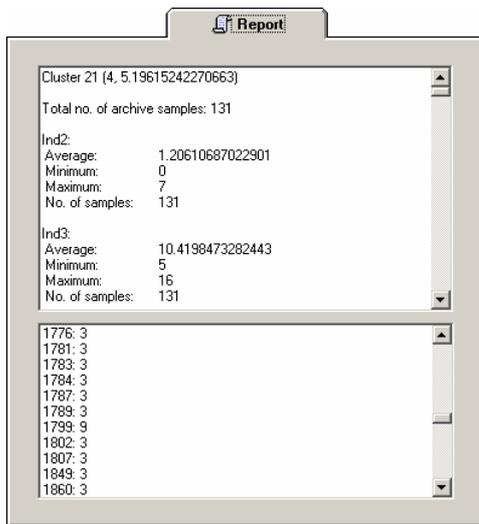


Figure D3.11 Report panel.

3.5.3. Archive samples panel

The *Archive samples* panel (Figure D3.12) enables the user to view information about a particular sample that was used in the MIR-max training process. (An ‘archive’ sample is one that appeared in the original data file used when training MIR-max.)

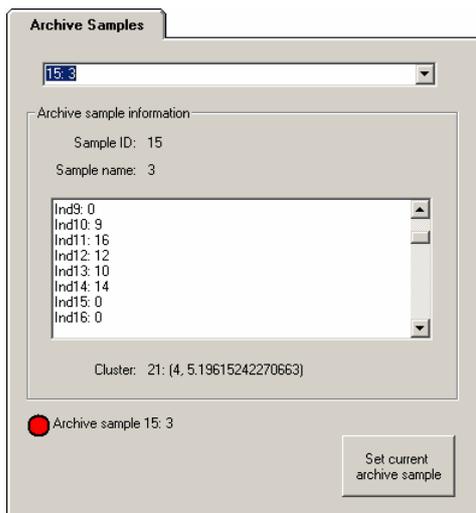


Figure D3.12 Archive samples panel.

An archive sample is selected using the drop-down list at the top of the panel. The sample ID number and name are shown, together with a list of values for each indicator, and the ID and coordinates of the cluster to which it belongs. Click *Set current archive sample* to set this sample as the current one (which is displayed on the output map and on the *Template* panel). The cluster to which the current archive sample belongs is highlighted on the output map by a red circle. (If this is also the current cluster, the red sample highlight overrides the blue cluster highlight.)

3.5.4. Input samples panel

A ‘new’ sample – one that did not appear in the original data file and so was not used in the MIR-max training process – is referred to as an ‘input sample’. Input samples are entered from comma-delimited files, using the *Input new data* button on the *Input samples* panel (Figure D3.13). Once the data has been entered, input sample information can be viewed in the same way as archive sample information. Click *Set as current input sample* to set the selected sample as the current one (which is displayed

on the output map and on the *Template* panel). As the input samples are 'new' to the system, MIR-max has to find which is the 'best matching' cluster and allocates the sample to this cluster. The cluster is highlighted by a green circle (which overrides the blue cluster highlight and red archive sample highlight if necessary). The output map displays which cluster is considered the 'best', and also gives an indication of which other clusters provide a possible match: clusters are colour-coded using a grading from white (no match) to yellow (best match).

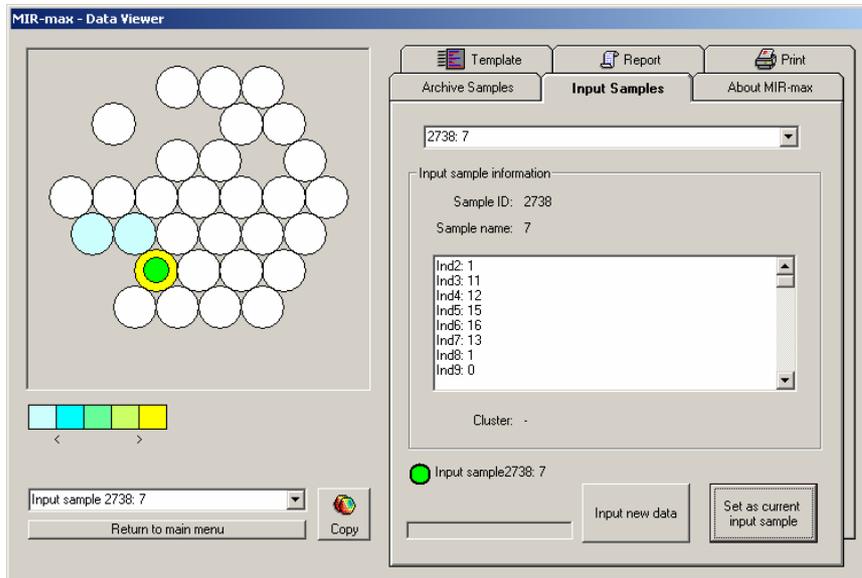


Figure D3.13 Input samples panel, with selected sample allocated to cluster.

3.5.5. Other functions

The *Print* panel enables the user to create a printout of the summary information for the current cluster, current archive sample, or current input sample.

The *Copy* button produces a new window that displays a copy of the current output map; this is useful for comparing feature maps.

Click *Return to main menu* to exit the MIR-max main data viewer.

4. File Formats

4.1 General

With the exception of 'configuration files', files used by MIR-max are comma-delimited files (which take a .csv filename extension). These can be edited using any standard text editor (such as Notepad) or spreadsheet (such as Excel). The data files contain the 'raw data' provided by the user; other files are created automatically using MIR-max (although they can still be edited using a text editor or spreadsheet). Each file begins with a line giving headings for the spreadsheet columns. If the first column heading starts with the letters 'ID', Excel may throw an error ('SYLK: File format is not valid'). The first heading should instead be 'Indicator ID' or 'Sample ID', for example, to avoid this error.

4.2 Data files

Comma-delimited (.csv) files.

Sample Name/ID, Indicator 1, Indicator 2 ...Indicator n.

4.3 Indicator files

Comma-delimited (.csv) files.

Indicator ID, Name, Units, Use in training?, Minimum, Maximum, Number of training categories, Training category bound 1, Training category bound 2 and so on, Number of display categories, Display category bound 1, Display category bound 2 and so on.

4.4 Base map files

Comma-delimited (.csv) files.

Node ID, X coordinate, Y coordinate.

4.5 Clustered data files

Comma-delimited (.csv) files.

Sample ID, Name, Indicator 1, Indicator 2 ...Indicator n, Cluster ID.

4.6 Cluster files

Comma-delimited (.csv) files.

Cluster ID, Measure, Indicator 1, Indicator 2 ...Indicator n.

Each cluster has information on seven 'measures' for each indicator, so seven lines are used:

Cluster ID, No. of samples, Ind. 1, Ind. 2, ... ,Ind. n
Cluster ID, Average value, Ind. 1, Ind. 2, ... ,Ind. n
Cluster ID, Minimum value, Ind. 1, Ind. 2, ... ,Ind. n
Cluster ID, Maximum value, Ind. 1, Ind. 2, ... ,Ind. n
Cluster ID, Average category, Ind. 1, Ind. 2, ... ,Ind. n
Cluster ID, Minimum category, Ind. 1, Ind. 2, ... ,Ind. n
Cluster ID, Maximum category, Ind. 1, Ind. 2, ... ,Ind. n

4.7 Mapping files

Comma-delimited (.csv) files.

Node ID, Cluster ID.

4.8 Configuration files

Plain text (.txt) files.

Indicator file name/location.

Archive samples (clustered) file name/location.

Clusters file name/location.

Base map file name/location.

Mapping file name/location.

5 Acknowledgements and contact details

5.1. Acknowledgements

The MIR-max software was written by Mark O'Connor and Bill Walley at Staffordshire University's Centre for Intelligent Environmental Systems (CIES), using Microsoft Visual Basic 6. Thanks are due to the Environment Agency and Staffordshire University for their support, and to Ray Martin and David Trigg from CIES for their work on aspects of the project.

5.2. CIES - Centre for Intelligent Environmental Systems

The Centre for Intelligent Environmental Systems is a Centre within Staffordshire University's School of Computing. The Centre specialises in the application of advanced computing techniques, especially artificial intelligence (AI), to problems affecting the natural environment. Projects to date have concentrated on the development of intelligent systems for the biological monitoring of river quality. The Centre's expertise in this field has grown out of the pioneering work carried out by Bill Walley and Bert Hawkes in the early 1990s. Although biomonitoring will remain the principal application domain of the group, some diversification into other environmental applications is envisaged. To find out more about the work of the Centre, visit and explore our extensive web pages at:

<http://www.cies.staffs.ac.uk>

Or, contact us at:

Centre for Intelligent Environmental Systems
Staffordshire University
School of Computing
The Octagon
Beaconside
Stafford ST18 0AD
UK

Dr Martin Paisley
m.f.paisley@staffs.ac.uk
+44 (0)1785 353510

Dr David Trigg
d.j.trigg@staffs.ac.uk
+44 (0)1785 353445

Appendix E

PISCES codes for sector, activity and pressure

Code	Sector
1	Agriculture (including forestry)
2	Water industry
3	Urban (built development)
4	Navigation (including ports and inland)
5	Mining, quarrying and aggregate extraction
6	Industry and business - other
7	Industry and business - power generation
8	Industry and business - manufacturing
9	Industry and business - construction
10	Transport - roads
11	Transport - rail
12	Transport - shipping
13	Private water treatment
14	Flood risk management
15	Natural processes
16	Waste
17	Recreation
18	Fishing (commercial)
19	Biodiversity/conservation
20	Unknown
21	Transport - aviation

Code	Activity
1	Abstraction (SW)
2	Abstraction (GW)
3	Hydro power flow alterations
4	Reservoir or augmentation related flow alterations
5	Water discharge from GW
6	Water discharge from river transfer
7	Managed flow regime (e.g. compensation, augmentation)
8	
9	Intermittent discharge (not CSO or SSO)
10	STW discharge (SW) - combined sewer overflow
11	STW discharge (SW) - septic tank
12	STW discharge (SW) - storm sewer overflow
13	STW discharge (SW) - treated STW effluent
14	Water treatment works discharge
15	Misconnections
16	Industrial discharge
17	Pollution incident
18	Air emissions (atmospheric deposition)
19	Coal mine drainage
20	Metal mine drainage
21	Mine spoil heaps
22	
23	Livestock bank-side erosion
24	Rural track run-off
25	Land drainage
26	Habitat management
27	Bank reinforcement
28	Culverts
29	Barrages
30	Weirs, penstocks, locks and sluices
31	Realignment, re-profiling and re-grading
32	Shoreline reinforcement
33	Land claim
34	Dredging (sediment management)
35	Dredging (commercial fishing and shellfish)
36	Aggregate extraction
37	Quarrying
38	Cable laying
39	
40	Farming - arable
41	Farming - livestock other (e.g. manure management)
42	Farming - other
43	Afforestation
44	Agri-industry - dairy
45	Agri-industry - food processing

46	Agri-industry - other
47	Water cress beds
48	
49	Fish stocking
50	Fishing - commercial (not habitat effect)
51	Fish farming
52	Fishing - recreation
53	
54	Weed management
55	
56	Illegal introduction of species
57	Natural expansion of species distribution
58	
59	Navigation (e.g. boat traffic/commercial shipping)
60	Boat moorings
61	
62	Non-ag diffuse pollution - airport
63	Non-ag diffuse pollution - roads
64	Non-ag diffuse pollution - current waste disposal
65	Non-ag diffuse pollution - historic waste disposal
66	Non-ag diffuse pollution - contaminated land
67	Non-ag diffuse pollution - industry
68	Non-ag diffuse pollution - railway
69	Non-ag diffuse pollution- urban/sub-urban development
70	
71	Natural processes e.g. natural mineralisation
72	Natural bank erosion
73	Problem with sample
74	
75	Climate change
76	
77	Unknown
78	Construction
79	Steel making, basic slag
80	Non-ag diffuse pollution – other

Code	Pressure
1	Reduced volume of water
2	Increased volume of water
3	Increased flow
4	Decreased flow
5	
6	Physical modification
7	Noise/vibration
8	
9	Alien species (add specific species)
10	Fish disease
11	Fish predation
12	Biota removal
13	
14	1,1,1-trichloroethane
15	1,1,2-trichloroethane
16	1,2,3-trichlorobenzene
17	1,2,4-trichlorobenzene
18	1,2-dichloroethane
19	1,3,5-trichlorobenzene
20	2,4 D Esters TOTAL
21	2,4,6-trichlorophenol
22	2,4-d (non ester)
23	2,4-d butyl ester (2,4 dichlorophenoxyacetic acid butyl ester)
24	2,4-d butylglycol ester (2,4 dichlorophenoxyacetic acid butyl glycol ester)
25	2,4-d iso-octyl ester (2,4 dichlorophenoxyacetic acid iso-octyl ester)
26	2,4-d methyl ester (2,4 dichlorophenoxyacetic acid methyl ester)
27	2,4-dichlorophenol
28	2,5-dichlorophenol
29	2,5-dimethylphenol
30	2-chloro-4-nitrotoluene
31	2-chloro-5-nitrotoluene
32	2-chloro-6-nitrotoluene
33	2-chlorophenol
34	2-methylphenol
35	3-methylphenol
36	4-chloro-2-nitrotoluene
37	4-chloro-3-methylphenol
38	4-chloro-3-nitrotoluene
39	Aldrin
40	Aluminium sulphate
41	Ammonia
42	Anthracene
43	Arsenic

44	Atrazine
45	Azinphos-methyl
46	Barium
47	Bentazone
48	Benzene
49	Benzo-[A]-pyrene
50	Benzo-[B]-fluoranthene
51	Benzo-[K]-fluoranthene
52	Benzo-ghi-perylene
53	Biphenyl
54	BOD
55	Boron
56	Cadmium
57	Carbon tetrachloride
58	Chloride
59	Chlorofenvinphos
60	Chloroform (trichloromethane)
61	Chloronitrotoluenes total
62	Chloropyrifos
63	Chromium
64	Colour
65	Conductivity
66	Copper
67	Cyanide
68	Cyfluthrin
69	Cypermethrin
70	DDE (pp) (dichlorodiphenyldichloroethylene)
71	DDT (op) (dichlorodiphenyltrichloroethane)
72	DDT (pp) (dichlorodiphenyltrichloroethane)
73	DDT total (dichlorodiphenyltrichloroethane)
74	Demeton-O
75	Demeton-O-methyl
76	Demeton-S
77	Demeton-S-methyl
78	Demeton-S-methyl sulphone
79	Demetons
80	Detergents anionic synthetic (surfactants)
81	Detergents non-anionic synthetic (surfactants)
82	Di(2-ethylhexyl)phthalate
83	Diazinon
84	Dichloromethane
85	Dichlorvos
86	Dieldrin
87	Dimethoate

88	Dissolved oxygen
89	Diuron
90	Drins total (Aldrin, Dieldrin, Endrin, Isodrin)
91	Endosulphan a
92	Endosulphan A&B
93	Endosulphan b
94	Endrin
95	Enterovirus
96	Faecal coliforms (Confirmed)
97	Faecal streptococci (Confirmed)
98	Fenitrothion
99	Flucofuron
100	Fluoranthene
101	Fluoride
102	HCH (hexachlorocyclohexane)
103	Hexachlorobenzene
104	Hexachlorobutadiene
105	Hydrocarbons
106	Indeno-[1,2,3-CD]-pyrene
107	Iron
108	Iron sulphate
109	Isodrin
110	Isoproturon
111	Lead
112	Linuron
113	Malathion
114	Manganese
115	MCPA
116	Mecoprop
117	Mercury
118	Mevinphos
119	Microbiology (not specified)
120	Napthalene
121	Nickel
122	Nitrate
123	Nitrite
124	Nonylphenol
125	Octylphenol
126	Omethoate
127	PAH total (polycyclic aromatic hydrocarbons)
128	Parathion
129	PCBs
130	PCSDs (polychloro chloromethyl sulphonamido diphenyl ether)
131	Pentachlorophenol

132	Perchloroethylene (tetrachloroethene)
133	Permethrin
134	pH
135	Phenol
136	Phosphate
137	Salinity
138	Salmonella
139	Sediment (including suspended solids)
140	Selenium
141	Silver
142	Simazine
143	Sulcofuron
144	Sulphate as SO ₄
145	TDE (pp) (tetrachlorodiphenylethane)
146	Temperature
147	Toluene
148	Total Coliforms (Confirmed)
149	Triazophos
150	Tributyltin
151	Trichlorobenzene total
152	Trichloroethylene
153	Trifluralin
154	Triphenyltin
155	Vanadium
156	Xylene
157	Zinc
158	
159	Disinfectant
160	Metals
161	Nutrients
162	Organics
163	Pesticides - sheep dip
164	Pesticides - other
165	Sanitary
166	
167	Unknown
168	Pesticides - herbicides
169	Cave
170	Drought
171	Flood
172	Freshwater but tidal
173	Heavily shaded
174	Lake or pond close u/s
175	Moorland drainage

176	Reedbed
177	Winterbourne/non-permanent stream
187	Other natural feature

Appendix F

Stress categories and associated activity, source and pressure codes

Stress code	Stress Source	Stress category	Stress type	PISCES pressure code	PISCES activity code	PISCES sector code
AD	Pollution	Acid	Acid deposition		018	007
EX	Pollution	Acid	Rock exposed by construction		078	009
CF	Pollution	Agricultural run-off	Forestry (conifer)		043	001
IA	Pollution	Agricultural run-off	Intensive arablisation		040	001
SL	Pollution	Agricultural run-off	Livestock slurry		041	001
SI	Pollution	Agricultural run-off	Silage		042	001
AO	Pollution	Agricultural run-off	Other (specify)		042	001
AB	Pollution	Agri-industry	Abattoir/meat processing/rendering		045	006
BR	Pollution	Agri-industry	Brewery		045	006
DA	Pollution	Agri-industry	Dairy		044	006
FL	Pollution	Agri-industry	Flour mill		045	006
MF	Pollution	Agri-industry	Mushroom farm		042	006
SU	Pollution	Agri-industry	Sugar refinery		045	006
TA	Pollution	Agri-industry	Tanning/leather		046	006
VE	Pollution	Agri-industry	Vegetable processing		045	006
WO	Pollution	Agri-industry	Wool		046	006
AI	Pollution	Agri-industry	Other (specify)		046	006
DF	Pollution	Farming	Disinfectant	159	042	001
FE	Pollution	Farming	Fertilisers	161	040	001
FF	Pollution	Farming	Fish farming		051	001
HE	Pollution	Farming	Herbicides	168	040	001
IN	Pollution	Farming	Insecticides	164	040	001
SD	Pollution	Farming	Sheep-dip	163	041	001
WC	Pollution	Farming	Water cress beds		047	001
FA	Pollution	Farming	Other (specify)		042	001
BW	Pollution	Industrial discharge	Brick works		016	008
CE	Pollution	Industrial discharge	Cement works		016	008
DY	Pollution	Industrial discharge	Coloration (dye)		016	008
CW	Pollution	Industrial discharge	Cooling water (warm)	146	016	007
DE	Pollution	Industrial discharge	Detergent		016	008
HI	Pollution	Industrial discharge	Heavy industry		016	008
LI	Pollution	Industrial discharge	Light industry/commercial		016	006
PM	Pollution	Industrial discharge	Paper mill		016	008
PC	Pollution	Industrial discharge	Petrochemicals (mfr & distribution)	162	016	008
PL	Pollution	Industrial discharge	Plating		016	008

ID	Pollution	Industrial discharge	Other (specify)		016	006
CB	Pollution	Mines, quarries & extractions	Brick-clay extraction		036	005
CC	Pollution	Mines, quarries & extractions	China-clay extraction		036	005
CM	Pollution	Mines, quarries & extractions	Coal mine drainage		019	005
MM	Pollution	Mines, quarries & extractions	Metal mine drainage		020	005
QA	Pollution	Mines, quarries & extractions	Quarry (acid rock)		036	005
QB	Pollution	Mines, quarries & extractions	Quarry (limestone/chalk)		036	005
SG	Pollution	Mines, quarries & extractions	Sand & gravel		036	005
MI	Pollution	Mines, quarries & extractions	Other (specify)		079	005
AF	Pollution	Run-off (non-agric.)/Leachate	Aircraft/airfield de-icing (specify)		062	021
BU	Pollution	Run-off (non-agric.)/Leachate	Building/road construction		063	009
DL	Pollution	Run-off (non-agric.)/Leachate	Domestic landfill		064	016
FY	Pollution	Run-off (non-agric.)/Leachate	Fly tipping		064	016
HR	Pollution	Run-off (non-agric.)/Leachate	Heavy industry		067	008
HY	Pollution	Run-off (non-agric.)/Leachate	Highway (incl. De-icing salt)	137	063	010
LR	Pollution	Run-off (non-agric.)/Leachate	Light industry/commercial		067	006
RU	Pollution	Run-off (non-agric.)/Leachate	Motorway (incl. De-icing urea)	162	063	010
RR	Pollution	Run-off (non-agric.)/Leachate	Railway		068	011
SY	Pollution	Run-off (non-agric.)/Leachate	Scrap yard		064	006
SH	Pollution	Run-off (non-agric.)/Leachate	Slag heap		079	007
TI	Pollution	Run-off (non-agric.)/Leachate	Toxic/industrial landfill		066	016
TY	Pollution	Run-off (non-agric.)/Leachate	Tyres		064	006
UR	Pollution	Run-off (non-agric.)/Leachate	Urban/suburban		069	003
RO	Pollution	Run-off (non-agric.)/Leachate	Other (specify)		080	
GR	Pollution	STW to aquifer	Via groundwater recharge		005	002
CS	Pollution	STW to river	Combined sewer overflow (CSO)		010	002
SE	Pollution	STW to river	Septic tank		011	002

SS	Pollution	STW to river	Storm sewer overflow (SSO)		012	002
TS	Pollution	STW to river	Treated STW effluent		013	002
ST	Pollution	STW to river	Other (specify)		080	002
AS	Pollution	WTW	Aluminium sulphate	040	014	002
FS	Pollution	WTW	Iron sulphate	108	014	002
SW	Pollution	WTW	Swimming pool		014	002
WT	Pollution	WTW	Other (specify)		014	002
SB	Activities	Artificial bank at site	Consolidated (stone/brick/concrete)	006	027	
GA	Activities	Artificial bank at site	Gabions	006	027	
SP	Activities	Artificial bank at site	Metal piling	006	027	
UC	Activities	Artificial bank at site	Unconsolidated (rip-rap/boulders)	006	027	
AT	Activities	Artificial bank at site	Other (specify)	006	027	
EC	Activities	Bank erosion at site	Clay	006	072	
EG	Activities	Bank erosion at site	Gravel, boulder	006	072	
ES	Activities	Bank erosion at site	Sand	006	072	
BM	Activities	Bank practices at site	Boat moorings	006	060	004
LV	Activities	Bank practices at site	Livestock poaching/overgrazing	006	023	001
MO	Activities	Bank practices at site	Mown/managed riparian zone	006	054	
BP	Activities	Bank practices at site	Other (specify)	006		
DI	Activities	Channel at the site	Artificial ditch or dyke	006	025	
BE	Activities	Channel at the site	Bedrock	015	071	
BG	Activities	Channel at the site	Bridge	006		021
CN	Activities	Channel at the site	Canal (artificial navigation)	006	027	004
CA	Activities	Channel at the site	Canalised stream/river (non-navigation)	006	027	
CH	Activities	Channel at the site	Choked channel (>33% plant)	006	071	
BD	Activities	Channel at the site	Concrete stream bed	006	027	
CU	Activities	Channel at the site	Culvert	006	028	
DN	Activities	Channel at the site	Dredging	006	034	004
RN	Activities	Channel at the site	River navigation (locks etc)	006	030	004
RA	Activities	Channel at the site	River restoration	006	026	019
WD	Activities	Channel at the site	Weed cutting	006	054	014
AN	Activities	Channel at the site	Other (specify)	006		

GS	Activities	Eroded material in channel	Gravel, boulder	139		
IS	Activities	Eroded material in channel	Inert siltation	139		
GW	Activities	Flow-related	Augmentation from groundwater	002	005	002
RT	Activities	Flow-related	Augmentation from river transfer	002	004	002
CD	Activities	Flow-related	Cessation of STW discharge	001	007	002
AG	Activities	Flow-related	Groundwater abstraction	001	002	002
HW	Activities	Flow-related	Hypolimnic water		004	002
PF	Activities	Flow-related	Ponded flow (lake/reservoir d/s)	004	007	002
RF	Activities	Flow-related	Regulated flow (lake/reservoir u/s)		007	002
AR	Activities	Flow-related	River abstraction	001	001	002
PN	Activities	Flow-related	Summer penning		030	001
WE	Activities	Flow-related	Weirs		030	002
FR	Activities	Flow-related	Other (specify)			
RI	Activities	Reclaimed land	Industrial		033	006
OC	Activities	Reclaimed land	Opencast		033	005
RL	Activities	Reclaimed land	Other (specify)		033	
CV	Natural	Natural features	Cave	169	071	015
DT	Natural	Natural features	Drought	170	071	015
FD	Natural	Natural features	Flood	171	071	015
FT	Natural	Natural features	Freshwater but tidal	172	071	015
HS	Natural	Natural features	Heavily shaded site	173	071	015
LP	Natural	Natural features	Lake or pond close u/s	174	071	015
MD	Natural	Natural features	Moorland drainage	175	071	015
RB	Natural	Natural features	Reedbed at the site	176	071	015
WI	Natural	Natural features	Winterbourne/non-permanent stream	177	071	015
LU	Natural	Natural features	Other (specify)	178	071	015
AC	Survey	Sampling difficulty	Access to one bank only *		073	019
AL	Survey	Sampling difficulty	Air-lift *		073	019
BO	Survey	Sampling difficulty	Bouldery site		073	019
DR	Survey	Sampling difficulty	Dredge *		073	019
MS	Survey	Sampling difficulty	Mobile substrate		073	019
DS	Survey	Sampling difficulty	Other (specify)		073	019
BS	Survey	Sorting problem	Bank-side sort *		073	019
PR	Survey	Sorting problem	Poorly preserved sample		073	019
NI	Negatives	No information	No information *			
NP	Negatives	No perceived stress	No perceived stress *			
MY	Negatives	Stress not identified	Unknown source	167		

EF	Effects	Eutrophication	Agriculture	161		001
EA	Effects	Eutrophication	Angling	161	052	017
EE	Effects	Eutrophication	Sewage	161		002
EW	Effects	Eutrophication	Wildfowl	161	071	
EO	Effects	Eutrophication	Other (specify)	161		
TX	Effects	Historical activity (now ceased)	Toxic sediment			
DC	Effects	No flow	Dry channel (caused by man)	001		
CO	Effects	Oils, petrochemicals	Crude	162		
FO	Effects	Oils, petrochemicals	Fuel (diesel/petrol)	162		
LO	Effects	Oils, petrochemicals	Lubricating	162		
TO	Effects	Oils, petrochemicals	Tar/bitumen	162		009
VO	Effects	Oils, petrochemicals	Vegetable	162		006
OI	Effects	Oils, petrochemicals	Other (specify)	162		
CL	Effects	Other indicators	<i>Cladophora</i>			
OH	Effects	Other indicators	Ochre			
SF	Effects	Other indicators	Sewage fungus			
IL	Effects	Saline	Industrial discharge	137	016	
IG	Effects	Saline	Inland geological	137	071	015
MA	Effects	Saline	Marine or estuarine origin	137	071	015
SA	Effects	Saline	Other (specify)	137		

Appendix G

BBN Creator

BBN Creator

**An aid for BBN construction and analysis.
Version 0.1 (alpha)**

User Guide

David Trigg & William J. Walley

**Centre for Intelligent Environmental Systems
School of Computing
Staffordshire University**

March 2002

An aid for BBN construction and analysis.
Version 0.1 (alpha)

User Guide

Contents

1. Introduction	G-4
1.1 Background.....	G-4
1.2 Bayesian Belief Networks	G-4
1.3 Using BBN output	G-6
2. Using BBN Creator	G-7
2.1 Installing BBN Creator	G-7
2.2 The BBN Creator Interface	G-7
2.3 Create BBN Information File.....	G-7
2.3.1 Selecting Variables	G-9
2.3.2 Checking Names.....	G-9
2.3.3 Choosing a data type	G-9
2.3.4 Defining States	G-10
2.4 Add Nodes to a New or Existing Network.....	G-11
2.5 Create links between Nodes.....	G-12
2.6 Calculate BBN probability values.....	G-13
2.7 Test Network.....	G-14
2.8 Calculate Mutual Information Ranking file	G-17
2.9 Automatically create links between Nodes	G-19
4. File Formats	G-21
4.1 General	G-21
4.2 BBN Information Files (bif)	G-21
4.3 Network Files	G-21
4.4 Ranking files.....	G-21
5. Acknowledgements and Contact Details	G-22
5.1 Acknowledgements	G-22
5.2 CIES - Centre for Intelligent Environmental Systems	G-22

1. Introduction

1.1 Background

BBN Creator is a system created to automate some of the tasks involved in the construction of Bayesian Belief Network (BBN) and works in collaboration with the HUGIN Bayesian Belief Network development software. The application itself is a more refined version of pieces of software used as part of a research project concerning river pollution in England and Wales.

The original software was used to produce a general pollution diagnostic BBN, which is the basis for the software package RPBBN (River Pressure Bayesian Belief Network). The software consisted of a number of VBA (Visual Basic for Applications) functions that were designed specifically to be used with the project data, Microsoft® Excel® and the HUGIN API (Application Programmers Interface). These functions automated construction, data analysis and testing tasks and significantly reduced network development time.

The need to use these functions on other datasets and third party interest led to the functions being made generic and grouped into a module. This BBN creation module could be imported into an Excel® workbook and used in conjunction with the data. Although the module was efficient and robust, it required some familiarity with computer programming. Therefore, the next step in this development process was to provide a graphical user interface to simplify interaction with the module. The result is the BBN Creator application, which provides the benefits of the BBN creation module without the need for computer programming skills.

Short-term development goals of the software will involve improving input and output and a move away from using the HUGIN API and toward integration with the dBBN, which is CIES's own implementation of BBN technology developed for RPBBN. In the long-term the two packages could be merged fully to provide a full development suite with task automation.

1.2 Bayesian Belief Networks

Bayesian Belief Networks (BBN) are a probabilistic reasoning tool. This means that given a number of observations they produce likelihood values for the possible states of other variables in the network. A simple example of a BBN is the 'Chest Clinic' network (see Figure G1.1). Observations about smoking, recent health and/or x-ray results can be combined to predict whether a patient is more likely to be suffering from tuberculosis or cancer. The strength of BBN technology is the flexibility of the reasoning process. Observational data can be entered or retracted from any part of the network. The implications of this change propagate in all directions throughout the network diagnostically and prognostically.

A BBN comprises of two distinguishable components

- A network structure, consisting of nodes/vertices and relationships/arcs.
- Tables of associated probabilities, consisting of the probability of a state occurring given the state of any 'parent' nodes.

Note: The relationships are directed in BBN; causal factors from which the relationship arrows originate are often called 'parents', while the affected nodes are called 'children'.

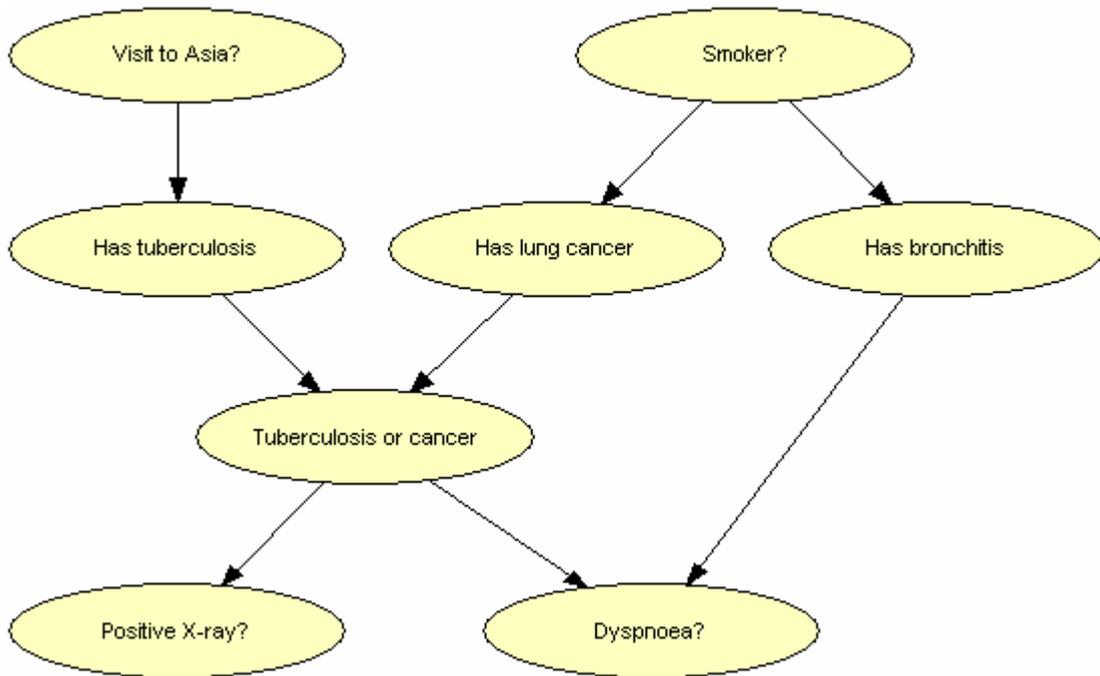
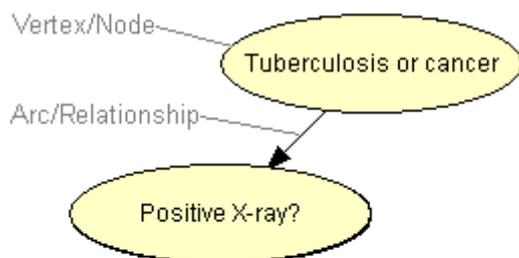


Figure G1.1 Chest clinic example network structure.

The structure of the network contains knowledge about the area of expertise. Causes and effects are mapped out providing a basis for the reasoning process. Each node in the network has an associated table of probability values. Figure G1.2 shows one such table, in this case the table associated with the node 'Positive X-ray?' taken from the example network.



Probability Table

Positive X-ray	Tuberculosis	
	yes	no
Positive	0.98	0.05
Negative	0.02	0.95

Figure G1.2 Probability table associated with positive X-ray?

The table is comprised of probability values for states of the variable, for each combination of states of the parent(s). For example here 'Positive X-ray' has the states 'Positive' and 'Negative' shown in the left column. It only has one parent 'Tuberculosis or cancer' that also has two states 'yes' and 'no' (the patient either has TB and/or Cancer or not). Therefore the number of values needed to complete a child node's probability table is directly linked to the number of states of the child and its parent(s).

From the network structure and probability tables, likelihood values can be calculated for each state of a node based on that of its parents and any observational data entered. The calculations are not complicated; multiplication is used in the causal direction as it is used with normal probability calculations. For calculations that go in the opposite direction, from effect to cause, the Bayes Theorem and equation is used.

$$\text{Probability of (B given A)} = \frac{\text{Probability of (A given B)} \times \text{Probability of (A)}}{\text{Probability of (B)}}$$

The advantage of using these methods of updating likelihood values is that they are mathematically sound. Therefore, unlike some other quasi-probabilistic methods, BBNs offer a reliable method of calculating and propagating likelihood values for any of the nodes throughout a network, based on new or updated observational data and/or likelihood values.

1.3 Using BBN output

The output form of a BBN is the updated likelihood values for the occurrence of a specific variable's states. In the case of the example network, this is the probability of the patient having a specific disease. The main difference between a BBN's output and that of a predictive system such as a multiple regression model is that the results are only a measure of likelihood and not a predicted value. For example, a BBN model designed to predict pH would produce a measure of likelihood of a specific state occurring, that is 75 per cent likelihood that the value is between pH 5 and 7. On the other hand a multiple regression model will predict an actual pH value, pH 5.67. Although this appears to be an obvious point as BBNs are probabilistic reasoning tool, it is worth reiterating that the output of BBN systems is distinctly different from predictive methods.

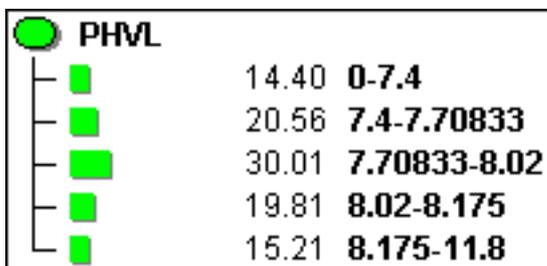


Figure G1.3 pH prediction chart.

The output from BBN is usually presented in both graphical and numeric formats. In the example illustrated in Figure G1.3 it appears that the pH value is most likely in the range 7.7 to 8.0, whilst the other nodes in the network are in their current states, although the level of likelihood means that this conclusion is far from certain.

Understanding the output of a BBN is largely dependent on appreciating the form and meaning of the results and using this information accordingly.

2. Using BBN Creator

2.1 Installing BBN Creator

BBN Creator is a Windows 32-bit application and runs on Windows 9.x, NT and 2000. The BBN Creator installation program is supplied on a CD-ROM that is configured to automatically start installation when the disk is inserted. If it does not, use Windows Explorer/My Computer to view the contents of the CD-ROM and then double click 'setup.exe'

2.2 The BBN Creator Interface

The main BBN Creator interface provides the option of performing seven automated tasks (see Figure G2.1). Used in conjunction with HUGIN the application is intended to act as an aide in the construction process.

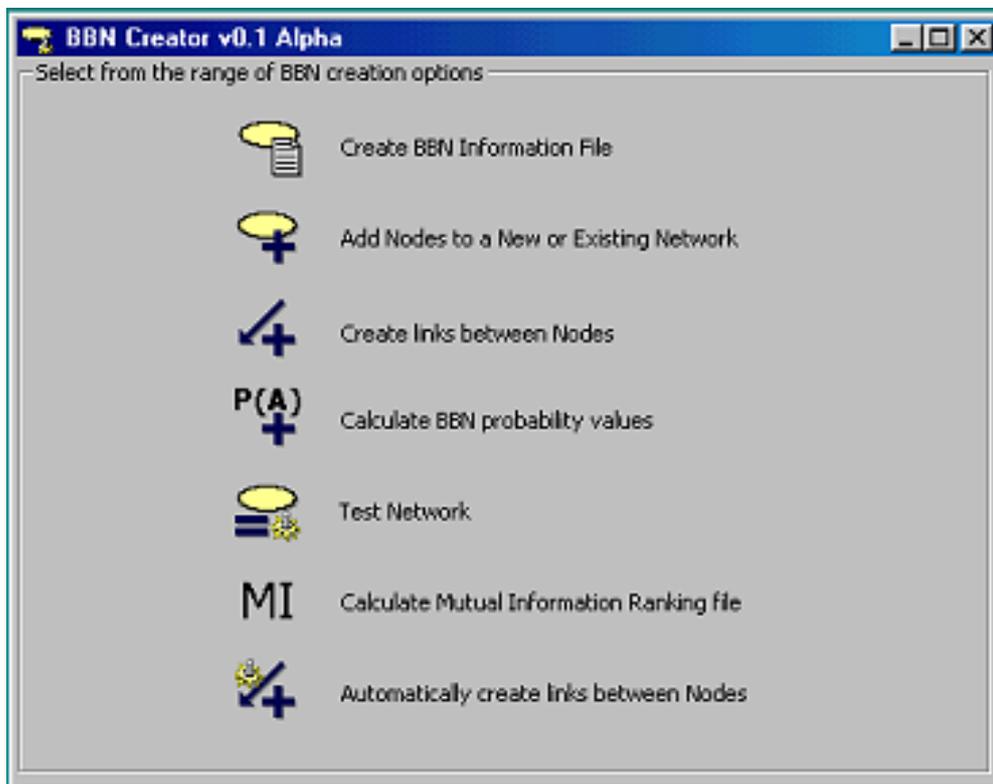


Figure G2.1 BBN Creator main screen.

The first four options are intended to simplify and/or automate tasks that are cumbersome to perform using the HUGIN interface. The remaining options perform more complicated tasks aimed to help the user to improve the BBN design.

2.3 Create BBN Information File

A BBN information file (bif) contains vital information about the variables used in a specific BBN, such as name, data type, number and name of states and state mean values. This file is used in the majority of tasks that BBN Creator performs and creating it is the initial task to be performed in the creation process. On clicking the 'Create BBN

Information File' link an open file dialogue box is shown, choose the data file on which the BBN is to be based. After selecting the appropriate data file the 'File specification' screen is displayed.

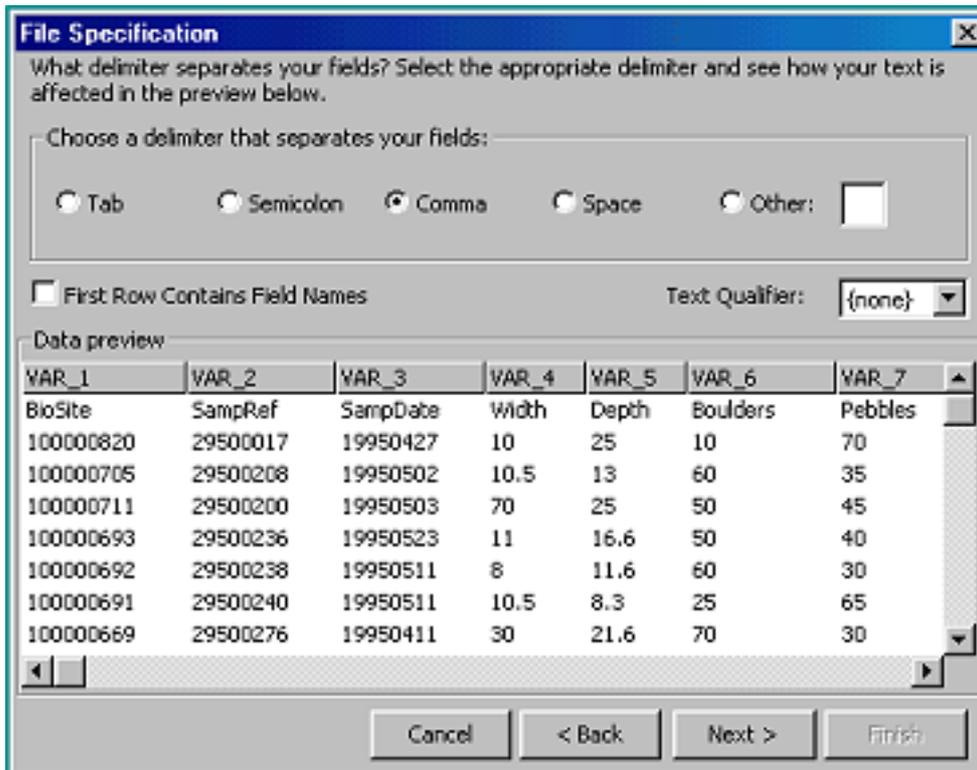


Figure G2.2 File specification screen.

The 'File specification' screen is used extensively in BBN Creator. Its purpose is to obtain information about the formatting of the data file from the user, such as delimiting characters and text qualifiers. The effect of specifying these values is shown in the 'Data Preview' box.

Clicking the 'Next' button leads on to the next stage where the data is extracted from the file and reformatted. In addition to these processes the program makes a decision about whether each variable is discrete or continuous data. The status of this process is shown on the processing screen (see Figure G2.3). Again this screen appears throughout BBN Creator to give the user feedback on the status and speed of the processing.

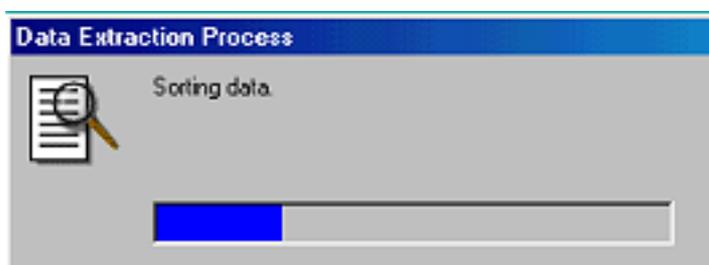


Figure G2.3 Processing status screen.

After all the processing has taken place the final screen in the process is shown (see Figure G2.4).

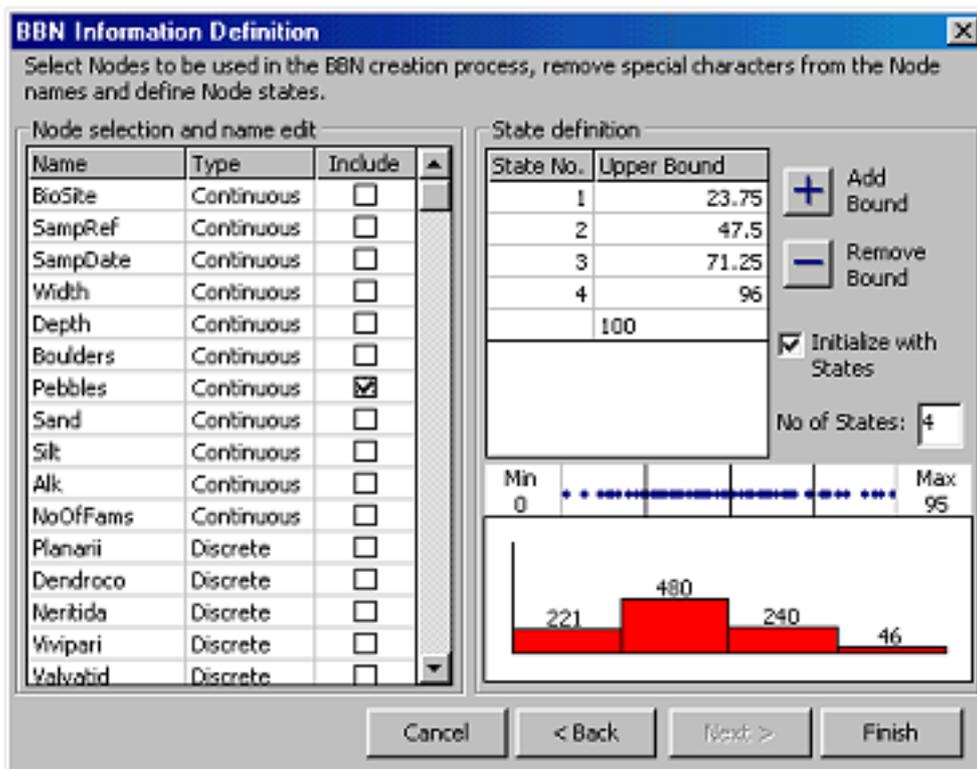


Figure G2.4 BBN information definition screen.

The screen has four main purposes:

1. Select variables to be used in the BBN creation process.
2. Check if the names of these variables can be used as node names or if they will have to be changed because they contain invalid characters.
3. Define the type of data for each variable, either discrete or continuous.
4. Define the states associated with each variable.

2.3.1 Selecting variables

Selecting and deselecting variables is achieved by clicking the corresponding check box.

2.3.2 Checking names

The range of characters that can be used in node names in HUGIN BBN is basically restricted to alphanumeric characters (“a-z”, “A-Z” and “0-9”) and the underscore character “_”. When names are loaded or changed they are checked for invalid characters. If they are invalid, names are greyed out. A variable with an invalid name will not be included in the BBN information file and so can take no part in the creation process. By double clicking on the correct box the value can be edited. When the editing process is finished, press ‘enter’ to update the name.

2.3.3 Choosing a data type

On loading the data, the system makes a decision on the data type of a particular variable, either discrete or continuous. The user has the option to alter the data type by clicking on the appropriate box and selecting the type from the list provided.

2.3.4 Defining States

The process of defining states is different depending on the data type (discrete/continuous).

Continuous data: states are defined by an upper bound value. These values are exclusive; the upper bound value is not part of the state. To help select states the maximum, minimum, spread of data and state bounds are shown along with a state graph (see Figure G2.4).

States are defined by double clicking the boxes in the 'Upper Bound' column, typing a value and pressing enter. Additional states are added by clicking the 'Add State' button and entering a new value. As the state values are updated the numbers of samples in each state are shown in red on the graph below, those not yet allocated to a state are shown in grey.

A state can be removed by clicking on the column and pressing the 'Remove State' button.

If a variable has no existing state information it can be automatically initialised with a number of states. This is achieved by choosing the 'Initialise with States' option and specifying the required number of states in the 'Number of States' text box. The state bounds are derived by equally dividing the total range into the required number of states.

Discrete data: the definition of discrete states is a different process and uses different methods than the definition of continuous states. A discrete variable produces a display similar to the one shown in Figure G2.5.

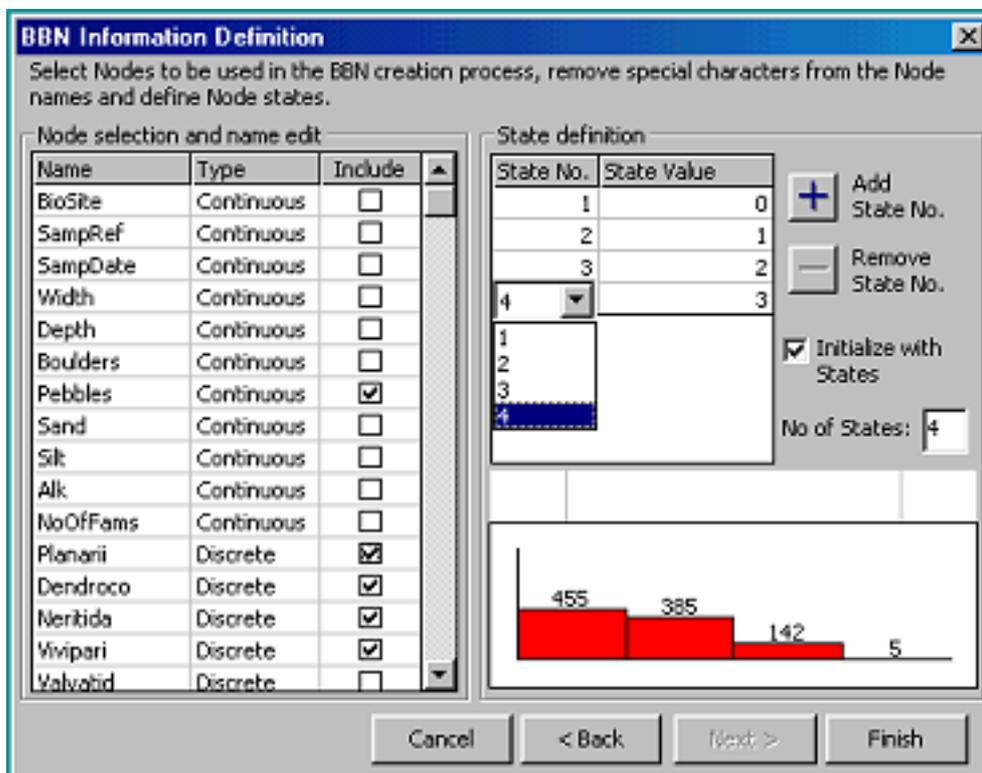


Figure G2.5 Defining discrete states.

If a variable is defined as discrete, each different value is stored and allocated a state number. When a discrete variable is selected the full range of values are shown with an

identifying 'State No.'. Due to the nature of the data, the only processes that can be performed on discrete states are merging two states or splitting merged states apart.

Unlike continuous variables, the states are altered by changing the 'State No.' to which they belong. Changing the 'State No.' to which a discrete value belongs either merges discrete values into a state or splits them apart.

To merge states click on the 'State No.' of the appropriate discrete value. A drop down list is shown with all the current 'State No.' values. By selecting the state number of one of the other states, the two will be merged. Using the example in Figure G2.5 the user has decided to merge the discrete value '3' with one of the other values. The corresponding 'State No.' box has been clicked and a list of other 'State No' values is shown. All the user has to do to merge value '3' with value '0' is choose its 'State No' value. In this case its 'State No.' value is one.

Splitting variables is just as simple. Normally the drop down list box of 'State No.' values only displays the current values, which means a split operation would not be possible. However by clicking the 'Add State' button, a new 'State No.' is created temporarily in the list. By selecting this new 'State No.' value the discrete value will be allocated to this new 'State No.', in effect splitting it into a new state.

When all the required states have been defined, press the 'Finish' button and specify the name of the new 'bif' file.

2.4 Add nodes to a new or existing network

The next step in the BBN creation process is to add the nodes defined in the BBN information file (bif) to a network. Using a 'bif' file, nodes can be created and information on the variable's states entered into these nodes automatically. The process can be performed using a data file, but as the data file contains no information on the states of variables, the process can only create and name the nodes.

On following the 'Add Nodes to a New or Existing Network' link, an open file dialogue box is shown. The default prompt is for a 'bif' file, but other data files can also be selected.

If a data file type is selected then the 'File Specification' screen is shown (see Figure G2.2) and so information on the formatting of the file can be obtained. The 'Add Nodes' screen is then displayed (see Figure G2.6).

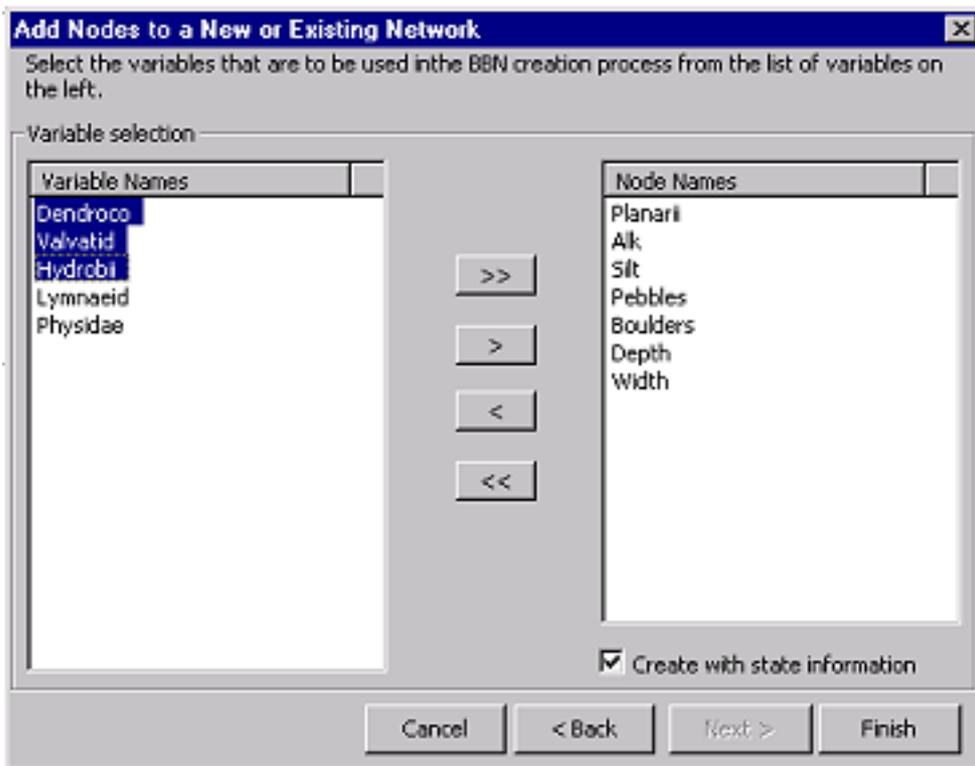


Figure G2.6 'Add Nodes' main screen.

The variables defined in the 'bif' file or data file are displayed on the left. If a data file has been used as the source of names it is possible that the names cannot be used as a node name because they contain invalid characters. In these cases the names are displayed in red and the invalid characters edited out before they can be used to create a new node. The only information required is which of the variable names are to be used to create a node with the corresponding name in the network. Individual or multiple selections of names can be moved to and from the 'Node Names' list using the '>' import and '<' export button. All the names can be moved by the '>>' import all and '<<' export all buttons.

If a 'bif' file has been used, the option to 'Create with State Information' is automatically selected, but it can be deselected if just the named nodes are required. In the case of data files, this option is not available as no state information exists.

Pressing the 'Finish' button finishes the process. A save dialogue box is shown requesting the name of the network to which the nodes are to be added. An existing network can be selected or the name of the new network entered.

2.5 Create links between nodes

The next step is to create causal links within a network. Although this is usually performed manually, in some cases where a node or nodes need to be linked to a large number of other nodes this process can become tedious and time-consuming. 'Create Links' automates this process by allowing the user to specify the nodes between which a link is required and then creating them automatically. This process does not require a 'bif' file, just a network.

If the 'Create Links between Nodes' link is open, the Network dialogue box is displayed. After the user chooses the network to which the links are to be added, the main 'Create Links between Nodes' is displayed (see Figure G2.7).

Selection of parent and child nodes between which the links will be created involves using the '>' import and '<' export buttons to move names between lists.

Once the selection of parent and child nodes has been made, the user can press 'Finish' to create the links and save the network. Alternatively if a number of linking operations are to be performed, the user can click the 'Create Links' button. This creates the currently selected links then resets the screen for further operations. When the Linking process is complete the user is prompted for a new file name for the updated network.

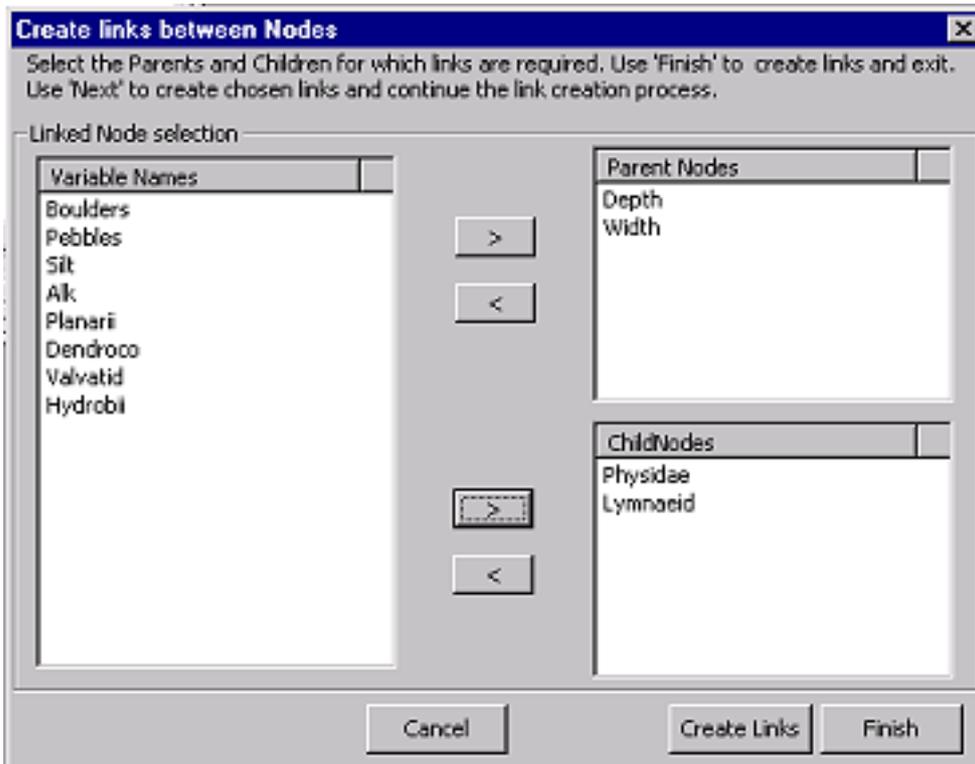


Figure G2.7 Main create links screen.

2.6 Calculate BBN probability values

Calculating the probability values for a network only requires the user to specify the network to be updated, the 'bif' associated with the network and the data file from which the probability values are to be obtained. In addition to this, the user specifies whether the probability values are to be 'raw' or 'zero-eliminated' probabilities and defines a 'M' constant if necessary.

The algorithm used to calculate zero-eliminated probabilities is based on the *m*-estimate (Cestnik, 1990). In the *m*-estimate approach, an extra *m* cases are added to the raw data distribution in proportion to the prior probability of (A_i, B_j) .

$$P'(A_i, B_j | C_k) = \frac{n_{ijk} + mP(A_i, B_j)}{N_k + m}$$

where:

C is a child node with K possible states, k = 1 to K,

A is a parent node with I possible states, i = 1 to I,

B is a parent node with J possible states, j = 1 to J,

$N_k = \sum_{i=1}^I \sum_{j=1}^J n_{ijk}$ and n_{ijk} is the number of times that C occurs in state k when A is in state I and B is in state j and m is an arbitrary constant.

Figure G2.8 Main calculate probabilities screen.

All that is required is that the user provides the paths of the network to be updated, plus the bif file and data file involved. The user can specify the path by typing the value into the box or using the browse button. If the zero-eliminated probabilities option is selected, it is possible to supply a 'm' value in the box provided. If not, a default value is used.

When the required information has been supplied, the 'File Specification' screen is displayed to obtain information on the formatting of the data file. Clicking 'Finish' on this form causes the processing form to be shown displaying the progress of the calculation process. The network file selected is simply updated with the probability values so there is no 'Save as' prompt.

2.7 Test network

Testing a network requires three separate sets of information. First the input files: network, bif and data files for the specific BBN. Second, the test input is required, that

is, the nodes that are going to have evidence entered from the dataset. Third, the type of test output and the output files must be defined.

Again, the process is simple and uses screens similar to those used elsewhere in BBN Creator. The form specifying the input file is almost the same as the 'Calculate Probabilities' screen. The path names for the files are required and the user can type the value in or use the browse button (see Figure G2.9).

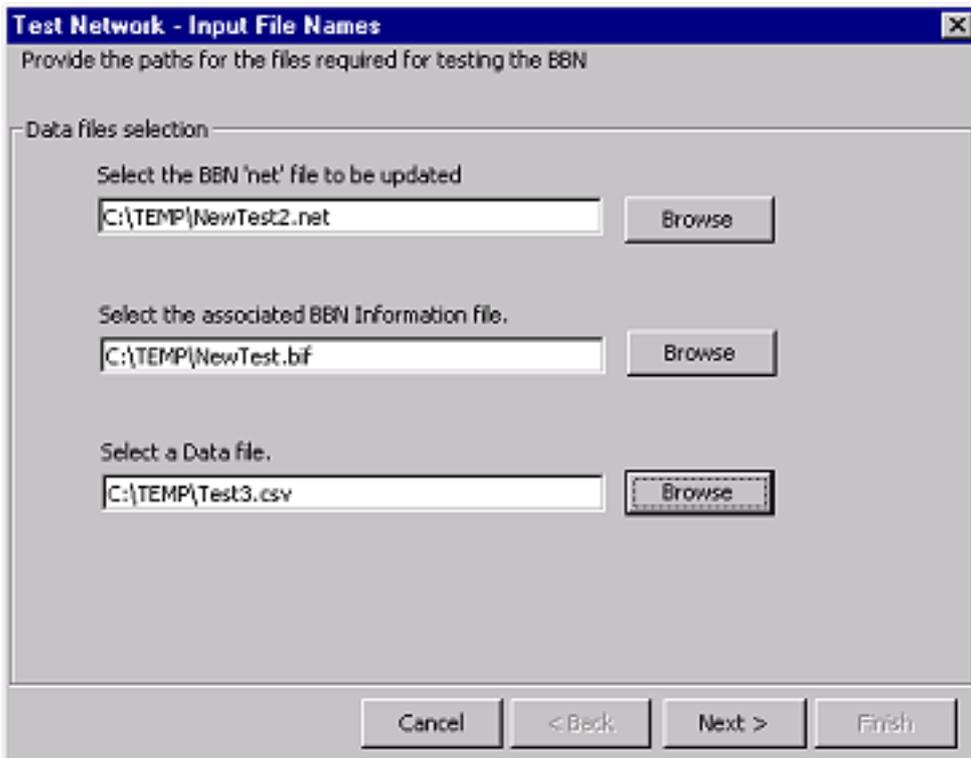


Figure G2.9 First screen of the test network process.

After pressing 'Next', the 'File Specification' screen is shown to obtain information on the formatting of the data file. The user is then required to enter the nodes that are to receive input during testing. This is done in the next screen 'Select Test Input Nodes' which uses a similar interface to other screens in BBN Creator. The names of the intended input nodes need to be imported into the 'Input Node Names' list. This list can be updated and changed via importing and exporting using the '>' and '<' buttons (see Figure G2.10).

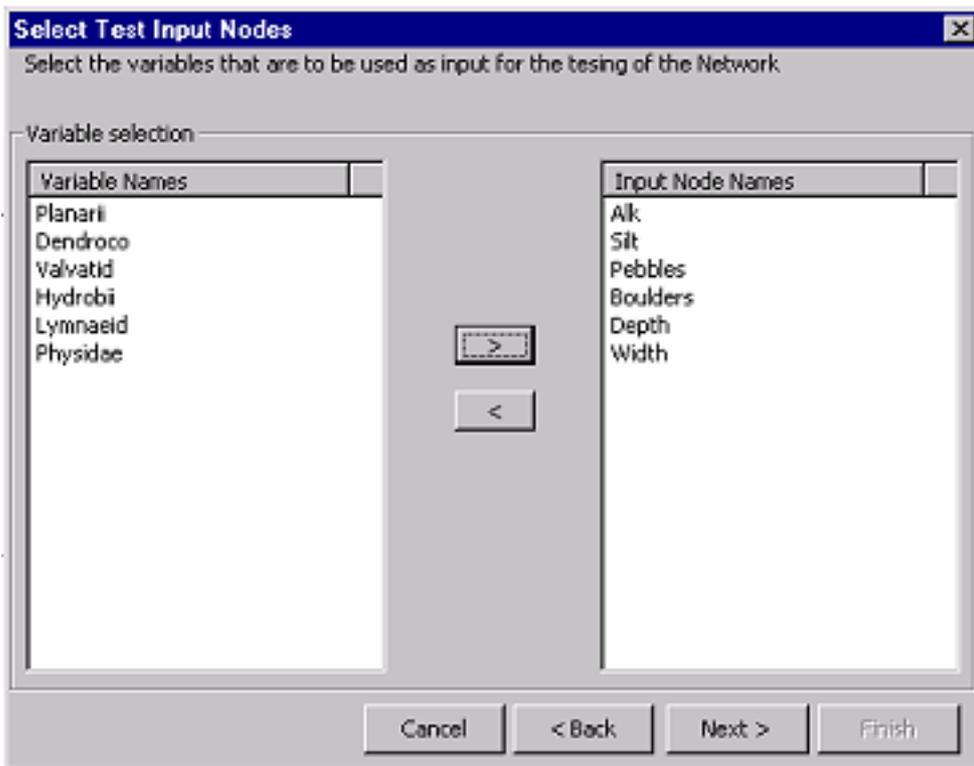


Figure G2.10 Select test input nodes screen.

Once the test input nodes have been selected, the next stage is to specify the types of result to record and the files in which to store the output values.

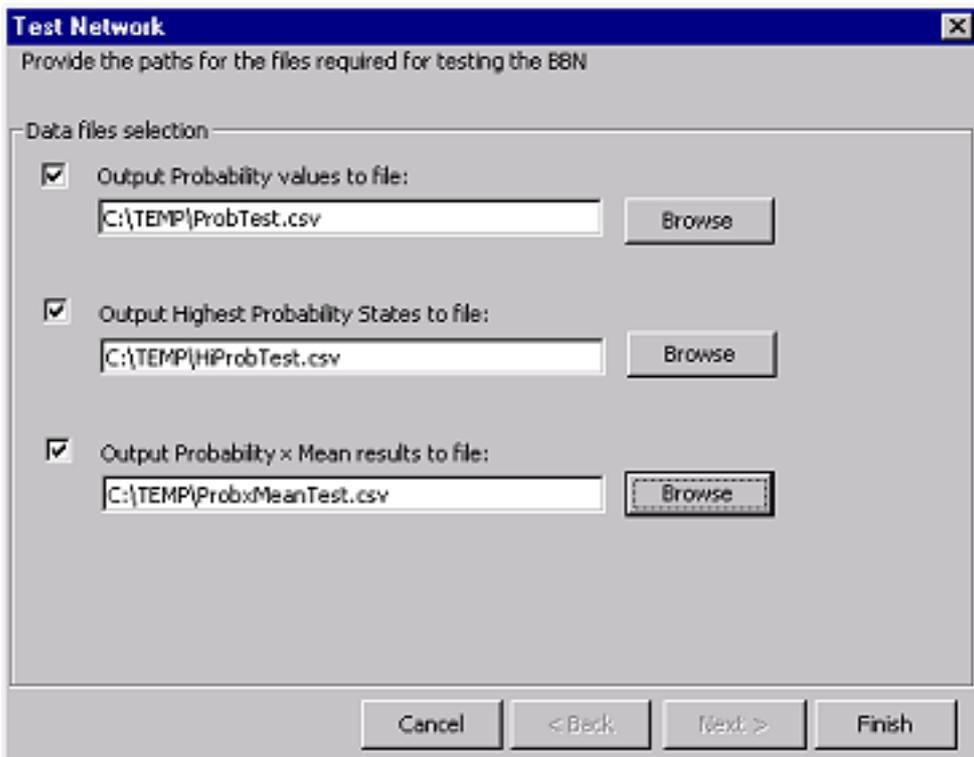


Figure G2.11 Specification of output type.

There are three main types of test result:

1. Probability values for each state of the result nodes.
2. A predicted state based on the highest probability.

3. A predicted value created by multiplying the probability value of each state by the mean value of the state obtained from the training data.

Note: The predicted value using the Probability × Mean method is only valid for continuous type variables.

The name and path of the new output file are then required. When complete, clicking 'Finish' shows the processing screen with the progress of the testing procedure.

2.8 Calculate mutual information ranking file

Calculating a mutual information ranking file is entirely separate from the BBN creation process. Its purpose is to provide a ranking metric by which the value of linking nodes can be assessed so that the automated linking algorithm has criteria on which to make a decision. As such, any other method which is able to rank the value of relationships between variables, that is multiple regression, can be used. The output of this process is a matrix of mutual information values for the variables selected for comparison. The columns contain the values for 'parent' variables, the rows the values for 'child' variables.

	1	2	3	4	5	6	7
1	Names	Alk	Silt	Pebbles	Boulders	Depth	Width
2	Physidae	0.0372	0.0367	0.0127	0.0411	0.0279	0.0067
3	Lymnaeid	0.0348	0.0281	0.0080	0.0311	0.0153	0.0029
4	Hydrobii	0.0171	0.0059	0.0199	0.0165	0.0107	0.0082
5	Valvatid	0.0853	0.0298	0.0076	0.0378	0.0340	0.0099
6	Dendroco	0.0154	0.0008	0.0024	0.0064	0.0015	0.0021
7	Planarii	0.0171	0.0065	0.0087	0.0087	0.0029	0.0103

Figure G2.12 Example mutual information ranking file.

The process involves two stages. Firstly, the selection of a 'bif' file (which is used for the state data it contains) and a data file and secondly the parent and child variables to assess.

The first screen is a standard BBN creator file selection form. Path and file names can be typed in or browsed for (see Figure G2.13).

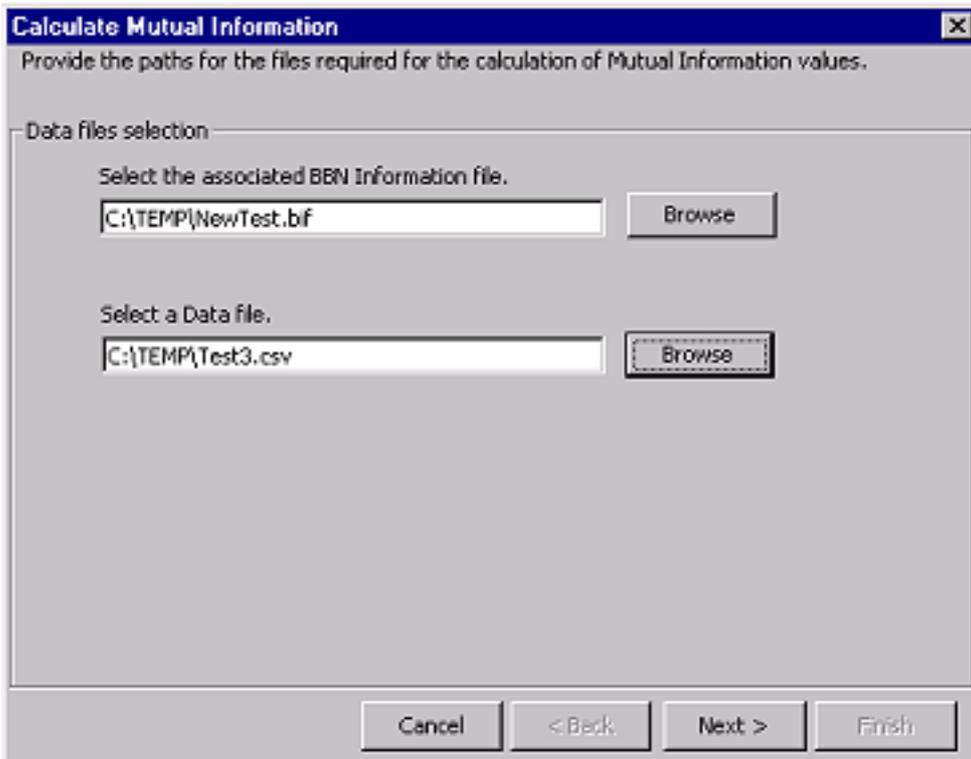


Figure G2.13 Calculate mutual information ranking file input screen.

The next screen 'File Specification' is used to obtain information about the formatting of the data file. Moving on again to the 'Next' screen, a standard BBN Creator variable selection screen is shown (see Figure G2.14).

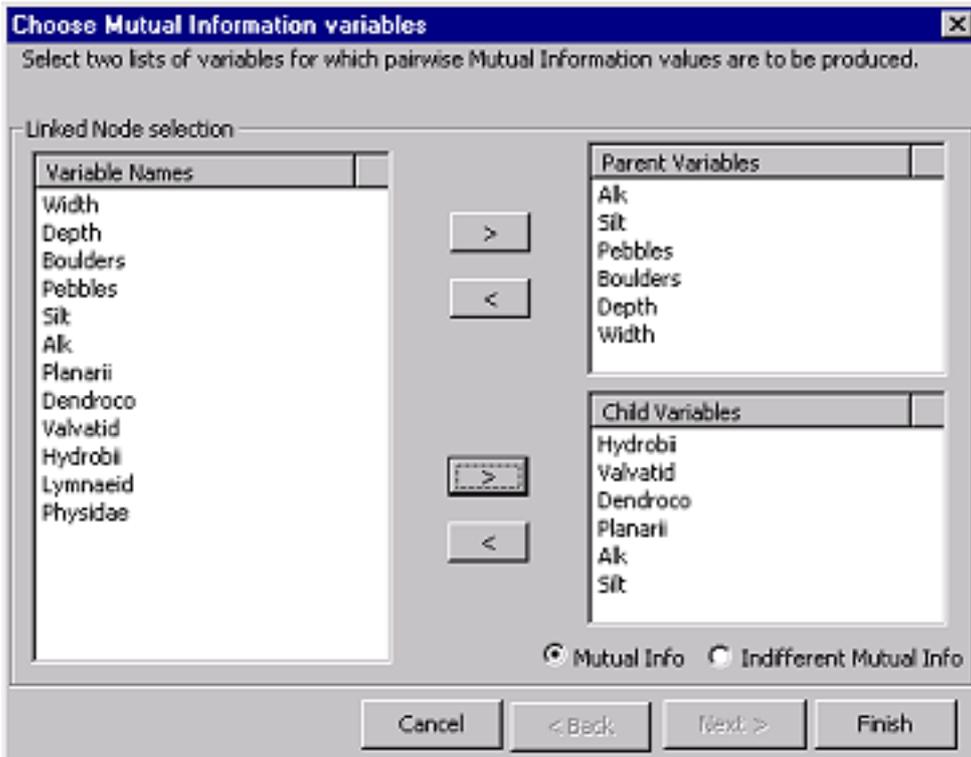


Figure G2.14 Mutual information ranking variable selection screen.

Values are not removed from the 'Variable Names' list, which allows an overlap between the parent and child selection. Pressing the 'Finish' button displays a 'Save As' dialogue box asking for a name for the new mutual information ranking file. Finally, the processing screen is shown with the progress of the procedure.

2.9 Automatically create links between nodes

This process automatically creates links between nodes based on a ranking file. The process involves two stages and requires three pieces of information. The first stage is specification of the files involved which are network, bif and ranking file. The second stage is designed to obtain the three pieces of information required:

1. List of potential parents and children involved in the process.
2. Number of links to be created between parents and children.
3. Whether the process should optimise links for predicting parents or children.

The first screen displayed is a standard BBN creator file selection form. Path and file names can be typed in or browsed for (see Figure G2.15).

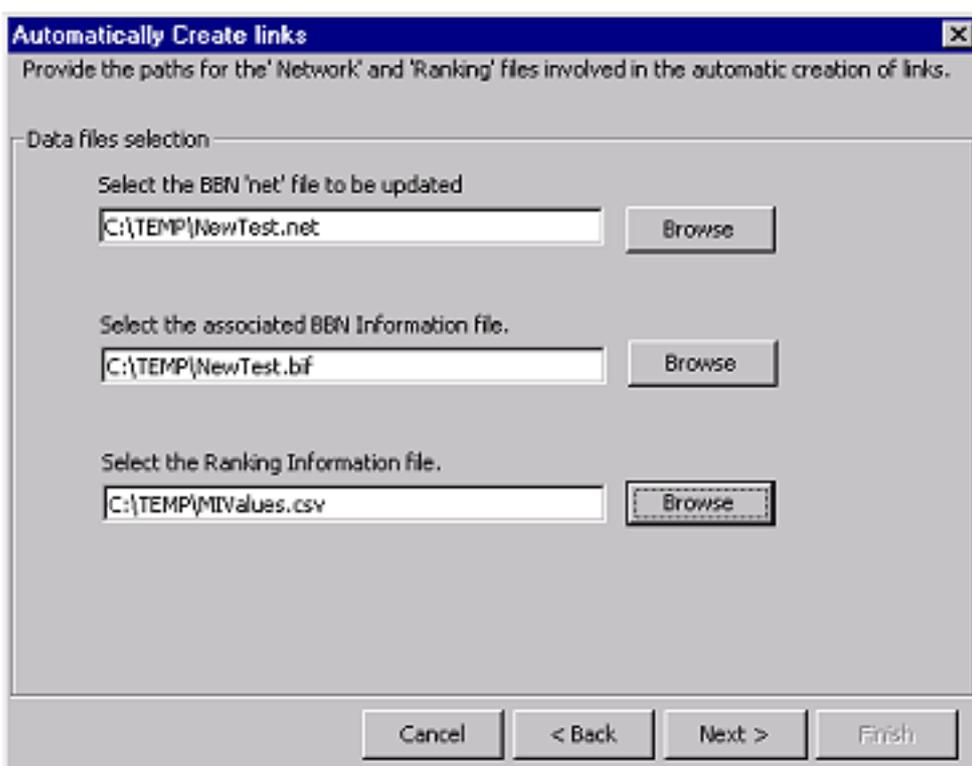


Figure G2.15 Automatically Create Links file input screen.

Once the files are specified the 'Next' screen displayed is a variation on the standard BBN Creator variable selection screen. In this case the parent and child variables specified in the ranking file are displayed on the left. Selecting these variables mean that the information will be used to link the corresponding parent or child node in the network. The screen also provides a box to specify the number of links to be created and option boxes for the optimisation of links, whether for predicting the parent or the child nodes (see Figure G2.16).

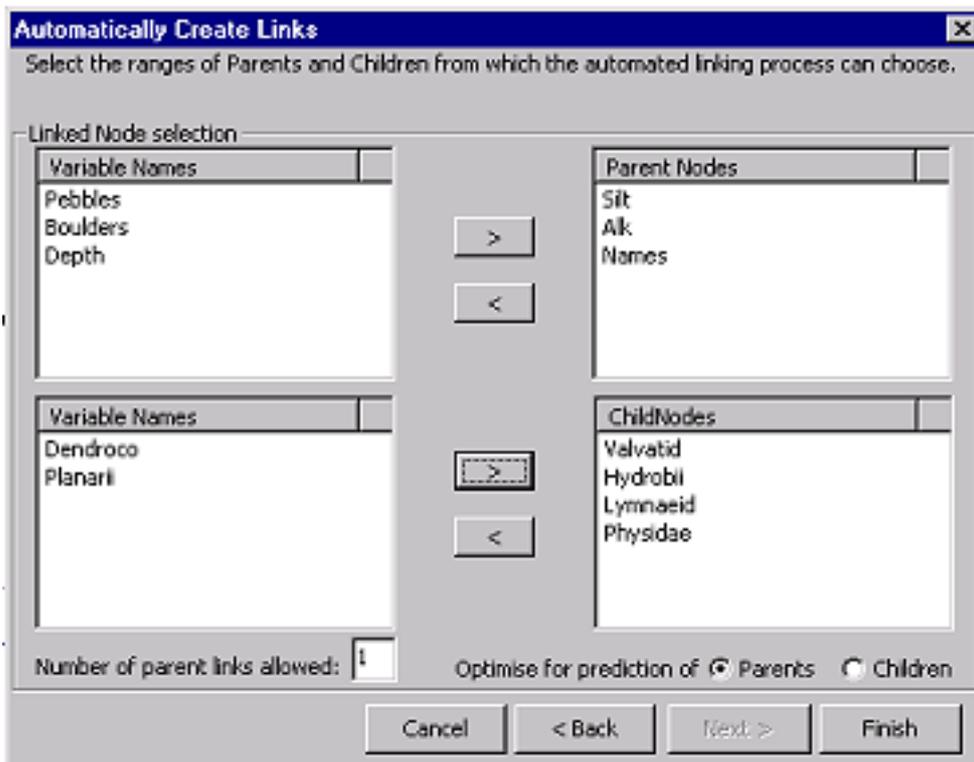


Figure G2.16 Specify parameters for Automatically Create Links process.

Clicking on the 'Finish' button shows the processing screen with the progress of automatically creating links procedure. There is no 'Save As' dialogue box as the network involved is updated and saved.

4. File Formats

4.1 General

All data files created by BBN Creator are comma-delimited files. The files can be opened and edited in a text editor (such as Notepad) or a spreadsheet application (such as Excel), although editing by hand is not recommended, especially in the case of the bif files.

4.2 BBN Information Files (bif)

Comma-delimited (.bif) files.

The format depends on the type of variable, whether it is continuous or discrete.

Continuous format:

```
Name, Type Name, State Count, Minimum Value, State Bound 1, State Bound 2, ...  
State Bound x, Maximum Value, State Mean 1, State Mean 2, ... State Mean x,  
State Median 1, State Median 2, ... State Median x
```

Discrete format:

```
Name, Type Name, State Count, State No 1, State No.1 Value Count, Discrete  
Value 1, Discrete Value 2, ... Discrete Value x, State No 2, State No.2 Value  
Count, Discrete Value 1, Discrete Value 2, ... Discrete Value x, ... etc
```

4.3 Network Files

The system used the HUGIN API to handle BBN. HUGIN uses 'hkb' and 'net' file formats. Discrepancies between the proprietary 'hkb' format used by different versions of HUGIN had caused problems previously so the 'net' file format was used. This file format saves information about the BBN in plain text and appears to be version independent, plus plain text format allows manual editing. More information on the 'net' format can be obtained from the HUGIN website at <http://www.hugin.com>.

4.4 Ranking files

Comma-delimited (.csv) files

```
<Ignored Value>, Parent Name 1, Parent Name 2, Parent Name 3, ... Parent Name x  
Child Name 1, Value 1, Value 2, Value 3, ... Value x  
Child Name 2, Value 1, Value 2, Value 3, ... Value x  
Child Name 3, Value 1, Value 2, Value 3, ... Value x  
...  
Child Name x, Value 1, Value 2, Value 3, ... Value x
```

5. Acknowledgements and Contact Details

5.1 Acknowledgements

BBN Creator was written by David Trigg at Staffordshire University's Centre for Intelligent Environmental Systems (CIES), using Microsoft Visual Basic 5. The software uses the HUGIN ActiveX API versions 1.0 to implement BBN. Thanks are due to the Environment Agency and Staffordshire University for their support, and to Professor William Walley (Centre Head), Mark O'Connor (Centre Manager) and Ray Martin for their work on this and linked projects.

5.2 CIES - Centre for Intelligent Environmental Systems

The Centre for Intelligent Environmental Systems is a centre within Staffordshire University's School of Computing. The Centre specialises in the application of advanced computing techniques, especially Artificial Intelligence (AI), to problems affecting the natural environment. Projects to date have concentrated on the development of intelligent systems for biological monitoring of river quality. The centre's expertise in this field has grown out of the pioneering work carried out by Bill Walley and Bert Hawkes in the early 1990s. Although biomonitoring will remain the principal application domain of the group, some diversification into other environmental applications is envisaged. To find out more about the Centre, visit and explore our extensive web pages at:

<http://www.cies.staffs.ac.uk>

Or, contact us at:

Centre for Intelligent Environmental Systems
Staffordshire University
School of Computing
The Octagon
Beaconside
Stafford ST18 0AD
UK

Dr Martin Paisley
m.f.paisley@staffs.ac.uk
+44 (0)1785 353510

Dr David Trigg
d.j.trigg@staffs.ac.uk
+44 (0)1785 35344

**Would you like to find out more about us,
or about your environment?**

Then call us on

08708 506 506* (Mon-Fri 8-6)

email

enquiries@environment-agency.gov.uk

or visit our website

www.environment-agency.gov.uk

incident hotline 0800 80 70 60 (24hrs)

floodline 0845 988 1188

* Approximate call costs: 8p plus 6p per minute (standard landline).
Please note charges will vary across telephone providers



Environment first: This publication is printed on recycled paper.