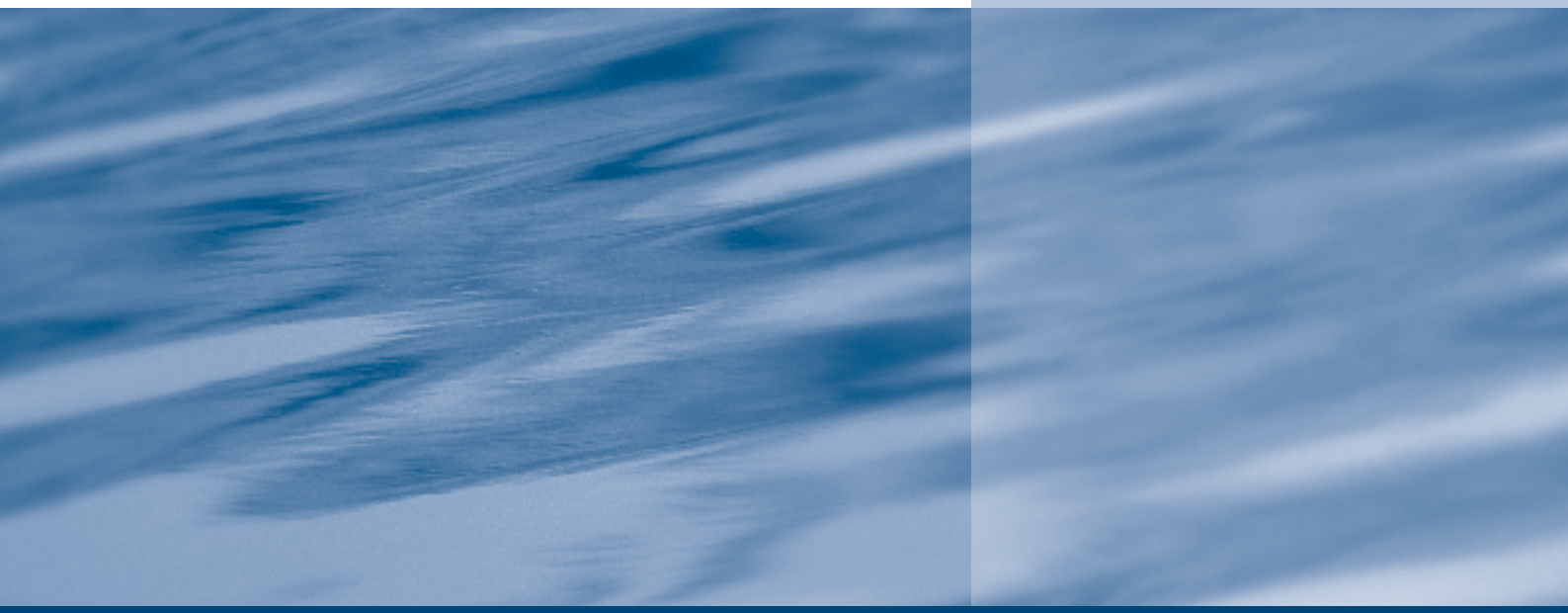

www.environment-agency.gov.uk

Further Development and Testing of Artificial Intelligence Systems for the Classification and Diagnosis of River Quality Based on Biological and Environmental Data

Science Report E1-056/SR2

ENVIRONMENT AGENCY

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's Science Programme.

Published by:
Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol, BS32 4UD
Tel: 01454 624400 Fax: 01454 624409
www.environment-agency.gov.uk

ISBN: 1 84432 368 4

This report is printed on Cyclus Print, a 100% recycled stock, which is 100% post consumer waste and is totally chlorine free. Water used is treated and in most cases returned to source in better condition than removed.

Further copies of this report are available from:
The Environment Agency's National Customer Contact Centre by emailing enquiries@environment-agency.gov.uk or by telephoning 08708 506506.

Author(s):
M.A. O'Connor, M.F. Paisley, D.J. Trigg and W.J. Walley

Dissemination Status:
Publicly available

Keywords:
Artificial Intelligence, Diagnostics, Software, River, Biology, Invertebrate, Survey, Quality, Environment

Research Contractor:
Centre for Intelligent Environmental Systems, School of Computing, Staffordshire University, The Octagon, Beaconside, Stafford ST18 0AD
Tel: + 44 (0) 1785 353510

Environment Agency's Project Manager:
John Murray-Bligh

Statement of use:
This report describes the testing and development of two computer-based systems for the prediction/diagnosis of river health previously delivered to the Environment Agency. The information in this document is for use by Agency staff and others interested in the ecological quality of rivers and systems for detecting causes of poor quality and modelling scenarios for improving that quality.

Science Project reference:
E1-056/SR2

Product Code:
SCHO1004BIPG-E-P

Science at the Environment Agency

Science underpins the work of the Environment Agency, by providing an up to date understanding of the world about us, and helping us to develop monitoring tools and techniques to manage our environment as efficiently as possible.

The work of the Science Group is a key ingredient in the partnership between research, policy and operations that enables the Agency to protect and restore our environment.

The Environment Agency's Science Group focuses on five main areas of activity:

- **Setting the agenda**: To identify the strategic science needs of the Agency to inform its advisory and regulatory roles.
- **Sponsoring science**: To fund people and projects in response to the needs identified by the agenda setting.
- **Managing science**: To ensure that each project we fund is fit for purpose and that it is executed according to international scientific standards.
- **Carrying out science**: To undertake the research itself, by those best placed to do it - either by in-house Agency scientists, or by contracting it out to universities, research institutes or consultancies.
- **Providing advice**: To ensure that the knowledge, tools and techniques generated by the science programme are taken up by relevant decision-makers, policy makers and operational staff.

Professor Mike Depledge Head of Science

Executive summary

This report outlines the results of an extension to R&D Project E1-056 (Walley *et al*., 2002), which was financed by the Environment Agency. The aims were to:

- test the computer-based systems River Pollution Diagnostic System (RPDS) and River Pollution Bayesian Belief Network (RPBBN) developed in E1-056 using data derived from the newly acquired 2000 general quality assessment (GQA) survey;
- update and improve RPDS by including a means of identifying and incorporating 'reference states' and a methodology to extend RPDS to produce a classification system;
- retrain RPDS using number of families in place of alkalinity in the input vector;
- produce an updated database to be used as input to the existing RPDS;
- identify means for updating and improving the current prototype RPBBN.

As proposed in the original specification of the extension, no changes were made to the existing RPDS and RPBBN software, as it would require considerable time, cost and effort to update the systems on the Environment Agency's network. Instead, the proposed changes remain largely theoretical. A new database has been produced for RPDS, including a retrained model with both 1995 and 2000 data. Although the database is not complete (e.g. it does not contain RIVPACS classifications of sites from the 2000 National River Quality Survey of England and Wales, or figures for feeding group composition at each site), it contains sufficient information to be used with the standard version of RPDS.

Contents

Figures

Tables

1. Introduction

1.1 Background and objectives

The work described in this report was carried out as an extension to National R&D Project E1-056, *Development of Artificial Intelligence Systems for the Classification and Diagnosis of River Quality based on Biological and Environmental Data*. The original contract resulted in the delivery of two software systems for the Environment Agency: River Pollution Diagnostic System (RPDS) and River Pollution Bayesian Belief Network (RPBBN). The development of these two systems is described in detail in R&D Technical Report E1-056/TR (Walley *et al*. 2002). Following the successful completion of the original contract, a number of further requirements were identified. The objectives of the additional work to be addressed in this extension to the project were:

1. Identify the best means of improving/updating the RPDS system to meet the needs of the Water Framework Directive (WFD). In particular, the development of:
a) a means of identifying 'reference states' and incorporating them into the model;
b) a classification system, an artificial intelligence (AI)-based equivalent of general quality assessment (GQA) that relates the actual biological condition of the river to its reference condition.

2. Retrain RPDS using number of families (NFAM) in place of alkalinity (ALK) in the input vector.

3. Identify any changes to the reported GQA quality classes that would result from the adoption of an AI-based (RPDS) classification system.

4. Test the existing systems using independent data derived from the 2000 GQA survey.

5. Update and extend the current prototype RPBBN.

It was not proposed that any changes should be made to the RPDS and RPBBN software code; rather, this extension was intended to explore possible future approaches and demonstrate the potential of further work on RPDS and RPBBN.

The work was carried out by the Centre for Intelligent Environmental Systems (CIES) in the Faculty of Computing, Engineering and Technology at Staffordshire University under the supervision of Mark O'Connor. Two research associates, David Trigg and Ray Martin, worked on the project. Mark O'Connor worked on the updating and retraining of RPDS, David Trigg on the testing of systems and updating of RPBBN, and Ray Martin on the collation and checking of data.

1.2 Summary of outcomes

The overall objectives of the project were achieved in that:

- Means of updating/improving RPDS were identified, including a means of identifying and incorporating 'reference states' and a methodology to extend RPDS to produce a classification system.
- RPDS was retrained using number of families in place of alkalinity in the input vector, and a new database was produced to be used as input to the existing RPDS.
- RPDS and RPBBN were tested using data derived from the 2000 GQA survey.
- Means for updating and improving the current prototype RPBBN were identified.

The new database resulting from the retraining of RPDS can be used with the original RPDS program. No changes were made to the existing systems (RPDS and RPBBN), and so the other updates and improvements that were identified remain largely 'theoretical'. A further project should incorporate the required changes to the software code, and more detailed testing of the theoretical possibilities for improvement of the systems. These objectives now form part of a new project, *Development of an Integrated Classification System for Rivers and Lakes*, funded by the Environment Agency's Environmental Monitoring, Classification and Reporting Project for the Water Framework Directive.

2 Data

2.1 Introduction

The project required good quality data for 2000 of the type that was provided for the original project E1-056: biological, environmental, chemical and stress data for the biological river quality monitoring sites throughout England and Wales. The most important of these were the biological data and environmental characteristics, because they were used as the input vector for the pattern recognition system developed in the original project. In order to retrain the system on 2000 data without changing the software code of RPDS, it was essential that the available data were of the same type. R&D Technical Report E1-056/TR provides a comprehensive overview of the data used in the original project (Section 2: Construction and Analysis of Project Databases, pp. 5–13). The following sections give brief details of the data that was compiled for this project.

2.2 Biological and environmental data

Each record in the biological and environmental database included the abundance category of the 76 families used in the Biological Monitoring Working Party (BMWP) score listed in Table 2.1 and the 13 environmental characteristics listed in Table 2.2 (these are the families and variables defined in BT001, Murray-Bligh (1999). The biological samples were collected and analysed according to standard RIVPACS methods used by the Environment Agency. The abundance categories used throughout the project are the same as those used previously (see Table 2.3). Table 2.4 shows the distribution of sites by Environment Agency region and season.

Table 2.1 The 76 BMWP families used in the study

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

Table 2.2 The 13 environmental variables used in the study

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} of slope (m/km)	SILT	Silt (% of substrate)
WIDTH	Average width of river (m)	ALK	Alkalinity (mg/l of $CaCO_3$)
DEPTH	Average depth of river (cm)		

Table 2.3 Abundance categories for BMWP taxa

Abundance category	Number of individuals
0	0 (none found)
1	1–9
2	10–99
3	100–999
4	1000 or more

Table 2.4 Distribution of sites used in the study by region and season. The 1995 sites were each sampled in spring and autumn, hence the numbers of sites for each season were identical. At the time the studies were undertaken, there was not an equivalent 'perfect match' between spring and autumn sites with 2000 data.

Region	1995			2000			Total	Total as %
	Spring	Autumn	Total	Spring	Autumn	Total		
Anglian	638	638	1276	621	588	1209	2485	11.0
North East	690	690	1380	650	491	1141	2521	11.2
North West	811	811	1622	672	603	1275	2897	12.8
Midlands	1033	1033	2066	959	789	1748	3814	16.9
Southern	496	496	992	475	430	905	1897	8.4
South West	1092	1092	2184	1015	887	1902	4086	18.1
Thames	484	484	968	442	446	888	1856	8.2
Welsh	795	795	1590	778	674	1452	3042	13.5
Total	6039	6039	12078	5612	4908	10520	22598	100
Total as %	26.7	26.7	53.4	24.8	21.7	46.6	100	

2.3 Chemical data

For those biological GQA sites in the database that could be matched with chemical GQA sites, the chemical data were added to the records, in the same way as previously (R&D Technical Report E1-056, Section 2.3). Table 2.5 shows the percentage of biological sites for which chemical data were available and the coverage across each chemical determinand. The values of the chemical determinands recorded in the database were the average of the 'raw' values recorded during the three months preceding the date on which the biological samples were taken.

Table 2.5 Data coverage by year and season, as numbers of samples and percentage of total samples, for chemical data used in the study. The data is listed in order of % overall coverage.

	1995				2000				All data	
	Spring		Autumn		Spring		Autumn			
Variable	No.	%	No.	%	No.	%	No.	%	No.	%
pH value	3538	58.59	3509	58.11	5326	94.90	4649	94.72	17022	75.33
TON	3538	58.59	3509	58.11	5324	94.87	4648	94.70	17019	75.31
Temperature	3536	58.55	3512	58.16	5319	94.78	4649	94.72	17016	75.30
Ammoniacal nitrogen (non-ionised)	3517	58.24	3501	57.97	5283	94.14	4631	94.36	16932	74.93
Ammoniacal nitrogen (total)	2811	46.55	2569	42.54	5327	94.92	4649	94.72	15356	67.95
BOD	3515	58.21	3480	57.63	3795	67.62	3332	67.89	14122	62.49
Oxygen (saturation)	2580	42.72	2567	42.51	3825	68.16	3297	67.18	12269	54.29
Chloride	2478	41.03	2069	34.26	3499	62.35	3097	63.10	11143	49.31
Oxygen (dissolved)	2108	34.91	2019	33.43	3492	62.22	2994	61.00	10613	46.96
Nitrite	2487	41.18	2472	40.93	2984	53.17	2566	52.28	10509	46.50
Zinc (total)	1849	30.62	2118	35.07	3399	60.57	2890	58.88	10256	45.38
Nitrate	2404	39.81	2395	39.66	2858	50.93	2398	48.86	10055	44.50
Hardness	1941	32.14	2210	36.60	2874	51.21	2450	49.92	9475	41.93
Calcium (total)	1428	23.65	1542	25.53	3340	59.52	2817	57.40	9127	40.39
Magnesium (total)	1428	23.65	1541	25.52	3336	59.44	2817	57.40	9122	40.37
Suspended solids	1762	29.18	1278	21.16	3389	60.39	2687	54.75	9116	40.34
Phosphate	3255	53.90	3440	56.96	895	15.95	764	15.57	8354	36.97
Copper (total)	606	10.03	537	8.89	3280	58.45	2777	56.58	7200	31.86
Conductivity	1435	23.76	1674	27.72	1586	28.26	1409	28.71	6104	27.01
Copper (dissolved)	1453	24.06	1502	24.87	747	13.31	640	13.04	4342	19.21
Cadmium (total)	513	8.49	537	8.89	692	12.33	743	15.14	2485	11.00
Nickel (total)	726	12.02	349	5.78	575	10.25	485	9.88	2135	9.45
Lead (total)	508	8.41	498	8.25	574	10.23	487	9.92	2067	9.15
Magnesium (dissolved)	235	3.89	238	3.94	822	14.65	716	14.59	2011	8.90
Calcium (dissolved)	235	3.89	238	3.94	818	14.58	711	14.49	2002	8.86
Chromium (total)	485	8.03	354	5.86	594	10.58	499	10.17	1932	8.55
Lead (dissolved)	348	5.76	427	7.07	430	7.66	354	7.21	1559	6.90
Nickel (dissolved)	423	7.00	195	3.23	422	7.52	348	7.09	1388	6.14
Iron (total)	285	4.72	267	4.42	432	7.70	338	6.89	1322	5.85
Chromium (dissolved)	221	3.66	196	3.25	434	7.73	354	7.21	1205	5.33
Iron (dissolved)	194	3.21	226	3.74	394	7.02	317	6.46	1131	5.00
Zinc (dissolved)	266	4.40	182	3.01	341	6.08	274	5.58	1063	4.70
Cadmium (dissolved)	172	2.85	256	4.24	289	5.15	236	4.81	953	4.22

2.4 Stress data

No stress data were available for the GQA sites in 2000. The 1995 stress data were of generally poor quality (Martin & Walley, 2000), and so the procedures for collecting the data needed to be modified. This was the subject of a separate project; the results of this were not available in time to incorporate into this project.

3. Testing and retraining RPDS

3.1. Background

It was noted in R&D Technical Report E1-056 (Walley *et al*., 2002, Sections 5.6.1 and 5.6.2) that:

- RPDS needed to be tested on independent data. In the original project, all the 1995 data were used to train the RPDS model. It was originally intended to test this model using the 2000 data, but these were not available in time.
- The models should be retrained using the combined 1995 and 2000 data sets, to produce a more reliable version of RPDS. This would not require any alterations to RPDS itself, only a change in the data files it uses.

3.2. Testing RPDS

The original test results for RPDS, based on the 1995 data, are given in Table 3.1 (this was Table 3.3, p. 34 in Walley *et al.*, 2002). The data were split into seasons so that spring and autumn were tested separately, and then the combined 'whole year' data set tested. The correlation coefficients (Pearson, r, and Spearman rank, r_s) were calculated between the predicted values of each chemical, taken to be the mean value for the samples in a particular cluster, and the recorded values for each sample in that cluster. As mentioned in Walley *et al.* (2002) this was not an independent test, since the predicted values were derived from the averages of the recorded values. On the other hand, it was not a dependent test in the usual sense either, since RPDS is an unsupervised-learning model and does not train to fit target (i.e. recorded) data.

In order to perform an independent test, RPDS was tested using the 2000 data; that is, each of the samples taken in 2000 was classified to one of the original RPDS clusters (i.e. the model based on the 1995 data). To ensure a fair comparison with the original system, the original input vector (i.e. abundance levels of 76 BMWP taxa and the environmental variables listed in Table 2.2 excluding eastings and northings) was initially retained. The predicted chemical values for 2000, defined as above, were compared with their recorded values, and the two correlation coefficients were then calculated for all chemicals. Table 3.2 gives the results of this exercise, which was performed for both spring and autumn samples separately, and then for combined samples over the whole year. The differences between these results and those of the original tests based on 1995 data (Table 3.1) are given in italics, and in general demonstrate a slight deterioration in performance, as is expected from a truly independent test. In quantitative terms, 13 of the 34 variables produced whole year r_s values greater than 0.6 (compared with 18 previously), and 4 of these were above 0.75 (compared with 11 previously).

Although Table 3.2 is headed by iron (dissolved), with an apparently large improvement in the prediction, this must be treated as spurious because of the very small numbers of samples involved. (Other variables occur at the other end of the table with apparently large deterioration in the prediction, but they too are based on similarly small sets of data.) Temperature, the second variable in Table 3.2, was predicted poorly in the individual seasons but very well over the whole year. This is explained by seasonal influence, where the data is grouped around a set of values for each season; both sets were predicted poorly on their own,

but much better in combination when the overall range of recorded temperatures was much greater. It is pleasing to note that some key variables were predicted well, such as those associated with eutrophication (total oxidised nitrogen, nitrate and phosphate) and some of the heavy metals (magnesium and copper in particular).

For the next exercise, 1995 and 2000 data sets were combined and used as training data for MIR-max (the software underlying RPDS, see Walley *et al.*, 2002) to produce an updated model for RPDS. Comparison of predicted values with recorded values was performed again and the results given in Table 3.3. This time, 41 variables were predicted, made up of the 34 chemicals plus average BMWP-score per taxon (ASPT), BMWP score, number of families, RIVPACS GQA class, and the five-year mean values of biochemical oxygen demand (BOD), dissolved oxygen (DO) and ammonia. The first four of the additional variables are closely associated with the biological data used for the clustering, and were predicted extremely well, as expected. Not being a truly independent test, for reasons explained earlier, the correlation coefficients for the chemicals were also generally high; 19 of the original 34 variables produced whole-year r_s values greater than 0.6, although only seven were above 0.75. Some of the higher correlation coefficients were attenuated (such as the previous spurious result for dissolved iron) because they were based on more training data. The mean change in the rank correlation coefficient for the data taken for the whole year from the original RPDS model was 0.0063, a very small improvement. The coefficients for the predictions for five variables (temperature, oxygen (dissolved), iron (dissolved), iron (total) and oxygen (saturation)) improved by more than 0.1, and those for five others (zinc (dissolved), cadmium (dissolved), chromium (dissolved), ammoniacal nitrogen (total) and ammoniacal nitrogen (non-ionised)) deteriorated by more than 0.1. The combined model would be expected to perform better on independent data, although this has not been tested yet.

The next test investigated the effect of removing alkalinity from the input vector. Although alkalinity was the most important predictor of biological composition, it is influenced not only by geology but also by effluent discharges such as sewage works and farm run-off. The ambiguity between influence by natural factors and those which might be termed pollution means that alkalinity is not robust as an environmental variable. This test was designed to assess its importance to the chemical predictions. The test used the same input vector as previously with the exception of alkalinity. This test was only run on the spring data and the results are given in Table 3.4. The differences in the correlation coefficients between this test and the previous one indicate that the predictions of the biological statistics (i.e. BMWP score, ASPT, number of families, biological GQA class) were improved a little while those of the chemical variables tended to be a little worse. Of the 34 original chemical variables, the predictions of 8 were better and 26 worse, although the mean change in rank correlation was only -0.0184. Of the 26 worse predictions, the decrease in the rank correlation was less than 0.05 for 20 variables, and greater than 0.1 for just one (cadmium (dissolved)). Alkalinity itself was one of five variables with a poorer rank correlation of between 0.05 and 0.1. Thus, removing alkalinity from the training vector had a minor detrimental effect on the predictive performance overall, as expected. However, the effect on most of the variables was only very slight. The removal of alkalinity actually improved the predictions of the concentrations of certain metals (lead and zinc) that need acidic conditions to dissolve and which would be sensitive to alkalinity. This may have been because some values of alkalinity were artificially boosted by effluent discharges masking the natural levels.

The final test included ‘number of families’ in the input vector, as an indicator of taxon richness. Since pollution reduces biodiversity, inclusion of ‘number of families’ should enable RPDS to distinguish between polluted samples with few families and those with very few, and hence enhance its predictive power at the polluted end of the spectrum. ‘Number of families’ is one of two variables used by RIVPACS to derive environmental quality indices (EQIs), but is free of the subjective component of the other, ASPT, which is based on BMWP scores. The test was performed on both spring and autumn data separately and for the whole year data, and the results given in Table 3.5. As in the previous test, the predictions of the biological statistics were slightly better, while those for the chemical variables were slightly worse overall. This time, 28 of the original 34 chemicals gave slightly worse rank correlations for the whole year data and six gave better values, although the mean deterioration was only -0.0180. Of the 28 worse rank correlations, 23 deteriorated by less than 0.05, one by more than 0.1, with four in between. As with the previous test, the effect on metals may be significant: three of the six improved predictions were zinc, iron and chromium. This again suggests an improved capacity to expose heavy metal pollution.

The replacement of ‘alkalinity’ by ‘number of families’ in the input vector resulted in slightly worse predictions of most chemical variables, but this was offset by slightly improved predictions of the heavy metals. Therefore, the change had a fairly neutral effect on the predictive capability of the model, but achieved the aim of removing a possible pollutant, alkalinity, from the input vector.

Table 3.1 Results of performance tests on RPDS 2.0, showing the Pearson correlation coefficient (r) and the Spearman rank correlation coefficient (r_s) between the predicted and recorded values of 34 chemical variables, listed in order of their whole year r_s. Also shown are the numbers of samples of each variable.

	Spring			Autumn			Whole year		
Variable*	**No. Sam**	***r***	r_s	**No. Sam**	***r***	r_s	**No. Sam**	***r***	r_s
CATL	*1326*	0.838	0.861	*1300*	0.815	0.846	*2626*	0.827	0.854
ALKN	*2874*	0.837	0.846	*2874*	0.829	0.839	*5748*	0.833	0.842
HARD	*1878*	0.615	0.851	*1867*	0.674	0.829	*3745*	0.642	0.840
MGDS	*151*	0.714	0.771	*155*	0.755	0.849	*306*	0.737	0.815
CADS	*151*	0.737	0.753	*155*	0.697	0.816	*306*	0.716	0.786
TOXN	*3531*	0.722	0.773	*3535*	0.707	0.752	*7066*	0.715	0.763
ZNDS	*121*	0.568	0.788	*90*	0.555	0.718	*211*	0.551	0.760
COND	*1337*	0.624	0.792	*1308*	0.531	0.725	*2645*	0.580	0.759
NO2N	*2453*	0.585	0.762	*2444*	0.524	0.752	*4897*	0.555	0.757
MGTL	*1326*	0.694	0.759	*1300*	0.686	0.748	*2626*	0.690	0.755
NO3N	*2363*	0.700	0.762	*2359*	0.685	0.745	*4722*	0.692	0.754
CDDS	*60*	0.744	0.673	*42*	0.793	0.766	*102*	0.767	0.706
AMNI	*2794*	0.393	0.705	*2801*	0.297	0.695	*5595*	0.345	0.700
PHVL	*3531*	0.732	0.691	*3535*	0.721	0.678	*7066*	0.726	0.685
PHOS	*3248*	0.448	0.676	*3249*	0.335	0.651	*6497*	0.395	0.663
CHLO	*2440*	0.383	0.675	*2451*	0.315	0.615	*4891*	0.337	0.646
CUDS	*1361*	0.471	0.645	*1343*	0.420	0.627	*2704*	0.436	0.637
CUTL	*408*	0.604	0.627	*424*	0.576	0.622	*832*	0.590	0.627
CRDS	*78*	0.562	0.644	*64*	0.584	0.545	*142*	0.569	0.584
AMTN	*3510*	0.376	0.583	*3514*	0.335	0.559	*7024*	0.355	0.571
PBDS	*168*	0.470	0.532	*158*	0.567	0.596	*326*	0.523	0.566
BOD5	*3508*	0.459	0.579	*3512*	0.438	0.550	*7020*	0.448	0.565
PBTL	*335*	0.510	0.476	*322*	0.313	0.499	*657*	0.390	0.486
CDTL	*342*	0.703	0.445	*324*	0.748	0.526	*666*	0.728	0.485
CRTL	*319*	0.364	0.488	*308*	0.464	0.472	*627*	0.396	0.481
TEMP	*3529*	0.494	0.486	*3533*	0.414	0.411	*7062*	0.455	0.450
SUSS	*1691*	0.278	0.457	*1712*	0.338	0.440	*3403*	0.310	0.449
ZNTL	*1801*	0.412	0.461	*1800*	0.446	0.427	*3601*	0.424	0.444
NIDS	*271*	0.653	0.418	*242*	0.431	0.465	*513*	0.544	0.442
OXDS	*2047*	0.439	0.455	*2058*	0.422	0.417	*4105*	0.430	0.437
NITL	*559*	0.465	0.443	*540*	0.487	0.409	*1099*	0.474	0.427
OXSA	*2543*	0.373	0.405	*2552*	0.353	0.382	*5095*	0.363	0.394
FEDS	*79*	0.183	0.378	*62*	0.346	0.388	*141*	0.205	0.374
FETL	*151*	0.441	0.440	*131*	0.397	0.222	*282*	0.418	0.345

Variable	Spring						Autumn						Whole year			
	No. samples		Correlation		Rank correlation		No. samples		Correlation		Rank correlation		No. samples		Correlation		Rank correlation	
Iron (dissolved)	29	*-50*	0.5720	*0.3890*	0.6659	*0.2879*	8	*-54*	0.8271	*0.4811*	0.9567	*0.5687*	37	*-104*	0.6546	*0.4496*	0.7929	*0.4189*
Temperature	5258	*1729*	0.0784	*-0.4156*	0.2250	*-0.2610*	4589	*1056*	0.2579	*-0.1561*	0.2659	*-0.1451*	9847	*2785*	0.6635	*0.2085*	0.7695	*0.3195*
Alkalinity	5612	*2738*	0.7806	*-0.0564*	0.7629	*-0.0831*	4908	*2034*	0.7611	*-0.0679*	0.7409	*-0.0981*	10520	*4772*	0.7713	*-0.0617*	0.7528	*-0.0892*
Calcium (total)	1717	*391*	0.7296	*-0.1084*	0.7834	*-0.0776*	1698	*398*	0.6641	*-0.1509*	0.7189	*-0.1271*	3415	*789*	0.6972	*-0.1298*	0.7524	*-0.1016*
Hardness	2286	*408*	0.7241	*0.1091*	0.7978	*-0.0532*	2098	*231*	0.6436	*-0.0304*	0.6903	*-0.1387*	4384	*639*	0.6881	*0.0461*	0.7492	*-0.0908*
Oxygen (dissolved)	2497	*450*	0.0994	*-0.3396*	0.2154	*-0.2396*	2166	*108*	0.3649	*-0.0571*	0.4366	*0.0196*	4663	*558*	0.5263	*0.0963*	0.6848	*0.2478*
Calcium (dissolved)	187	*36*	0.6432	*-0.0938*	0.7382	*-0.0148*	170	*15*	0.4891	*-0.2079*	0.5870	*-0.2290*	357	*51*	0.5901	*-0.1259*	0.6826	*-0.1034*
Conductivity	1041	*-296*	0.4713	*-0.1527*	0.7512	*-0.0408*	1033	*-275*	0.3963	*-0.1347*	0.6556	*-0.0694*	2074	*-571*	0.4017	*-0.1783*	0.6751	*-0.0839*
Nitrate	2582	*219*	0.6138	*-0.0862*	0.6917	*-0.0703*	2192	*-167*	0.4908	*-0.1942*	0.5833	*-0.1617*	4774	*52*	0.5660	*-0.1260*	0.6499	*-0.1041*
TON	5263	*1732*	0.6236	*-0.0984*	0.6801	*-0.0929*	4588	*1053*	0.4754	*-0.2316*	0.5639	*-0.1881*	9851	*2785*	0.5637	*-0.1513*	0.6380	*-0.1250*
Magnesium (dissolved)	187	*36*	0.6446	*-0.0694*	0.7280	*-0.0430*	170	*15*	0.5181	*-0.2369*	0.4733	*-0.3757*	357	*51*	0.5792	*-0.1578*	0.6197	*-0.1953*
Nitrite	2743	*290*	0.5149	*-0.0701*	0.6749	*-0.0871*	2367	*-77*	0.2569	*-0.2671*	0.5716	*-0.1804*	5110	*213*	0.3032	*-0.2518*	0.6097	*-0.1473*
Magnesium (total)	1713	*387*	0.5156	*-0.1784*	0.6786	*-0.0804*	1678	*378*	0.3698	*-0.3162*	0.5227	*-0.2253*	3391	*765*	0.4363	*-0.2537*	0.6053	*-0.1497*
Copper (dissolved)	363	*-998*	0.4885	*0.0175*	0.6148	*-0.0302*	327	*-1016*	0.1787	*-0.2413*	0.5370	*-0.0900*	690	*-2014*	0.1890	*-0.2470*	0.5930	*-0.0440*
Phosphate	876	*-2372*	0.3957	*-0.0523*	0.6079	*-0.0681*	758	*-2491*	0.3005	*-0.0345*	0.4909	*-0.1601*	1634	*-4863*	0.3337	*-0.0613*	0.5561	*-0.1069*
pH value	5265	*1734*	0.6563	*-0.0757*	0.5946	*-0.0964*	4589	*1054*	0.5350	*-0.1860*	0.5011	*-0.1769*	9854	*2788*	0.6075	*-0.1185*	0.5520	*-0.1330*
Chromium (dissolved)	31	*-47*	0.4756	*-0.0864*	0.6374	*-0.0066*	13	*-51*	-0.3046	*-0.8886*	-0.1184	*-0.6634*	44	*-98*	0.4394	*-0.1296*	0.5273	*-0.0567*
Chromium (total)	172	*-147*	0.0015	*-0.3625*	0.5167	*0.0287*	94	*-214*	0.1069	*-0.3571*	0.5104	*0.0384*	266	*-361*	0.0363	*-0.3597*	0.5187	*0.0377*
Copper (total)	526	*118*	0.1875	*-0.4165*	0.4724	*-0.1546*	299	*-125*	0.3970	*-0.1790*	0.4343	*-0.1877*	825	*-7*	0.2534	*-0.3366*	0.4827	*-0.1443*
Chloride	3176	*736*	0.1836	*-0.1994*	0.5540	*-0.1210*	2674	*223*	0.1121	*-0.2029*	0.4603	*-0.1547*	5850	*959*	0.1038	*-0.2332*	0.4629	*-0.1831*
BOD	3746	*238*	0.4029	*-0.0561*	0.4856	*-0.0934*	3288	*-224*	0.2864	*-0.1516*	0.4052	*-0.1448*	7034	*14*	0.3383	*-0.1097*	0.4486	*-0.1164*
Iron (total)	47	*-104*	0.0952	*-0.3458*	0.0531	*-0.3869*	37	*-94*	0.3180	*-0.0790*	0.5180	*0.2960*	84	*-198*	0.1784	*-0.2396*	0.4349	*0.0899*
Lead (total)	175	*-160*	0.0748	*-0.4352*	0.3223	*-0.1537*	109	*-213*	0.0483	*-0.2647*	0.4940	*-0.0050*	284	*-373*	0.0609	*-0.3291*	0.4189	*-0.0671*
Ammoniacal nitrogen (non-ionised)	5222	*2428*	0.2152	*-0.1778*	0.5180	*-0.1870*	4571	*1770*	0.1220	*-0.1750*	0.2851	*-0.4099*	9793	*4198*	0.1541	*-0.1909*	0.4094	*-0.2906*
Ammoniacal nitrogen (total)	5008	*1498*	0.2338	*-0.1422*	0.4632	*-0.1198*	4177	*663*	0.1959	*-0.1391*	0.3935	*-0.1655*	9185	*2161*	0.1646	*-0.1904*	0.3892	*-0.1818*
Suspended solids	2538	*847*	0.1338	*-0.1442*	0.3416	*-0.1154*	1636	*-76*	0.2111	*-0.1269*	0.3287	*-0.1113*	4174	*771*	0.1548	*-0.1552*	0.3793	*-0.0697*
Oxygen (saturation)	3187	*644*	0.1549	*-0.2181*	0.1693	*-0.2357*	2757	*205*	0.3792	*0.0262*	0.4223	*0.0403*	5944	*849*	0.3790	*0.0160*	0.3667	*-0.0273*
Zinc (dissolved)	37	*-84*	0.2513	*-0.3167*	0.5112	*-0.2768*	22	*-68*	0.0201	*-0.5349*	0.3666	*-0.3514*	59	*-152*	0.1489	*-0.4021*	0.3291	*-0.4309*
Nickel (total)	191	*-368*	0.2301	*-0.2349*	0.2159	*-0.2271*	94	*-446*	0.3869	*-0.1001*	0.2300	*-0.1790*	285	*-814*	0.3340	*-0.1400*	0.2807	*-0.1463*
Zinc (total)	2360	*559*	0.0291	*-0.3829*	0.2961	*-0.1649*	2306	*506*	0.1453	*-0.3007*	0.2453	*-0.1817*	4666	*1065*	0.0675	*-0.3565*	0.2780	*-0.1660*
Cadmium (total)	221	*-121*	0.5879	*-0.1151*	0.0267	*-0.4183*	165	*-159*	0.3665	*-0.3815*	0.4786	*-0.0474*	386	*-280*	0.4377	*-0.2903*	0.2417	*-0.2433*
Lead (dissolved)	60	*-108*	0.4188	*-0.0512*	0.1291	*-0.4029*	69	*-89*	-0.0209	*-0.5879*	0.2101	*-0.3859*	129	*-197*	0.0099	*-0.5131*	0.2296	*-0.3364*
Cadmium (dissolved)	23	*-37*	-0.8663	*-1.6103*	-0.3437	*-1.0167*	23	*-19*	0.3289	*-0.4641*	0.5093	*-0.2567*	46	*-56*	0.1714	*-0.5956*	0.0077	*-0.6983*
Nickel (dissolved)	49	*-222*	0.0109	*-0.6421*	-0.2376	*-0.6556*	13	*-229*	-0.1454	*-0.5764*	-0.0087	*-0.4737*	62	*-451*	-0.0233	*-0.5673*	-0.0946	*-0.5366*

Table 3.2 Results of tests on the original RPDS 2.0 model applied to 2000 data, showing Pearson correlation coefficient and Spearman rank correlation coefficient between the predicted and recorded values of 34 variables. The numbers of samples for which a prediction was available are also shown. The figures in italics show the change from the original results (i.e. tested on 1995 data). Data is listed in order of whole year rank correlations.

Table 3.3 Results of tests on RPDS using combined 1995 and 2000 samples trained with the original input vector, showing Pearson correlation coefficient and Spearman rank correlation coefficient between the predicted and recorded values of 41 variables. The numbers of samples for which a prediction was available are also shown. Data is listed in order of whole year rank correlations.

	Spring			Autumn			Whole year			
Variable	No. of Samples	Corr.	Rank corr.	No. of Samples	Corr.	Rank corr.	No. of Samples	Corr.	Rank corr.	Diff from 1995
ASPT	6032	0.9356	0.9411	6033	0.9218	0.9290	12065	0.9303	0.9374	N/A
BMWP score	6032	0.9328	0.9356	6033	0.9236	0.9274	12065	0.9285	0.9321	N/A
Number of families	11651	0.8905	0.8865	10947	0.8831	0.8807	22598	0.8870	0.8856	N/A
Temperature	8855	0.2746	0.4121	8161	0.4733	0.4648	17016	0.7879	0.8371	0.3871
RIVPACS GQA class	6032	0.8644	0.8231	6033	0.8543	0.8149	12065	0.8594	0.8340	N/A
Hardness	4787	0.7481	0.8387	4639	0.7447	0.7998	9426	0.7488	0.8224	-0.0176
Calcium (total)	4752	0.7831	0.8260	4319	0.7284	0.7887	9071	0.7585	0.8099	-0.0441
Conductivity	2889	0.5688	0.8120	2915	0.4373	0.7920	5804	0.4804	0.8018	0.0428
Alkalinity	11651	0.8213	0.8109	10947	0.8055	0.7905	22598	0.8137	0.8014	-0.0406
Calcium (dissolved)	840	0.7892	0.7862	775	0.7651	0.7898	1615	0.7809	0.7958	0.0098
Oxygen (dissolved)	5570	0.3537	0.4280	4993	0.5040	0.5214	10563	0.6682	0.7581	0.3211
Magnesium (dissolved)	842	0.6508	0.7444	780	0.6426	0.7400	1622	0.6565	0.7495	-0.0655
TON	8862	0.7128	0.7646	8157	0.5331	0.6369	17019	0.6288	0.7146	-0.0484
Nitrate	5219	0.7105	0.7503	4731	0.5072	0.6621	9950	0.6015	0.7133	-0.0407
Magnesium (total)	4748	0.6010	0.7367	4318	0.5741	0.6702	9066	0.5885	0.7083	-0.0467
Nitrite	5438	0.5962	0.7357	4995	0.4797	0.6619	10433	0.5062	0.6967	-0.0603
Phosphate	4111	0.4639	0.6994	4179	0.5149	0.6563	8290	0.5415	0.6899	0.0269
Copper (total)	3836	0.5612	0.6924	3241	0.5476	0.6793	7077	0.5591	0.6888	0.0618
BOD (3-year mean)	3576	0.5373	0.6509	3565	0.5609	0.6661	7141	0.5492	0.6587	N/A
pH value	8864	0.7302	0.6853	8158	0.6617	0.6170	17022	0.7016	0.6584	-0.0266
Chloride	5928	0.2902	0.6649	5097	0.2463	0.6280	11025	0.2615	0.6493	0.0033
Iron (dissolved)	293	0.5912	0.6522	228	0.4564	0.6434	521	0.4696	0.6486	0.2746
Ammonia (3-year mean)	3576	0.4345	0.6373	3565	0.4369	0.6530	7141	0.4357	0.6469	N/A
Lead (dissolved)	438	0.5705	0.5217	478	0.4541	0.5229	916	0.4813	0.6378	0.0718
Copper (dissolved)	2024	0.5179	0.7071	1986	0.4468	0.5956	4010	0.4615	0.6348	-0.0022
DO (3-year mean)	3576	0.5549	0.5756	3565	0.5562	0.5630	7141	0.5556	0.5701	N/A
Suspended solids	5138	0.3437	0.4828	3909	0.3286	0.4746	9047	0.3575	0.5446	0.0956
Chromium (total)	808	0.3053	0.5729	581	0.4255	0.5236	1389	0.3137	0.5429	0.0619
Zinc (dissolved)	307	0.4870	0.5077	195	0.5236	0.5978	502	0.5074	0.5417	-0.2183
Cadmium (dissolved)	186	0.6999	0.3243	244	0.4603	0.6859	430	0.5815	0.5303	-0.1757
Iron (total)	451	0.3679	0.4930	297	0.5127	0.5530	748	0.4156	0.5159	0.1709
Oxygen (saturation)	6405	0.3670	0.3796	5864	0.5236	0.5385	12269	0.5202	0.5154	0.1214
BOD	7027	0.4058	0.5208	6423	0.3822	0.4920	13450	0.3937	0.5074	-0.0576
Lead (total)	806	0.4542	0.4497	694	0.4611	0.5463	1500	0.4609	0.5048	0.0188
Nickel (total)	1020	0.5037	0.4108	561	0.6038	0.4506	1581	0.5552	0.4611	0.0341
Cadmium (total)	925	0.6324	0.3915	1013	0.4713	0.5420	1938	0.5309	0.4594	-0.0256
Nickel (dissolved)	505	0.4478	0.4247	251	0.5216	0.4465	756	0.5001	0.4524	0.0104
Chromium (dissolved)	350	0.4523	0.5659	256	0.4130	0.4102	606	0.4510	0.4485	-0.1355
Ammoniacal nitrogen (total)	8138	0.3567	0.4443	7218	0.2971	0.4695	15356	0.3346	0.4450	-0.1260
Zinc (total)	5248	0.3572	0.4217	4993	0.3665	0.4362	10241	0.3612	0.4385	-0.0055
Ammoniacal nitrogen (non-ionised)	8800	0.2709	0.3365	8132	0.2605	0.3622	16932	0.2649	0.3384	-0.3616

Table 3.4 Results of tests on RPDS using combined 1995 and 2000 spring samples trained with alkalinity removed from the original input vector, showing Pearson correlation coefficient and Spearman rank correlation coefficient between the predicted and recorded values of 41 variables. The difference in correlation from that achieved using the full original input vector is also shown. Data is listed in order of rank correlations.

Variable	No. of samples	Correlation	Difference	Rank correlation	Difference
ASPT	6039	0.9358	*0.0002*	0.9413	*0.0002*
BMWP score	6039	0.9339	*0.0011*	0.9361	*0.0005*
Number of families	11651	0.8937	*0.0033*	0.8922	*0.0057*
RIVPACS GQA class	6039	0.8686	*0.0042*	0.8408	*0.0177*
Hardness	4815	0.7117	*-0.0365*	0.7942	*-0.0444*
Conductivity	3021	0.5700	*0.0012*	0.7902	*-0.0218*
Calcium (total)	4768	0.7398	*-0.0432*	0.7798	*-0.0462*
Calcium (dissolved)	1053	0.7423	*-0.0468*	0.7680	*-0.0182*
Alkalinity	11651	0.7620	*-0.0593*	0.7463	*-0.0646*
TON	8862	0.6964	*-0.0163*	0.7457	*-0.0189*
Nitrate	5262	0.7028	*-0.0078*	0.7410	*-0.0092*
Nitrite	5471	0.5801	*-0.0161*	0.7298	*-0.0059*
Magnesium (total)	4764	0.5757	*-0.0253*	0.7164	*-0.0203*
Magnesium (dissolved)	1057	0.6251	*-0.0257*	0.7070	*-0.0374*
Phosphate	4150	0.4420	*-0.0219*	0.6872	*-0.0122*
Copper (total)	3886	0.3811	*-0.1800*	0.6825	*-0.0099*
Copper (dissolved)	2200	0.4497	*-0.0682*	0.6735	*-0.0336*
Chloride	5977	0.2987	*0.0085*	0.6710	*0.0060*
pH value	8864	0.7069	*-0.0233*	0.6481	*-0.0372*
BOD (3-year mean)	3615	0.5264	*-0.0109*	0.6440	*-0.0068*
Ammonia (3-year mean)	3615	0.4020	*-0.0325*	0.6413	*0.0040*
Lead (dissolved)	778	0.5394	*-0.0311*	0.6085	*0.0868*
Zinc (dissolved)	607	0.5433	*0.0563*	0.5971	*0.0894*
Iron (dissolved)	588	0.5642	*-0.0270*	0.5731	*-0.0791*
DO (3-year mean)	3615	0.5439	*-0.0109*	0.5635	*-0.0121*
Chromium (total)	1079	0.3779	*0.0726*	0.5419	*-0.0311*
BOD	7310	0.4171	*0.0113*	0.5321	*0.0112*
Chromium (dissolved)	655	0.4065	*-0.0458*	0.4762	*-0.0897*
Suspended solids	5151	0.3248	*-0.0189*	0.4741	*-0.0088*
Lead (total)	1082	0.4835	*0.0293*	0.4562	*0.0065*
Oxygen (dissolved)	5600	0.3658	*0.0121*	0.4292	*0.0011*
Iron (total)	717	0.3712	*0.0033*	0.4289	*-0.0642*
Ammoniacal nitrogen (total)	8138	0.3259	*-0.0307*	0.4270	*-0.0173*
Zinc (total)	5248	0.3212	*-0.0359*	0.4230	*0.0013*
Temperature	8855	0.2711	*-0.0035*	0.4086	*-0.0035*
Nickel (total)	1301	0.4441	*-0.0596*	0.3899	*-0.0209*
Cadmium (total)	1205	0.6381	*0.0056*	0.3821	*-0.0094*
Oxygen (saturation)	6405	0.3511	*-0.0159*	0.3695	*-0.0101*
Nickel (dissolved)	845	0.4059	*-0.0419*	0.3437	*-0.0809*
Ammoniacal nitrogen (non-ionised)	8800	0.2695	*-0.0013*	0.3406	*0.0041*
Cadmium (dissolved)	461	0.5827	*-0.1172*	0.1801	*-0.1443*

Table 3.5 Results of tests on RPDS using combined 1995 and 2000 samples trained with number of families included in the input vector, showing Pearson correlation coefficient and Spearman rank correlation coefficient between the predicted and recorded values of 41 variables. The numbers of samples for which a prediction was available are also shown. Data is listed in order of whole year rank correlations, and the differences in the whole year correlations (see Table 3.3) shown in the last column in italics.

	Spring			Autumn			Whole year			
Variable	No. of samples	Corr.	Rank corr.	No. of samples	Corr.	Rank corr.	No. of samples	Corr.	Rank corr.	Diffs
BMWP score	6036	0.9527	0.9551	6026	0.9428	0.9482	12062	0.9480	0.9522	*0.0201*
ASPT	6036	0.9351	0.9425	6026	0.9253	0.9327	12062	0.9315	0.9396	*0.0022*
Number of families	11651	0.9295	0.9340	10947	0.9226	0.9273	22598	0.9263	0.9311	*0.0455*
RIVPACS GQA class	6036	0.8741	0.8464	6026	0.8580	0.8324	12062	0.8661	0.8400	*0.0060*
Temperature	8855	0.2744	0.4073	8161	0.4577	0.4499	17016	0.7865	0.8340	*-0.0031*
Hardness	4797	0.7032	0.7933	4628	0.6980	0.7399	9425	0.7034	0.7700	*-0.0525*
Conductivity	2881	0.5470	0.7760	2929	0.4180	0.7625	5810	0.4603	0.7689	*-0.0329*
Calcium (total)	4754	0.7458	0.7842	4319	0.6914	0.7368	9073	0.7213	0.7635	*-0.0464*
Oxygen (dissolved)	5594	0.3593	0.4349	4977	0.5174	0.5248	10571	0.6722	0.7583	*0.0002*
Calcium (dissolved)	856	0.7394	0.7527	742	0.6762	0.7255	1598	0.7132	0.7470	*-0.0488*
Alkalinity	11651	0.7570	0.7415	10947	0.7363	0.7169	22598	0.7471	0.7299	*-0.0715*
Magnesium (dissolved)	858	0.6412	0.7173	743	0.6026	0.7006	1601	0.6250	0.7159	*-0.0336*
Nitrate	5221	0.7018	0.7382	4735	0.4867	0.6341	9956	0.5867	0.6947	*-0.0187*
TON	8862	0.6978	0.7461	8157	0.5116	0.6084	17019	0.6113	0.6918	*-0.0227*
Nitrite	5449	0.5823	0.7311	4990	0.4515	0.6449	10439	0.4812	0.6849	*-0.0118*
Magnesium (total)	4750	0.5799	0.7124	4318	0.5514	0.6382	9068	0.5667	0.6799	*-0.0284*
Phosphate	4113	0.4575	0.6910	4179	0.4843	0.6281	8292	0.5171	0.6706	*-0.0193*
Copper (total)	3852	0.4133	0.6940	3240	0.3391	0.6250	7092	0.4027	0.6680	*-0.0208*
BOD (3-year mean)	3558	0.5228	0.6486	3584	0.5401	0.6488	7142	0.5315	0.6490	*-0.0097*
Ammonia (3-year mean)	3558	0.3829	0.6471	3584	0.3850	0.6462	7142	0.3840	0.6468	*-0.0001*
Chloride	5935	0.2793	0.6504	5105	0.2374	0.6128	11040	0.2519	0.6336	*-0.0157*
Iron (dissolved)	257	0.4900	0.5798	241	0.6266	0.6447	498	0.5459	0.6170	*-0.0315*
Copper (dissolved)	2042	0.3449	0.6673	1975	0.3388	0.5669	4017	0.3413	0.6148	*-0.0200*
pH value	8864	0.7036	0.6437	8158	0.6115	0.5610	17022	0.6655	0.6063	*-0.0520*
Zinc (dissolved)	297	0.4643	0.6104	173	0.4360	0.5474	470	0.4653	0.5904	*0.0487*
DO (3-year mean)	3558	0.5416	0.5624	3584	0.5475	0.5610	7142	0.5446	0.5620	*-0.0080*
Lead (dissolved)	414	0.5619	0.4881	438	0.3485	0.5569	852	0.4886	0.5403	*-0.0975*
Suspended solids	5150	0.3338	0.4780	3911	0.3515	0.4503	9061	0.3588	0.5333	*-0.0113*
Chromium (total)	781	0.3124	0.5922	559	0.3910	0.4251	1340	0.3178	0.5298	*-0.0131*
Iron (total)	429	0.3574	0.4835	318	0.4646	0.5748	747	0.3847	0.5230	*0.0071*
BOD	6952	0.4177	0.5307	6413	0.3674	0.4910	13365	0.3918	0.5087	*0.0013*
Oxygen (saturation)	6405	0.3504	0.3742	5857	0.5034	0.5258	12262	0.5039	0.5075	*-0.0079*
Lead (total)	793	0.5006	0.4748	656	0.4052	0.5108	1449	0.4340	0.4981	*-0.0067*
Chromium (dissolved)	338	0.4505	0.5832	238	0.3910	0.3336	576	0.4487	0.4865	*0.0380*
Cadmium (total)	922	0.5331	0.4173	1003	0.4809	0.4861	1925	0.4974	0.4434	*-0.0159*
Nickel (total)	983	0.5661	0.4137	547	0.5042	0.4432	1530	0.5392	0.4372	*-0.0240*
Zinc (total)	5242	0.2789	0.4051	5001	0.3220	0.4366	10243	0.2973	0.4294	*-0.0092*
Ammoniacal nitrogen (total)	8138	0.3330	0.4272	7218	0.2888	0.4382	15356	0.3171	0.4283	*-0.0167*
Cadmium (dissolved)	182	0.7547	0.2105	224	0.5464	0.6082	406	0.5824	0.4266	*-0.1037*
Nickel (dissolved)	488	0.4398	0.4144	224	0.3771	0.3354	712	0.4163	0.4130	*-0.0394*
Ammoniacal nitrogen (non-ionised)	8800	0.2551	0.3456	8132	0.2636	0.3749	16932	0.2612	0.3561	*0.0177*

3.3 Proposed improvements to RPDS and MIR-max

A number of improvements have been suggested to enhance RPDS and its underlying algorithms, MIR-max.

3.3.1 Temporal tracking of sites in RPDS

Temporal tracking is a complex task in terms of data visualisation; the problem is illustrated for a hypothetical case in Figure 3.1. In Figure 3.1, a GQA-equivalent RPDS classification has been derived for the clusters, with quality classes 'a' to 'f' and 'reference' classes (within class 'a') for those clusters representing the very best conditions. The desired reference class for any sample can be determined through examination of the physical characteristics of the sampling site. Its corresponding reference class is the one that most closely matches the site characteristics. This then becomes the 'target' class towards which the sampled site tracks as its quality improves.

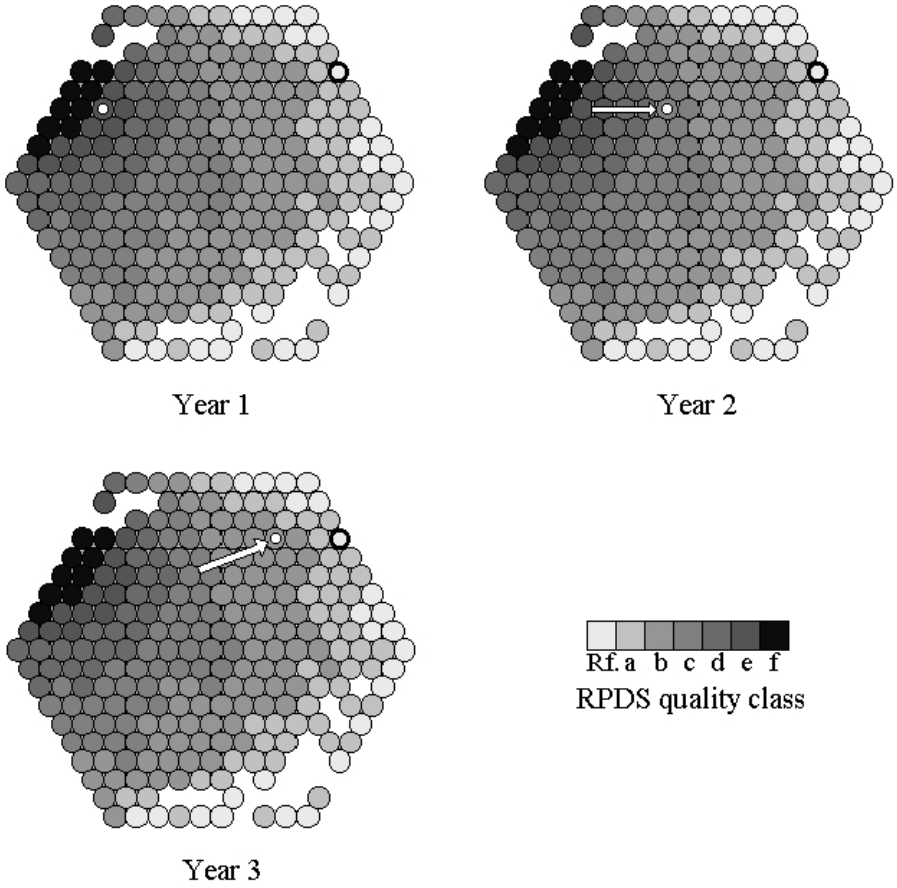

Figure 3.1 Tracking a site through time, for a hypothetical case. The desired reference class for the site (i.e. the reference class most closely matching the environmental conditions of the site) is circled in black, and its current class is highlighted by a white circle. The progression of the site across the output map towards its 'target' reference class is tracked over time as its condition changes.

3.3.2 Identification of reference condition and development of classification system

The Water Framework Directive (WFD) demands a means of identifying 'reference states' and a classification system relating actual biological condition to its reference.

Although RPDS is not based on a reference condition, it would be possible to select the 'best' RPDS clusters as representative of reference conditions by interpreting the cluster map according to predefined criteria and expert opinion. The precise mechanism for selecting these clusters would need to be decided following further research and discussion, but would probably rely on the biological and chemical characteristics of the clusters given in the 'Report' page of RPDS. By setting threshold values for key variables such as EQI(ASPT), EQI(NFAM), AMTN (total ammoniacal nitrogen), BOD, DO, PHOS (total phosphate), TOXN (total oxidised nitrogen) and heavy metals concentrations, the 'top' clusters could be identified and assigned as 'reference clusters', subject to final confirmation by a panel of experts. EQI(ASPT) and EQI(NFAM) could be derived from RIVPACS predictions of ASPT and NFAM or an approach based on AI such as the neural networks model developed by Walley and Fontama (1998). The thresholds set for the chemical variables would be the upper limits above which the cluster could not be considered to represent a reference state. The range of variables would be wide enough to prevent a change in a single variable, such as an increase in biological diversity following mild enrichment, from having undue influence. The measures used would also need to take account of the 'site type', to ensure that reference clusters exist for all possible types. The contents of each cluster would then be examined in detail to ensure that each site satisfies the requirements of a reference site. Any sites deemed not to meet the reference criteria would be removed from the reference clusters.

This method has been used to produce a 'prototype' system in MIR-max, to illustrate the possibilities. Thirty-four clusters were identified as potential reference clusters, because of the predominance of GQA class 'a' sites within them. (Note that, for this prototype, the choice was based on the average GQA class of sites within each cluster. In any 'final' system, the choice would be made using a suite of measures, taking account of differing site types.) Sites that could not be regarded as being of reference condition were then removed from these clusters – any sites that were not in the top GQA classification group were rejected. The group of reference clusters was then arranged around the perimeter of a hexagonal output map (Figure 3.2i), although this constraint need not be imposed. The reference clusters were then considered to be 'fixed'. The remaining data (i.e. all the site samples not allocated to reference clusters) were then grouped into 216 clusters using MI-max (to maintain 250 clusters total), and arranged in the hexagonal output space using R-max but under the restriction that the perimeter remained 'fixed' (Figure 3.2ii).

Figure 3.2(i) Reference clusters arranged on perimeter of MIR-max output space. (Spring samples; colour coding refers to average ASPT – bottom left of the map roughly corresponds to reference 'pools', top right to reference 'riffles'.)

Figure 3.2(ii) Non-reference clusters arranged in output space, with reference clusters on the perimeter. (Spring samples; colour coding as in Figure 3.2i)

It would also be possible to develop a classification system based on RPDS. The classification of the clusters into quality bands could be achieved at two levels. First, it could be based on EQIs in a similar way to that used in RIVPACS classifications. Second, it could be based on more detailed profiling of typical characteristics of each WFD quality class, using the approach described earlier for the definition of RPDS reference states. In this case, the basis of the classification would be the comparison of exemplar values for key characteristics of a cluster with predefined threshold values. This approach makes use of a much wider range of recorded data and could be achieved with or without the use of reference states. We consider it to be far more reliable than the RIVPACS approach, since EQIs alone provide a very simplistic model of the degree of impact. However, it could only be considered WFD compliant if it incorporated EQRs in its key characteristics.

Dealing with site types for which there is no reference state in the data would not be a serious issue since reference conditions would not be essential to the classification. Although RPDS could classify sites relative to reference states, the formation and location of clusters in the output map would not be dependent upon the existence of a corresponding reference cluster. Thus, the classification of sites with no corresponding reference state would still be possible provided the boundaries between the five WFD classes can be drawn on the output map. This would not be difficult, provided that there are sufficient clusters near to the boundaries that do have corresponding reference states. The site in question would then be classified (in probabilistic terms) to the WFD class of the cluster(s) to which it is assigned. Furthermore, this classification could be tested against a classification based on thresholds for key variables, similar to that suggested earlier for use in the definition of reference states.

3.3.3 Classification statistics

Error associated with the assignment of clusters into WFD quality bands The exemplar values for a cluster are the mean values of each variable for the samples in that cluster. Clearly, these values are prone to uncertainty, especially if the number of samples in the cluster is small, or the range of values is large. Confidence intervals (say 95%) could be given for each variable to indicate the level of confidence with which a particular cluster belongs to its allocated WFD quality class. Hence, near the boundaries of the class there may be 'grey areas' containing clusters that do not belong to the classes on either side at a given level of confidence.

Error associated with assignment of a sample to a cluster Predictions in RPDS are based on the average values of 'archive' sites belonging to the cluster to which the site under consideration is allocated. A number of measures relate to the level of 'confidence' that can be given to these predictions:

- the mutual information (MI) value achieved for the classification (i.e. the confidence with which the site has been allocated to the cluster);
- the MI achieved for other potential clusters;
- the consistency of 'archive' values within the cluster;
- the homogeneity of 'archive' values across all clusters.

Each of these measures is readily available from the information in RPDS; the difficulty is in presenting the variety of measures as a single figure that will make sense to the end-user. An overall confidence scale using percentages is proposed as the most 'user-friendly' option (possibly with a set of bandings to indicate e.g. 'very good', 'good', 'fair', etc.), with further options to examine any particular aspect (from those listed above) in more detail. Alternatively, the distance (i.e. Euclidean, or possibly an MI equivalent) of the site from its neighbouring clusters in data space could be used to provide probabilities of the sites belonging to each cluster, in a similar way to that used in RIVPACS. These probabilities would then be summed across all clusters in each WFD class, to give a probability of the site belonging to each class.

Error associated with sampling The sensitivity of a classification due to small variations in the sample composition, either from natural sources or due to sampling error, could be assessed using Monte Carlo simulation.

RPDS is a purely data-driven system based on pattern recognition rather than the rule-based approach associated with a traditional 'expert system'. It is founded on objective data analysis and not the subjective elicitation of rules from experts. This ensures that any necessary subjective input, such as expert appraisal of the 'ecological reality' of the clusters, only occurs at the end of the modelling process. Past experience has indicated that experts find RPDS output maps meaningful, and often enlightening. Justification of an RPDS classification system in terms of 'ecological reality' would therefore depend on the degree to which its classifications instilled confidence in experienced limnologists. The additional features of RPDS, such as feature maps, templates and reports should help in this regard, especially if these were to be extended and used in conjunction with the expertise inherent in a system such as RPBBN, because they provide 'back-up' evidence to support the 'basic' classification. This would ensure that an RPDS-based classification system would satisfy the needs of WFD.

In summary, RPDS has been demonstrated to be a viable alternative to classification for WFD based on the RIVPACS approach. Classification can be achieved by comparison with defined reference conditions, and appropriate confidence statistics derived, using the methodologies outlined. The main advantages of the approach are that the processes are: (i) holistic and maximise the amount of information that can be extracted from the data (unlike RIVPACS where much of the information is lost by the use of BMWP scores); and (ii) totally objective until expert opinion is used at the end to define thresholds values (unlike RIVPACS where subjective judgement enters via both the use of pre-determined reference sites and the reliance on BMWP scores).

3.3.4 Improvements to the MIR-max algorithms

O'Connor (2004) suggested the following improvements to MIR-max, and discussed the possible future development of MIR-max in more detail.

The MIR-max algorithm, like most optimisation algorithms, does not guarantee convergence to a global optimum, only a near-optimum. A possible improvement would be to allow groups of samples, not just individual samples, to change clusters during the training process. This approach may achieve a higher overall mutual information, because it is possible either for single 'true' clusters to be split across two or more MI-max classes, or for multiple 'true' clusters to be merged into a single MI-max class.

Two possible approaches to the achievement of this potential improvement are:

(1) Perform a check at intervals throughout the training process, whereby clusters could be either split or merged.

(2) Enable groups of samples to be moved 'en bloc' (e.g. on selection of a single random sample, also select those samples in the same cluster that are sufficiently 'similar'), rather than restricting the movements to single samples.

The precise way in which such solutions could be implemented efficiently would require further research.

R-max is currently based on a distance measure in data space. Ordinal biological data, such as abundance ratings, are not ideally suited to the use of distance measures, although R-max provides an acceptable solution since the ordering process is considered secondary to that of clustering. An information-theoretic version of R-max would be possible, using the mutual information between clusters in place of a distance measure. However, when only small numbers of samples are involved, the mutual information between two clusters may be difficult to justify as a reliable measure. Further research would be required to investigate the feasibility and utility of an information-theoretic approach to ordering.

4 RPBBN

4.1 Testing RPBBN using the 2000 river survey data

4.1.1 Introduction

The River Pollution Bayesian Belief Network (RPBBN) was created using the data from the 1995 River Survey of England and Wales (this data set is referred to as N2R95). At the time of development, this was the only data set of its kind in existence. Therefore the system could only be subject to either dependent testing (i.e. using the training data) or cross-validation testing, where the data set is split into parts that are used alternately for training and testing. However, neither of these options gave a true indication of the accuracy of the full RPBBN system trained using all the N2R95 data set. R&D Report E1-056/TR (Walley *et al.*, 2002) provides background on BBNs and details the initial development and testing of RPBBN.

The compilation of the data for the 2000 River Survey of England and Wales (N2R2000) provided the opportunity to perform tests on the full RPBBN system and gain some indication of its potential.

4.1.2 Aim of testing

The aim of testing was to evaluate how accurately the RPBBN system (trained using the 1995 river survey data) could predict the chemical attributes **total ammoniacal nitrogen** (AMTN), **dissolved oxygen – percentage saturation** (OXSA), **total phosphate** (PHOS), **pH** (PHVL) and **total oxidised nitrogen** (TOXN).

4.1.3 Methods of analysis

The predictions produced by BBN systems are sets of probability values that correspond to each of a variable's possible states. That is, each prediction corresponds to a probability distribution rather than a single value. This makes it necessary to perform several tests in order to extract enough information to make a reasoned judgement about the performance of the BBN. In the following tests, three different types of analysis were used:

1. The derivation of the correlation coefficient '*r*' value for the actual values of the variable and a predicted value produced by summing the product of the probability and mean values (obtained from the test set) of each state (Equation 4.1).

$$\bar{x} = \sum_{i=1}^{N} p_i x_i \qquad \text{(Eqn 4.1)}$$

   where:
   N = the number of states of the variable
   p_i = the probability of the variable being in the i^{th} state
   x_i = the centroid value of the i^{th} state.

2. The percentage of correct categorical classifications, where the categorical classification is taken to be that with the highest probability value.

3. The mean of the highest probability values for each state, for the correct categorical predictions alone and for all the predictions (i.e. correct and incorrect).

The weighted mean analysis (test 1) assesses the whole probability distribution, the correct categorical classification (test 2) assesses the most likely state, and the mean highest probability (test 3) assesses the overall confidence in the predictions.

4.1.4. Results

Table 4.1 shows both the Pearson and Spearman rank correlation coefficients of the weighted means predictions compared with those obtained previously in tests using 1995 data. The Spearman rank values are included as they usually provide a better appraisal of the weighted mean prediction because of the truncation of the predictable range.[1] The results of the categorical classification analysis are shown in Table 4.2, which details both the number and the percentage of those correct classifications.

Table 4.1 The Pearson (r) and Spearman rank (r_s) correlation coefficients for the weighted mean analysis

	AMTN		OXSA		PHOS		PHVL		TOXN	
		r_s	r	r_s	r	r_s	r	r_s	r	r_s
2000	0.2181	0.4270	0.4151	0.4015	0.4472	0.6073	0.5919	0.5558	0.5613	0.6277
1995	0.2631	0.4798	0.4327	0.4716	0.4252	0.6415	0.6405	0.6071	0.5406	0.6497

Table 4.2 The number and percentage of correct categorical classifications

AMTN		OXSA		PHOS		PHVL		TOXN	
No. Correct	% Correct	No. Correct	% Correct	No. Correct	% Correct	No. Correct	% Correct	No. Correct	% Correct
2958	29.84	2802	39.35	646	38.96	3829	38.39	4066	40.78

Figures 4.1–4.5 show confusion matrices of the categorical classification results for AMTN, OXSA, PHOS, PHVL and TOXN, respectively. These matrices show the spread of the predicted classifications in relation to their actual classifications. The matrices provide additional information on the incorrect classifications.

[1] The probability values are limited to the range [0,1] and must sum to unity. The results produced by the weighted mean equation are thus limited to values between the lowest and highest mean state value. This often leads to a clustering of predictions at the maximum and minimum values, which can have a detrimental effect on the Pearson correlation coefficient. Spearman's coefficient analyses the rank ordering of the predictions and so should in theory be less affected by clustering.

Figure 4.1 Confusion matrix for total ammoniacal nitrogen (AMTN)

Figure 4.2 Confusion matrix for dissolved oxygen – percentage saturation (OXSA)

Figure 4.3 Confusion matrix for phosphate (PHOS)

Figure 4.4 Confusion matrix for pH (PHVL).

Figure 4.5 Confusion matrix for total oxidised nitrogen (TOXN)

To provide a baseline against which to compare the certainty of the predicted probabilities, Table 4.3 shows the prior probability values for the states of the five chemical variables. Table 4.4 shows the mean highest probability values for all the predictions and Table 4.5 for the correct classifications alone.

Table 4.3 Prior probability values for each state of the five chemical variables in RPBBN

AMTN				
0–0.03	*0.03–0.05*	*0.05–0.12*	*0.12–0.295*	*0.295–33*
0.065	0.187	0.347	0.263	0.139
OXSA				
0–80.9	*80.9–91.5*	*91.5–99.8*	*99.8–107.2*	*107.2–235*
0.147	0.213	0.279	0.214	0.148
PHOS				
0–0.0275	*0.0275–0.06*	*0.06–0.24*	*0.24–0.98*	*0.98–14*
0.141	0.210	0.283	0.218	0.149
PHVL				
0–7.4	*7.4–7.7*	*7.7–8*	*8–8.2*	*8.2–14*
0.143	0.212	0.282	0.215	0.147
TOXN				
0–0.9	*0.9–2.5*	*2.5–5.9*	*5.9–9.5*	*9.5–160*
0.146	0.213	0.280	0.214	0.147

Table 4.4 The mean and standard deviation (SD) of the highest probability values for all classifications

AMTN									
0–0.03		*0.03–0.05*		*0.05–0.12*		*0.12–0.295*		*0.295–33*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.502	0.141	0.587	0.124	0.577	0.117	0.472	0.068	0.608	0.145
OXSA									
0–80.9		*80.9–91.5*		*91.5–99.8*		*99.8–107.2*		*107.2–235*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.549	0.132	0.443	0.072	0.580	0.142	0.494	0.112	0.422	0.116
PHOS									
0–0.0275		*0.0275–0.06*		*0.06–0.24*		*0.24–0.98*		*0.98–14*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.670	0.163	0.570	0.137	0.577	0.135	0.461	0.088	0.484	0.098
PHVL									
0–7.4		*7.4–7.7*		*7.7–8*		*8–8.2*		*8.2–14*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.662	0.165	0.525	0.108	0.608	0.140	0.487	0.105	0.434	0.093
TOXN									
0–0.9		*0.9–2.5*		*2.5–5.9*		*5.9–9.5*		*9.5–160*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.719	0.167	0.606	0.123	0.567	0.117	0.490	0.088	0.516	0.127

Table 4.5 The mean and standard deviation (SD) of the highest probability values for correct classifications

AMTN									
0–0.03		*0.03–0.05*		*0.05–0.12*		*0.12–0.295*		*0.295–33*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.493	0.112	0.587	0.126	0.570	0.119	0.472	0.070	0.638	0.151
OXSA									
0–80.9		*80.9–91.5*		*91.5–99.8*		*99.8–107.2*		*107.2–235*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.569	0.133	0.455	0.066	0.593	0.143	0.516	0.103	0.422	0.144
PHOS									
0–0.0275		*0.0275–0.06*		*0.06–0.24*		*0.24–0.98*		*0.98–14*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.716	0.155	0.582	0.132	0.583	0.141	0.465	0.090	0.433	0.028
PHVL									
0–7.4		*7.4–7.7*		*7.7–8*		*8–8.2*		*8.2–14*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.711	0.161	0.544	0.104	0.619	0.140	0.487	0.102	0.446	0.084
TOXN									
0–0.9		*0.9–2.5*		*2.5–5.9*		*5.9–9.5*		*9.5–160*	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.758	0.160	0.623	0.116	0.573	0.118	0.493	0.091	0.541	0.137

4.1.5. Discussion

The results produced from the testing of the full RPBBN using independent data are similar to those produced during the cross-validation tests in the original study, based on RPBBN models trained on half of these records (Walley *et al*., 2002). Although the results for the cross-validation tests were slightly better, the differences were not pronounced. Therefore, the results were again characterised by some reasonable correlation coefficients in the weighted mean analysis, a relatively poor performance in the categorical classification analysis and low means overall for the highest probability values, which indicate a fair degree of residual uncertainty in the predictions made.

The probability values for the full RPBBN were derived from the complete N2RPlus1 database (see Walley *et al.*, 2002) containing 7230 records, whereas the cross-validation RPBBN models were trained on half of these records. Therefore, it might have been expected that the full RPBBN would have performed better; the conditional probability matrices were derived from a much larger data set and so could be expected to produce better estimations of the underlying probability distribution. However, the results showed little discernible difference in the performance of the two models. Two distinct factors most probably serve to explain this:

1. *Diminishing returns.* As with many statistical methods, the improvements that can be gained by increasing the sample population from which the estimate is derived quickly decrease.

2. *Impact of Dither smoothing.* In the initial RPBBN study, it was noted that although the effect of Dither smoothing was better than previous methods it produced fairly cautious

models with improved performance in the mid-range predictions but a 'reluctance' to predict extreme events.

In combination, these two factors imply that, although the full RPBBN is based on twice as much data as the cross-validated models, the differences between the two were unlikely to be significant.

4.1.6. Conclusions

The similarity in performance between the full and cross-validated models could be interpreted as an indication of the failure of BBN to improve predictive accuracy even with additional data. However, even though the RPBBN model was tested using all the 1995 river survey data, thus overcoming the drawbacks of previous testing, various issues identified as having significant impact on performance have not been addressed by the tests undertaken in this study. These issues include:

- the development of new and improved methods of representing conditional probability distributions to reduce the difficult task of deriving the large number of probability values that are required for the probability matrices;
- the development of new and more sensitive methods of smoothing probability distributions.

If these issues are addressed (even partially) it should be possible to develop a BBN model capable of better performance than that achieved in this study.

4.2 Improvements to RPBBN

4.2.1 Viewing updated probabilities alongside previous ones

The previous version of RPBBN (version 1.2) was limited in that it could only display the current state of the variables in the system. The inability to save and display previous states of the system meant there was no simple method of visualising the differences between two states. The addition of a *Store Probabilities* feature (in the *Options* menu) rectified this problem by enabling current probabilities to be saved and stored as blue probability bars alongside the updated probability bars. Figure 4.6 shows the revised system, displaying a stored set of probabilities alongside the current probabilities of the variables. This provides a simple mechanism for comparing different scenarios.

This feature is of particular use when experimenting with 'What if?' situations, where the effects of different proposed actions (e.g. reducing ammoniacal nitrogen) on the current state of the system can be assessed. The process would simply store the state of the system associated with the sample data, and then the states of variables could be adjusted to achieve the proposed actions, for example the reduction of ammoniacal nitrogen. The system would then predict the changes that would occur as a result of the proposed action, and the new states could be quickly compared against the original states of the variables.

The storage of previous states could be extended to allow a number of them to be saved concurrently. The main problem with this proposal would be the display of more than one or

two stored states at a time, given the limited amount of space allocated to the probability bars and the potential for confusing the user. The solution to this problem may be the development of an easy to use and unobtrusive *Stored Probabilities* manager, which would allow several previous states to be stored, indexed and then selected individually for display.

Figure 4.6 Screen shot of RPBBN with the *Store Probabilities* feature in action, assessing the effects of changes in a reduction in ammoniacal nitrogen on the biological community

4.2.2 Identifying anomalies in recorded data for quality assurance or sensitivity testing

The identification of anomalies in recorded data is an important task, both as part of quality assurance and for sensitivity testing. Anomalous values are identified in RPBBN by comparing the recorded values for each variable with the predicted value for that variable made by RPBBN using the rest of the sample data as evidence. The values are defined to be anomalous if the predicted probability for the state that corresponds to the recorded value is less than the threshold value of 0.05. This value was chosen because it is a standard level of

significance used in a variety of statistical tests to prove the *null* hypothesis (i.e. that the prediction was not similar to the recorded value).

The option to enable the anomalies to be identified in the loaded records is under the *Data* menu. By default, this option is disabled as it involves a substantial amount of additional processing. Figure 4.7 shows a screen shot of this feature in action. The RPBBN software has identified the recorded value for *Ephemeridae* as anomalous. This is indicated as follows:

- the bar corresponding to the anomalous state is highlighted in the bar-chart;
- the variable is marked in bold in the listing of the recorded values in the sample.

Figure 4.7 Screenshot of RPBBN with a sample record loaded and the *Identify Anomalies* options enabled. In this example, the recorded value for *Ephemeridae* has been identified as anomalous and is indicated as such both in the bar chart for this variable (highlighted in yellow) and in the listing of the sample data (text in bold).

The identification of anomalous values could be developed further in several ways. Probably the most obvious is the development of a more sophisticated method of defining the values that are considered anomalous than the arbitrary threshold value of a probability less than

0.05 currently used. Another useful development would be a batch-processing feature that could work through an entire data file and identify the anomalous values in each sample and then produce a report at the end. This would allow RPBBN to be used as a data validation tool, although final validation of a sample with potentially anomalous data would require examination by an experienced biologist.

5 Summary and conclusions

5.1 Summary

This extension to R&D Project E1-056 has addressed the following aims:

- to test the computer-based systems RPDS and RPBBN developed in E1-056 using data derived from the newly acquired 2000 GQA survey;
- to update and improve RPDS by including a means of identifying and incorporating 'reference states' and a methodology to extend RPDS to produce a classification system;
- to retrain RPDS using 'number of families' in place of 'alkalinity' in the input vector;
- to produce an updated database to be used as input to the existing RPDS;
- to identify means for updating and improving the current prototype RPBBN.

A new database has been produced for use with RPDS, which includes both 1995 and 2000 data. The new RPDS model defined in this database has been trained with both sets of data, using 'number of families' in place of 'alkalinity' in the input vector.

5.2 Recommendations for further research and development

A number of recommendations for further research were made in the original R&D Technical Report that have not been addressed by this extension.

Fully integrated surveys. As recommended in the initial R&D Technical Report E1-056, the most valuable action that could be taken to further the development of advanced bio-monitoring systems would be to fully integrate future biological, chemical, stress and other Environment Agency surveys. It is hoped that the current and ongoing redevelopment/update of the Environment Agency's main B4W database will go some way to addressing the problems of 'matching' sites from the various surveys.

Include other types of bio-indicator. Indicators other than macroinvertebrates (e.g. aquatic plants, algae and fish) may provide valuable information, and where such data is available, it should be investigated as a possible enhancement to the current input vector.

Integration of River Biology Monitoring System (RBMS), RPDS and RPBBN. It would be useful to integrate RPDS and RPBBN into a single system to create an expert system using both pattern recognition and plausible reasoning. Ideally, this would involve some method of combining their conclusions/predictions. It would also seem sensible to include user-friendly information-retrieval facilities, as exemplified by RBMS, in the same system, to produce a single software resource for use by the Environment Agency.

Other recommendations arising particularly from this project extension are discussed in Sections 3.3 and 4.2:

- further research and development of MIR-max algorithms;
- further investigation of the use of MI and other RPDS outputs as measures of confidence;
- improvements to enhance RPBBN user interaction.

5.3 Conclusions

Conclusions from the RPDS tests

- When samples taken in 2000 were classified into the clusters based on the 1995 data, comparisons of predicted chemical variables with their recorded values indicated only a slight deterioration from the results of previous tests on the 1995 data only. Given the truly independent nature of this test, this is a convincing demonstration of the predictive capabilities of the system.
- When RPDS was trained on the combined data sets from 1995 and 2000, comparisons of predicted chemical variables with their recorded values indicated a very slight overall improvement from the results of previous tests with the 1995 data only.
- Removing alkalinity from the input vector had only a minor detrimental effect on overall performance, although predictions of some potentially sensitive metals were improved. Thus little predictive capacity was lost while removing a potential pollutant from the input vector.
- Replacing alkalinity by 'number of families' in the input vector also had a fairly neutral effect on overall predictive performance, but again predictions of some potentially sensitive metals were improved.

Updates and improvements to RPDS

- A methodology for identifying reference conditions and other WFD quality bands has been proposed, based either on a set of EQIs, or on the interpretation of the cluster map according to expert opinion. Confidence intervals could be given to indicate the level of confidence with which a cluster belongs to a particular WFD band.
- A probabilistic method for classifying a new sample has been described with probabilities based on a distance measure of the site from its neighbouring clusters. These probabilities would be summed across all clusters in each class to give the probability of belonging to each class.
- Errors associated with sampling could be assessed using Monte Carlo simulation.
- Reference clusters in the output map could be constrained to lie on the boundary as described, although distortion would be avoided if unconstrained. Temporal tracking of improvements could be easily visualised.

Conclusions from the RPBBN tests

- The results of testing the full RPBBN (trained on the whole 1995 database) with the samples taken in 2000 as independent data were similar but slightly worse than those produced during the cross-validation tests (based on splitting the 1995 database) in the original study.
- Further work is needed to develop better methods of smoothing the distributions in the conditional probability matrices.

Updates and improvements to RPBBN

- A *Stored Probabilities* option has been added, which enables previous probability bars to be shown alongside new bars. This allows comparison between different scenarios to be visualised easily.
- An *Identify Anomalies* option has been added, which compares the recorded values for each variable with predictions made using the rest of the database.

6 Acknowledgements

This study was funded by the Environment Agency. The authors are grateful to Dr John Murray-Bligh, who was the Environment Agency's Project Manager (National R&D Contract E1-056). His assistance and support throughout the project were much appreciated.

7 References

Martin, R.W. and W.J. Walley (2000) *Distribution of Perceived Stresses in English and Welsh Rivers based on the 1995 Survey: The quality of a preliminary dataset*. R&D Technical Report E126. Environment Agency, Bristol.

Murray-Bligh, JAD (1999) *Procedures for collecting and analysing macro-invertebrate samples*. Quality Management Systems for Environmental Monitoring: Biological Techniques, BT001. (Version 2.0, 30 July 1999.) Environment Agency, Bristol.

O'Connor, M.A. (2004) *The Development of a Pattern Recognition and Data Visualisation System for the Diagnosis of River Health from Biological and Environmental Data*. PhD Thesis, Staffordshire University.

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

Walley, W.J., M.A. O'Connor, D.J. Trigg & R.W. Martin (2002) *Diagnosing and Predicting River Health from Biological Survey Data using Pattern Recognition and Plausible Reasoning*. R&D Technical Report E1-056/TR. Environment Agency, Bristol.