Self-Organising Maps for the Classification and Diagnosis of River Quality from Biological and Environmental Data

William J. Walley, Raymond W. Martin, Mark A. O'Connor
Centre for Intelligent Environmental Systems, School of Computing,
Staffordshire University, Beaconside, Stafford, ST18 0DG.
Tel: +44 1785 353510 Fax: +44 1785 353497 E-mail: W.J.Walley@staffs.ac.uk

Key words: Neural networks, pattern recognition, self-organising maps, river quality, biomonitoring, RIVPACS, pollution, classification, diagnosis.

Abstract: The paper addresses the problem of how to classify and diagnose the state of health of a river from the composition of its biological community. It is claimed that experts use two complex mental processes when interpreting such data, knowledge-based reasoning and pattern recognition. It is argued that existing classification methods are inadequate and that the application of advanced computer-based techniques is vital to the realisation of the full potential of biological monitoring. The paper then concentrates on a pattern recognition approach and demonstrates how Self Organising Maps (SOM), a type of unsupervised-learning neural network, can be used to classify and diagnose river quality. A brief introduction is given to the theory of SOMs and the interpretation of their output, as expressed in feature maps and class templates. SOMs are developed using two different methods of accounting for the confounding effects of environmental factors, and their relative performances are compared. Some improvements to the SOM architecture and functionality that are currently being implemented are briefly described, together with plans to use information theory for the assessment of performance. Finally, it is concluded that the methods of classification / diagnosis described in the paper have considerable potential not only in river quality monitoring, but also in other environmental fields.

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35503-0_29

R. Denzer et al. (eds.), *Environmental Software Systems*

1. INTRODUCTION

1.1 River Quality Monitoring

River quality was traditionally monitored by chemical analysis, but increasingly this has been supplemented by biological monitoring. The reasons are twofold: a) there are now so many different chemicals polluting our rivers that chemical monitoring alone is too onerous; and b) chemical data only relate to the state of the river at the time the sample was taken, whereas biological data relate to its state over preceding weeks. Thus, polluters who discharge at midnight may escape detection by chemical monitoring but not by biological monitoring, because the flora and fauna of the aquatic community act as witnesses of their crimes. However, biological data have to be interpreted into river quality terms. Existing data interpretation systems use relatively simple algorithms based upon numeric averages or look-up. The current system used in the UK is the General Quality Assessment (GQA) classification, which is based on a development of the Biological Monitoring Working Party (BMWP) system that takes account of site characteristics using a system called RIVPACS (Moss *et al.*, 1987). A comprehensive review of various systems used in Europe was given by DePauw and Hawkes (1993). None of these systems is capable of representing the complex non-linear relationships between composition of the aquatic community and river quality, and all fail to model some of the essential characteristics of both the data and its interpretation (Walley, 1993, 1994 ; Walley and Hawkes, 1996, 1997; Walley and Fontama, 2000). For example, none of the existing systems takes account of the fact that the absence of some biota provides valuable information, or that different levels of abundance indicate different water qualities. Thus, the development of computer-based systems capable of interpreting the complex relationships involved is central to the realisation of the full potential of biological monitoring. This paper demonstrates one way in which this might be achieved.

1.2 Confounding Factors

Unfortunately, river quality is not the only factor affecting community composition. Seasonal variations and environmental factors like altitude, current velocity and the nature of the river bed also affect it. Consequently, these factors confound attempts to interpret community data into river quality terms, unless they are properly accounted for in the interpretation process. In addition, the effects of spatial isolation and extreme hydrological events (i.e. severe flooding and prolonged drought) are important factors in

some regions, but in the United Kingdom they are only of secondary importance.

1.3 Classification and Diagnosis

In the past, the main reason for routine biomonitoring was to provide data for the classification of river quality into one of *n* quality bands for the purpose of national audits. However, improvements in data collection and interpretation, together with the need to maximise the value gained from recorded data, has resulted in a drive to develop diagnostic systems. That is, systems capable of diagnosing the state of health of a river in terms of specific pollutants or types of pollutant. Such systems would enable chemical sampling and analytical effort to be directed more effectively. The overall result would be a more efficient and effective means of protecting our rivers, and added value from existing data collection programmes.

1.4 An Artificial Intelligence Approach

A new approach to the interpretation of biological data, based upon Artificial Intelligence (AI), has been under development in the United Kingdom since 1989 (Walley, 1994; Walley and Fontama, 2000). Early work with river ecologist Dr. H. A. Hawkes concluded that experts interpret data using two complementary mental processes. They use their scientific knowledge to draw a reasoned interpretation and their experience to recognise patterns that they have seen before. They then subconsciously combine these two approaches to draw a final conclusion. Walley and Fontama (2000) showed that inherent uncertainty in the meaning of the data was a key characteristic of the interpretation task, and that it has important implications for the selection of modelling techniques. They argued that Bayesian methods provide the best means of modelling the reasoning approach, and that a form of neural network, called a Self-Organising Map (SOM), is capable of modelling the pattern recognition approach. Since 1995 this work has been carried out under contract to the Environment Agency for England and Wales, and the systems presented here were developed as part of a National R&D Project that investigated potential application of AI in river quality monitoring (Walley *et al.*, 1998).

This paper describes two different approaches that were used to develop SOMs for the classification and diagnosis of river quality. The first approach accounted for the environmental factors through the definition of five site types and the development of a separate SOM for each type. In this case, each SOM had a 10×10 output array and its input consisted of biological data only. The second approach accounted for the environmental

factors by including them alongside the biological data in the input to a single SOM with a 20×20 output array. The effects of seasonal variation were eliminated in both cases through the use of combined spring and autumn samples.

2. SELF-ORGANISING MAPS

SOMs are trained using unsupervised learning. That is, they learn to identify and categorise patterns in data without knowing what the patterns represent. This is very useful in cases where the available data do not included details of their correct interpretation (e.g. a list of symptoms without their cause). This is the case in biomonitoring, because interpretations produced to date are either incomplete or unreliable.

A trained SOM allocates each input to a pattern category represented by one of the bins in the output array. However, these cannot be used for classification or diagnosis until the cases allocated to each output category have been examined and labelled by experts to indicate the specific condition they represent. The process is similar to a baby learning to recognise the faces of family members and then later learning the names to associate with them.

2.1 Structure and Function

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

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

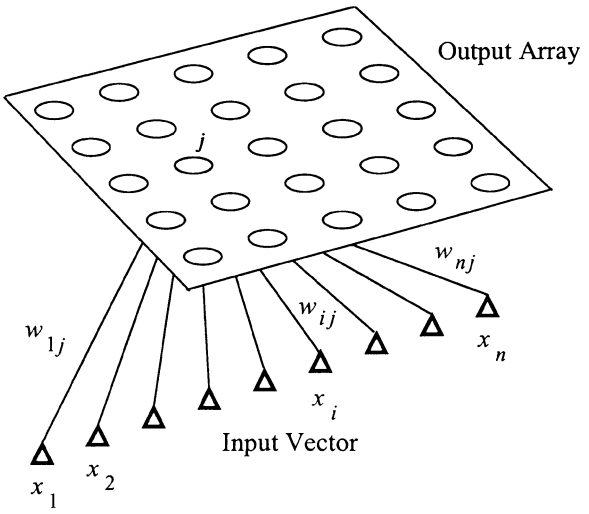

Figure 1. Topology of a Self-Organising Map with a 5x5 output array. Note that all output nodes are connected to the input vector, not just node *j* as shown.

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

$$w'_{ij} = w_{ij} + \alpha_1 \alpha_2 (x_i - w_{ij}) \qquad (2)$$

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

Both coefficients are less than or equal to unity and decay with training time. In addition, the neighbourhood coefficient decreases with distance from the winning node, thus the node's distant neighbours are modified less than its close neighbours. The neighbourhood coefficient is typically defined by a bell-shaped function with its maximum value (i.e. unity) centred on the winning node. Initially the bell is very wide, covering a large neighbourhood, but as training time proceeds its diameter gradually shrinks, thus confining the neighbourhood to an ever tightening circle around the winning node. Hence, equation (2) has the effect of gradually reducing both the size and spatial extent of the modifications as training proceeds. The final result is that neighbouring nodes represent very similar patterns and well-separated nodes represent very different ones. Readers seeking a more detailed introduction to the use of neural networks for pattern recognition are referred to the texts by Haykin (1994), Bishop (1995) and Kohonen (1997).

2.2 Feature Maps

Any element of the exemplar patterns (e.g. w_{2j}) can be plotted on the output array as a contoured map, commonly referred to as a feature map. If the input variable (or attribute) corresponding to this element is an important discriminating factor the feature map will be well defined, otherwise it will appear more like random noise. Figure 2 shows the feature maps of two input variables to a river quality SOM having a 10×10 output array (SOM10/2 described later). The maps show the abundance levels of two aquatic families, Asellidae (the water hog louse) and Heptageniidae (a mayfly). Each grid intersection represents an output category (or bin) and the contours show the variation in abundance level across the array. Note that Heptageniidae, which is indicative of good quality waters, tends to occupy a different part of the array than Asellidae, which is more indicative of poor quality waters. Both are key discriminators and therefore produce well-defined feature maps.

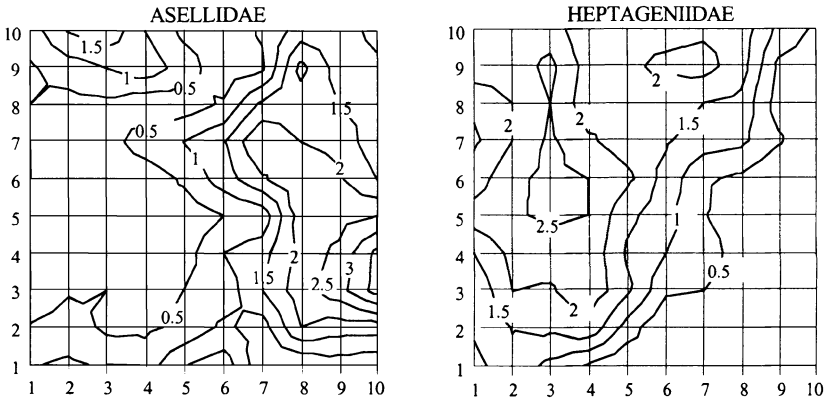

Figure 2. Feature maps for two of the input variables to SOM10/2. The vertical and horizontal scales merely serve to provide the 'x' and 'y' co-ordinates of the 100 'classification bins' located at the grid intersections.

3. THE DATA

The study was based on a validated database derived from the 1995 River Quality Survey of England and Wales. It consisted of biological and environmental data for 6038 monitoring sites covering all six GQA biological quality classes from 'a' (good) to 'f' (bad). Table 1 shows the distribution of sites with respect to GQA class and site type. The definition of the site type is given later, but basically ranges from fast flowing upland rivers (type 1) to slow flowing lowland rivers (type 5).

Table 1. Distribution of the sites by biological GQA class.

	Biological GGA Class							
Site Type	a	b	c	d	e	f	Total	Percentage
1	463	475	148	58	62	17	1223	20.3%
2	387	341	203	123	125	30	1209	20.0%
3	307	270	259	168	138	23	1165	19.3%
4	357	304	336	161	73	19	1250	20.7%
5	248	357	325	128	90	43	1191	19.7%
Total	1762	1747	1271	638	488	132	6038	100%
Percentage	29.2%	28.9%	21.0%	10.6%	8.1%	2.2%	100%	

The biological data consisted of spring and autumn samples of macroinvertebrates living in or on the river bed, and covered 76 BMWP families in all, each of which was recorded as either absent (0) or present in one of four abundance categories (i.e. 1 = one to nine individuals present, 2 = 10 to 99, 3 = 100 to 999 and 4 = over 1000). For the purpose of this study, the spring and autumn samples were combined to give a single 'combined sample' for the year as a whole. This was in keeping with the procedure used by the Environment Agency when deriving GQA quality classes. All sites had data on 13 environmental variables, namely:

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

4. SITE-SPECIFIC SOM OF RIVER QUALITY

4.1 Site Types

Initially, it was felt that the best way of accounting for environmental factors would be to divide the data into several subsets based upon the environmental characteristics of the sites, and to develop a separate SOM for each site type. Since each subset needed to contain sufficient sites to make the development of separate systems viable, the data was divide into five approximately equal subsets based upon the site's predicted 'unpolluted' average score per taxon (ASPT). The 'unpolluted' ASPT provides an indication of the type of community supported by the site in the absence of pollution. The method used to predict 'unpolluted' ASPT was similar to that

used by Walley and Fontama (1998). Full details of the method used and the geographic distributions of site types are given in Walley *et al.*, (1998).

The key physical characteristics that determined site type were found to be alkalinity, altitude and the percentage of silt in the substrate of the river (Walley and Fontama, 1998). The variations of mean alkalinity, altitude and percentage silt across the five site types are given in Table 2.

Table 2. Mean values of alkalinity, altitude and percentage silt by site type.

Environmental	Site Type (*i*)				
Variable	1	2	3	4	5
Alkalinity	34.3	95.7	167.3	206.7	221.2
Altitude	121.3	73.6	57.9	45.9	24.6
Percentage Silt	3.7	7.7	12.5	22.5	56.6

The five site types represent a transition from upland rivers and streams having low alkalinity and a substrate consisting mainly of boulders and pebbles (Type 1) to lowland rivers and streams having high alkalinity and a substrate consisting mainly of sand and silt, but especially silt (Type 5). The former are typically fast flowing 'riffles', whereas the latter are typically sluggish main river channels or 'pools' on smaller rivers / streams.

4.2 SOM10/i

Five different networks, one for each site type (*i*), were trained over 60,000 cycles using the combined spring and autumn biological data (i.e. abundance levels of the 76 BMWP families) as the input vector. A 10×10 output array was used in each case, thus giving 100 pattern categories to each site type for the classification of the inputs. Once trained, the networks were used to allocate each of the 6038 sites to one of the 100 output categories of the SOM representing its particular site type. The SOM algorithm ensured that sites allocated to neighbouring bins were similar in terms of community composition. Thus, attributes closely related to community composition (e.g. river quality, alkalinity or the abundance of a given family) varied in a relatively smooth and continuous way across each 10x10 output array, as illustrated by the two examples of SOM10/2 feature maps shown in Figure 2.

The standard deviations of each attribute within each bin were derived to provide a measure of the relative performance of the five site-specific SOMs. The average standard deviation of each attribute over the 100 output bins was then derived for each site type. This value represented the *noise* between the attribute's feature map and the samples allocated to the bins. Thus, the lower the *noise* the better the fit between the model and the data with respect to the particular attribute. Table 3 shows the *noise* levels on the

SOM10/*i* feature maps for 11 key attributes (i.e. the three key site-type variables - alkalinity, altitude and silt - and the top eight indicator families). Also given is the average *noise* level over the output arrays of all five site types.

The large variations in the average *noise* levels of alkalinity, altitude and percentage silt with respect to site type were entirely in keeping with the definition of site type. Mean alkalinity and percentage silt increase from site type 1 to site type 5, whereas altitude decreases from site type 1 to site type 5. The variation in *noise* level simply reflects the fact that the higher means were associated with higher standard deviations.

Table 3. Standard deviations of 11 key attributes averaged over the 100 output nodes of the SOM10/i feature maps for each of the five site types (i = 1 to 5).

	Site Type (*i*)					SOM10
Attribute	1	2	3	4	5	Avg
Environmental						
Alkalinity	18.58	33.60	39.06	42.47	50.98	36.94
Altitude	58.48	40.37	31.89	26.71	19.20	35.33
Percentage Silt	3.78	6.88	10.92	16.89	25.55	12.80
Biological						
Leptoceridae	0.52	0.50	0.47	0.49	0.46	0.49
Gammaridae	0.63	0.71	0.65	0.63	0.58	0.64
Elmidae	0.44	0.46	0.47	0.54	0.48	0.47
Baetidae	0.49	0.54	0.58	0.60	0.61	0.56
Caenidae	0.48	0.54	0.48	0.52	0.46	0.49
Hydrobiidae	0.61	0.74	0.70	0.67	0.60	0.66
Limnephilidae	0.54	0.53	0.57	0.57	0.54	0.55
Hydropsychidae	0.49	0.56	0.57	0.58	0.47	0.53

5. A GENERAL SOM OF RIVER QUALITY

In view of the encouraging results from SOM10/*i*, an attempt was made to develop a general SOM classifier of river quality using an input vector consisting of the combined biological and environmental data. It was felt that a SOM with a larger output array might be capable of identifying patterns in the combined data in a way that would obviate the need for prior classification of the sites into site types. Consequently a SOM with a 20×20 output array (SOM20) was trained over 60,000 cycles using an input vector consisting of the abundance levels of the 76 BMWP and numeric values of the 13 environmental variables.

The *noise* levels on the feature maps produced by SOM10/*i* and SOM20 for the 11 key attributes are compared in Table 4. The overall average ratio of *noise* levels between the two sets of data (i.e. 1.015) indicates that SOM20 was only marginally worse than SOM10/*i*. Thus it was concluded that the single classifier, SOM20, achieved a similar level of performance to

the five site-specific classifiers that made up SOM10/*i*. An interesting feature of the results shown in Table 4 is the difference in relative importance of alkalinity and percentage silt between SOM10/*i* and SOM20. Alkalinity is the most important environmental factor in SOM10/*i*, because alkalinity was the primary factor in the prediction of ASPT, and hence the site type classifications (Walley *et al.*, 1998; Walley and Fontama, 1998). However, in SOM20 the importance of silt was much increased at the expense of alkalinity. The ratios of 1.259 for alkalinity and 0.727 for percentage silt, indicate that SOM20 fits the silt data much better than SOM10, and *vice versa* for the alkalinity data.

Table 4. Average standard deviations across the SOM10/i and SOM20 feature maps of 11 key attributes. The SOM10 values are the average across the five site types, as per Table 3.

Attribute	SOM10/*i* Avg.(S1)	SOM20 Avg.(S2)	Ratio S2/S1
Environmental			
Alkalinity	36.94	46.52	1.259
Altitude	35.33	35.45	1.003
Percentage Silt	12.803	9.307	0.727
Biological			
Leptoceridae	0.489	0.498	1.018
Gammaridae	0.640	0.634	0.991
Elmidae	0.474	0.486	1.025
Baetidae	0.564	0.573	1.016
Caenidae	0.493	0.511	1.036
Hydrobiidae	0.663	0.692	1.044
Limnephilidae	0.551	0.562	1.020
Hydropsychidae	0.535	0.550	1.029
		Average	1.015

The benefit of being able to classify the river quality of a site directly from its biological and environmental data, without first having to classify its site type, was seen as a very useful feature of SOM20.

6. INTERPRETING THE OUTPUT

The output array of a SOM not only classifies input patterns to one of *n* bins, each representing different characteristic patterns in the data, but also arranges them so that neighbouring bins represent very similar patterns. The former is an essential part of classification and diagnosis, whereas the latter is an added benefit that permits effective visualisation of the data and possibly a better understanding of the underlying system. In fact, this produced an unexpected spin-off benefit in this study, since it was found that by comparing the feature maps of different attributes much could be learned about the environmental requirements of the different families. Thus the

feature maps were found to provide a valuable tool for use in basic biological research. Software (called SOM Viewer) was developed to facilitate the comparison of feature maps. This is now available on our home page (http://www.soc.staffs. ac.uk/research/groups/cies/).

However, classification and diagnosis can only be performed once the output bins have been labelled in terms of the conditions they represent. For a SOM with a large output array, like those in this study, the task of labelling the bins is substantial and may require considerable input from experts in the particular domain. In the case of river quality, the pattern associated with each bin (i.e. its biological community and site characteristics) has to be interpreted into water quality terms. If the SOM is to be used for diagnosis, the specific pollutants or types of pollutant, if any, that are responsible for a particular pattern have to be identified and its bin labelled accordingly. The labelling of a SOM for classification purposes alone is far less demanding, since it only requires that the patterns are allocated to one of a relatively small number of quality classes (e.g. the UK uses six GQA quality classes).

6.1 Use of feature maps and class templates

The task of labelling the bins is made easier by the use of feature maps and class templates. Feature maps enable us to examine how each attribute (e.g. the occurrence of a given taxon) varies across the output array, whereas class templates provide a visual image of the pattern represented by each bin. Figure 3 shows partial class templates for three of the bins in the output array of SOM10/2. The templates represent three very different river qualities. Template (a) indicates the presence of pollution sensitive creatures like Ephemeridae, Caenidae and Perlodidae and relatively few Asellidae, which tend to thrive in organically polluted waters. It represents bin (3, 5) and can be labelled "GQA class a - Unpolluted". Template (c) indicates an abundance of the families that are highly tolerant of organic pollution (Oligochaeta, Asellidae and Chironomidae), relatively few of the fairly tolerant families (Hydrobiidae, Glossiphoniidae, Gammaridae and Baetidae) and a total absence of the pollution sensitive families. It represents bin (10, 3) and can be labelled "GQA class e - severe pollution, probably organic". Template (b) indicates a community that lies somewhere between those of templates (a) and (c). It represents bin (7, 4) which lies almost midway between bins (3, 5) and (10, 3), and may be labelled "GQA class c - moderate pollution, probably organic".

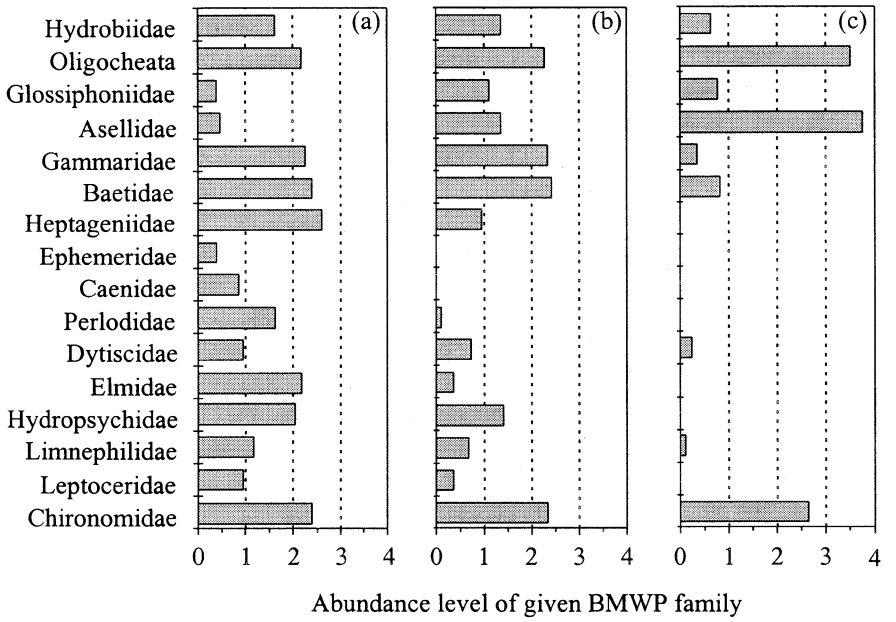

Figure 3. Pattern templates for three of the bins in the output array of SOM10/2. In the interest of brevity only 16 of the 76 families are shown. The bins represented are: a) node (3,5); b) node (7,4) and c) node (10,3) in the output array shown in Figure 2.

Although this level of pattern identification is adequate for the classification of river quality, much more detail is required if the system is to be used to diagnose the cause of any degradation of river quality. For this we need to identify patterns in the biological community that are characteristic of specific pollutants. This is clearly possible for moderately polluted conditions, because expert river ecologists are able to do so. However, there is little prospect of identifying the specific cause of severe pollution from biological data, because under these conditions, whatever the cause, little life remains on which to base a diagnosis. On the other hand, the diagnosis of severe pollution is not our prime objective. The majority of rivers are of good or moderate quality and the diagnosis of problems arising in these rivers is a major concern of river managers, and hence the prime objective of this work.

The systems described above are capable of classifying river quality, but are not yet capable of diagnosing the causes of pollution. This aspect of the project, being a substantial task, has been postponed until after planned improvements to the function and architecture of the SOM have been made. These developments form the basis of National R&D Project E1-056 of the Environment Agency, which is currently in progress.

7. CURRENT AND FUTURE DEVELOPMENTS

7.1 Improvements to the SOM

Two improvements to the SOM's architecture and functionality are planned. Firstly, the form of the output array is being extended to permit the use of a triangular grid in addition to the current rectangular one. This will permit much greater flexibility in the choice of output topology, allowing the use of triangles, hexagons, diamonds, stars as well as squares and rectangles. Thus an output topology which best suits the data will be more easily achieved with the new architecture, and in addition the triangular grid will allow neighbourhoods to be defined on a purely radial basis during the training phase. Secondly, a new similarity metric is being developed which permits greater emphasis to be placed upon the most discriminating variables and less on the least discriminating ones. The new similarity metric is based upon the Euclidean distance but the scale of each axis is modified to match the relative importance of the variable it represents, thus increasing the influence of important variables and reducing that of weak variables. Once these improvements have been implemented and tested, a new set of SOMs with improved output topologies will be trained.

7.2 Adding The Diagnostic Capability

In order to facilitate the labelling of the output bins in sufficient detail to make the SOMs useful diagnostic tools, additional data is being gathered on the pollutants and perceived environmental stresses that affected the 6038 sites during the 1995 survey period. Firstly, the regional biologists who surveyed the rivers in 1995 are giving their subjective assessments of the pollutional and / or environmental stresses that affected the sites. These cover a wide range of stresses from sewage effluent to various agricultural and industrial pollutants, motorway run-off, engineering works, weed control and the effects of drought. Secondly, the chemical data held by the Environment Agency are being matched as closely as possible, on a site to site basis, so as to provide as much data as possible on the chemical nature of the sites prior to the taking of biological samples in 1995. With these data at hand, it will be possible to produce feature maps and class templates for the chemical and stresses data in addition to the biological and environmental input variables. This should enable a more precise identification to be made of the water quality condition represented by each output bin. Once this has been achieved the system will be tested under operational conditions by regional biologists. Final refinement of the labels

by experts, with the benefit of feedback from the field test, should then be a far less onerous task than was originally thought.

7.3 Use of Information Theory

Assessing the performance of systems based on supervised learning is fairly straightforward because the system's outputs can be compared directly with a set of target outputs (i.e. 'correct' answers), but in unsupervised learning this is not possible since there are no target outputs. The method based on *noise* levels that was used to assess the relative performance of the SOMs, although adequate for its purpose in this study, is not suitable for the more demanding tasks that lie ahead. Information theory offers a more theoretically sound approach to questions like "Which output topology is best suited to the data?" and "Which system gives the best overall performance?". Thus such assessment will in future be based upon the mutual information between the output categories and the data. In addition information theory is being used as the basis of the rescaling of the Euclidean axes in the modified similarity metric. Readers seeking an introduction to information theory are referred to the text by Cover and Thomas (1991). Hakin (1994) discusses the use of information theory in the context of neural networks.

8. CONCLUSION

Pattern classification based upon Self Organising Maps, a type of unsupervised-learning neural network, have been shown to provide a means of classifying and diagnosing the state of health of rivers from biological and environmental data. A useful benefit of SOMs is that they do not require target outputs for their training. However, their classifications are meaningless until their output arrays have been labelled, a task that can prove substantial. The use of a large output array (i.e. many classification bins) enables the complex relationships that exist between the composition of the aquatic community and the quality of the river to be modelled more closely. This permits finer classifications to be made and thereby gives a diagnostic capability to the SOM, but the large size of the array makes the task of labelling the bins more onerous. A method of minimising the involvement of experts in this labelling task, based upon the use of feature maps and class templates, has been outlined. Feature maps have also been shown to provide a powerful means of visualising the multi-variate data, and of facilitating a greater understanding of the domain.

Although the methods presented are designed for use in the monitoring and control of river pollution, they offer considerable potential for use in other environmental fields.

9. ACKNOWLEDGEMENTS

The authors wish to thank the Environment Agency for granting permission to publish this work, which was carried out as part of National R&D Project E1/i621. Thanks are also due to Dr J. Murray-Bligh (Project Manager) from the Environment Agency for his assistance on various aspects of the project, and to Dr. V. N. Fontama (former Research Associate, Staffordshire University) for his contribution in the early stages of the project.

10. REFERENCES

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

Cover T. M. and Thomas J. A. (1991) *Elements of Information Theory*. Wiley.

DePauw N. and Hawkes H. A. (1993) Biological monitoring of river water quality. In *River Water Quality Monitoring and Control,* eds. Walley W. J. and Judd S. Aston University, Birmingham. 87-111. ISBN 1-85449-115-6.

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

Kohonen T. (1997) *Self-Organising Maps*. Second Edition. Springer-Verlag.

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

Walley W. J. and Fontama V. N. (2000) New approaches to the biological classification of river quality based upon artificial intelligence. In: *Assessing the biological quality of fresh waters. RIVPACS and other techniques.* Eds. Wright J. F., Sutcliffe D. W. and Furse M. T. Freshwater Biological Association, Ambleside. 263-279.

Walley W. J., Fontama V. N. and Martin R. W. (1998) Applications of Artificial Intelligence for the Biological Surveillance of River Quality. R&D Technical Report E52, Environment Agency, Bristol.

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

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

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

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

Walley W. J. (1993) Artificial Intelligence in River Water Quality Monitoring and Control. In *Proceedings of the Freshwater Europe Symposium on River Water Quality Monitoring and Control,* eds. Walley W. J. and Judd S. Aston University. February 1993. ISBN 1-85449-115-6.