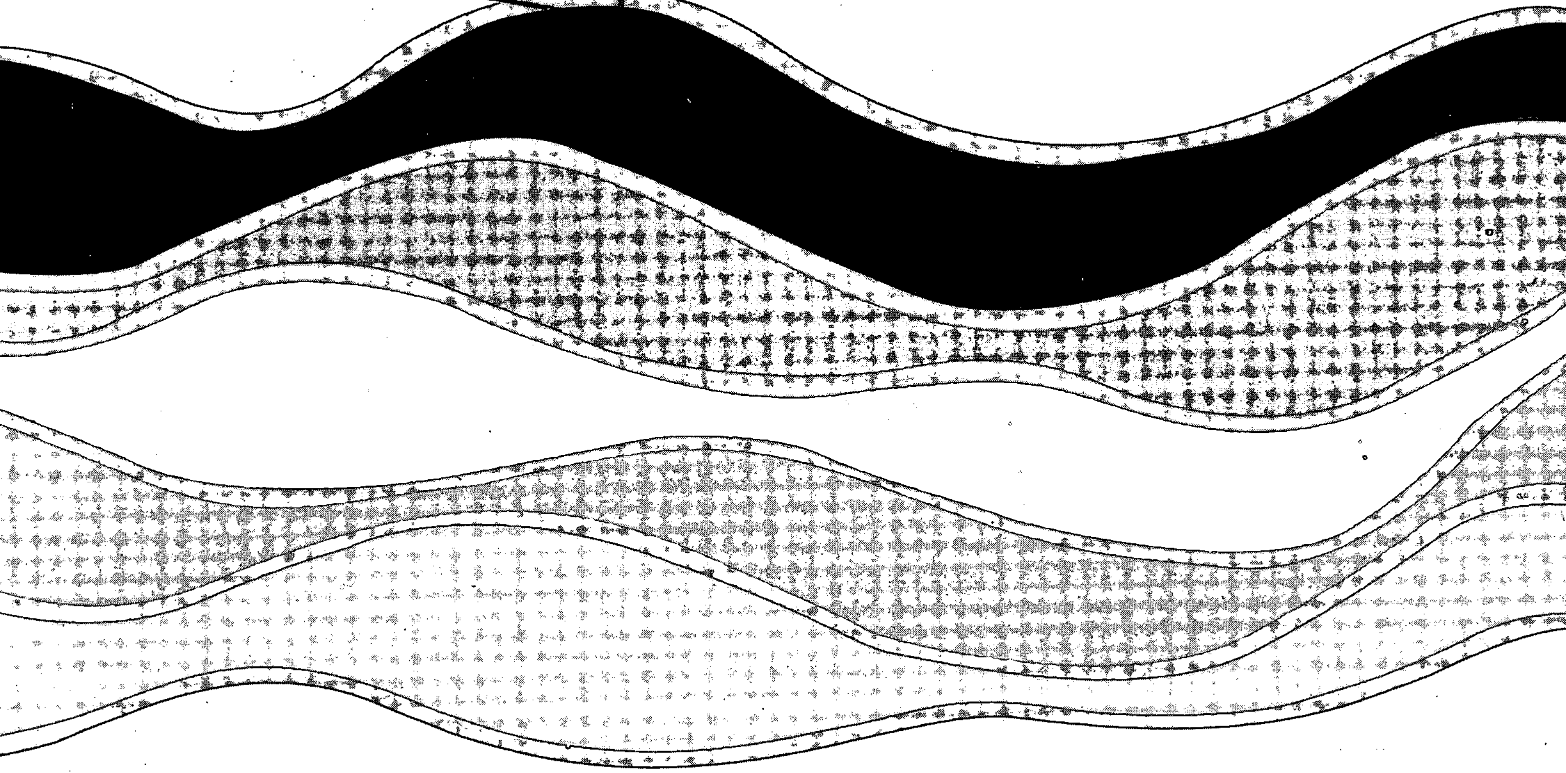
NATIONAL
WATER
RESEARCH
INSTITUTE

INSTITUT
NATIONAL
de RECHERCHE
sur les
EAUX



A NEURAL NETWORK APPROACH TO PREDICT MISSING ENVIRONMENTAL DATA

I.W. Wong, D.C.L. Lam, A. Storey, P. Fong &
D.A. Swayne

NWRI Contribution No. 93-153

A NEURAL NETWORK APPROACH TO PREDICT MISSING ENVIRONMENTAL DATA

I.W. Wong[1], D.C.L. Lam[1], A. Storey[2], P. Fong[2] & D.A. Swayne[2,3]

[1]National Water Research Institute
867 Lakeshore Road, Box 5050
Burlington, Ontario, CANADA L7R 4A6

[2]ESA Inc.
489 Enfield Road, Burlington, Ontario, CANADA L7T 2X5

[3]Computing & Information Sciences Dept.
University of Guelph, Guelph, Ontario, CANADA N1G 2W1

NWRI Contribution No. 93-153

MANAGEMENT PERSPECTIVE

This is a paper for the World Congress on Neural Networks, to be held in San Diego, California, U.S.A., June, 1994. It represents practical applications of applying neural network techniques to environmental problems. The research highlights a useful approach to predict missing data from other known data. It is found that the neural network approach, together with cluster analysis and data transformation, produces satisfactory results. Other environmental applications can be easily adapted within the existing neural network framework in RAISON. The application was used in a joint study between the National Water Research Institute and the State of Environment Reporting.

ABSTRACT

We discuss some preliminary results of a neural network approach to predict missing environmental data. One of the main problems in environmental modelling and expert system application is the lack of useful data. The neural network approach will no doubt provide more useful data. It is found that the neural network, when used with proper pre-screening processes, produces satisfactory results. The preprocessing techniques used here are the cluster analysis to filter noisy data and the transformation to align the data in the appropriate range. Design procedures for this application are given and its performance is discussed by means of a sensitivity analysis.

I. INTRODUCTION

Environmental science is a complex field and multidisciplinary in nature. When determining how to correctly analyze any collection of data, the first consideration must be the characteristics of the data themselves. After understanding the data, we can apply these data in applications such as environmental modelling, expert systems and statistics. Most often, environmental data are measured frequently or in real time, but others are often incomplete and noisy. Most statistical software permit the entry of the missing data. Sometimes, more than one code might be used to identify particular types of missing data, such as don't know, no measurement or out of legitimate range. Analytical packages typically exclude the missing data for any of the variables in an analysis. This approach is found to be inappropriate, since the researcher is usually interested in understanding the entire database, rather than the portion of the database that would provide measurements to all relevant variables in the analysis.

In the past few years neural networks have received a great deal of attention in many areas. Recently there have been many practical applications to apply neural networks techniques to environmental science [1]. Much of the success is contributed to the ability of neural networks to predict and draw conclusions when presented with complex, noisy, irrelevant, and even partial and missing information. In particular, we are interested in dealing with the data poor situation, i.e. finding the appropriate values of the missing data. As a result, a neural network can be trained to predict missing values from a network trained to examples.

II. NEURAL NETWORK APPROACH

We have developed a knowledge-based system, RAISON, using the C++ language [2]. This system is composed of a geographical information system (GIS), a database, a spreadsheet and most importantly, an artificial intelligence (A.I.) module. In this paper, we describe how to implement neural networks within the A.I. module.

The neural network that we have implemented is a variation of the back-propagation network model [3] because of its simplicity and flexibility required to interface with other modules in RAISON. In this network model, data from the input layer are fed into one or more hidden layers and a set of connections weights are continually adjusted under the supervised training mode. In particular, the feedforward output state calculation is combined with backward error propagation (e.g. using the conventional sum of squared errors based on the square of the difference between the output value and observed value). In addition, it is found that the back-propagation network is the most appropriate one in dealing with missing data. However, the training time tends to be very slow in this kind of network. Since environmental data sets are often very large, we have opted for a modified form of back-propagation network [4], called quickprop.

The following are some of the essential features of this neural network model:

(i) The weight is updated based on the sum of the errors affected by that weight over the entire training set. This gives the gradient of the composite error function in the weight space.

(ii) Different learning rates are prescribed for different weights in the network.

(iii) We set the weight change proportional to the derivative of the sigmoid function and when a node output is close to the extremes of the function, the derivative is close to zero. When an output node's error is large, the small derivative allows a slight change to be propagated back, hence slowing down the convergence. To overcome this, a constant of 0.1 is added to the derivative value.

(iv) While standard back-propagation takes the first derivative of the error surface with respect to the weights and adds a constant step size in that direction, quickprop uses the gradient of the error function at two consecutive points and the weight change between them, fits a parabola and sets the step size to the minimum of this parabola.

(v) To further improve the performance, a weight decay term is added to the error function which optimizes the weights on this slightly changed error surface.

III. PREPROCESSING TECHNIQUES

It is not always necessary to have all the environmental parameters in the data set used in the neural networks. Some parameters have a stronger relationship than others. For example, pH and alkalinity measurements should be closely related. Therefore, it is desirable to pre-screen the parameters before feeding them into the neural network. One of the most commonly used preprocessing methods is cluster analysis.

Statistical methods are used to select clusters in order to examine whether significant differences exist between the objects of different clusters. One of the approaches is to consider a criterion for measuring the tightness of the clusters and to plot its value against the number of clusters [5]. A sudden marked flattening of the curve indicates a significant number of clusters since there is relatively little gain from a further increase of number of clusters. In our case, the correlation coefficients between parameters are used as a criterion to determine the number of significant clusters.

It is also found that the ranges of some environmental parameters vary over several orders of magnitude. In practice, careful treatment of the network representation is usually required to obtain an efficient network. When the range of a parameter is large (many orders of magnitude), the logarithm of the variable should be used. This transformation makes small changes in small inputs as important as large changes in large inputs. In addition, some of the parameters are logarithmic in nature, e.g. pH. We must deal with these parameters properly to achieve optimum results.

IV. AN EXAMPLE

We use a data set of water quality data (NAQUADAT) in the Great Lakes Ecoregions as an example. This set of data consists of the following parameters: Na, K, Ca, Mg, SO4, Cl, alkalinity, pH, colour, specific conductivity, Al, Fe, NO_3, NH_4 and DOC. Some of the pH data are not available. The problem is to show how to make use of the existing known data for all parameters to predict the missing pH values.

First, we have to identify the parameters which have a close relationship with pH and use them as the input parameters in the neural network. This is done using the cluster analysis. A cluster distance is assigned to each pair of parameters and then the cluster process uses these distances to classify the parameters into different clusters and to determine the appropriate number of clusters. In this example, the cluster distance is based on the correlation coefficient from the regression analysis. Figure 1 shows the plot of cluster distance (regression coefficient) against the number of clusters. The significant value is identified when there is a abrupt change in slope. The results of the cluster analysis are summarized in Figure 1. By inspecting this figure, the appropriate number of clusters is 10. It is found that pH formed a cluster with Ca, specific conductivity and alkalinity.

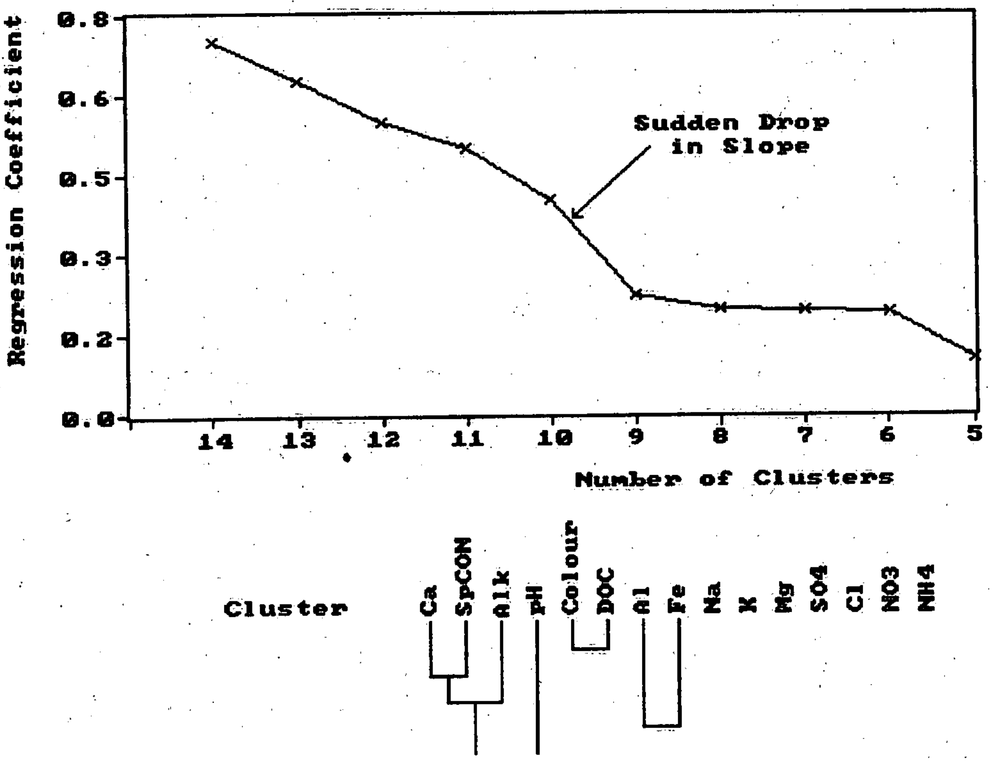


Figure 1: Cluster results of the 14 parameter data set

Once we identify the parameters that are related to pH, we can set up a neural network to estimate the pH values from these parameters. To validate the preprocessing techniques, a sensitivity analysis is performed. Three different network topologies are used. The first network uses all 14 parameters to predict pH. The second network uses seven randomly picked parameters that are unrelated with pH as determined from the cluster analysis as inputs. Finally, we use the three parameters from the cluster results as the input parameters. In addition, each topology has two variations: the output parameter pH is either normal or log-transformed. The number of hidden layers can be set to one to speed up the training. As a rule of thumb, the number of hidden nodes is set to the geometric mean of the sum of the inputs and outputs nodes [6]. Table 1 summaries the layout of the sensitivity analysis.

Case	Input	Hidden	Output	pH	Relationship	Median Rel.
	Nodes	Nodes	Node	Tranform.	with pH	Error
a	14	4	1	Yes	Don't Care	1.8%
b	14	4	1	No	Don't Care	3.2%
c	7	3	1	Yes	Unrelated	4.5%
d	7	3	1	No	Unrelated	8.2%
e	3	2	1	Yes	Highly related	1.6%
f	3	2	1	No	Highly related	2.8%

Table 1: Layout of the sensitivity analysis

V. RESULTS AND DISCUSSION

There are a total of 2250 good records in the data set. We randomly select 400 records as our training set and the rest is for verification. Training is complete when either the network reaches one million epochs or the sum-squared error satisfies a prescribed threshold. We use the median values of the relative error between the predicted pH and the observed pH as a benchmark of the sensitivity analysis. Table 1 displays the results of the six network topologies.

It is found that cases (c) and (d) are the worst cases as expected. Although the results of case (a) and (e) are close, the training time on case (e) is about one-tenth of case (a). In addition, case (e) shows slightly better results. It confirms that cluster analysis based on regression coefficient is an essential tool for preprocessing environmental data. It is also found that cases (a), (c) and (e) generate better results than their counterparts, i.e. cases (b), (d) and (e). This shows that proper transformation on certain environmental parameters will yield vast improvement over simply using the data.

The neural network provides a new approach to predict missing values of environmental data. It has been successfully implemented in RAISON as part of the artificial intelligence component. The sensitivity analysis shows that preprocessing of the data is crucial to the performance of the neural network. Although neural networks require a great deal of training and computational resources, it is very fruitful when dealing with large data sets in environmental science. Neural networks can be applied to provide more information with reasonable accuracy. Future work include the improvement of the training time and in formulating imprecise measurements with fuzzy logic algorithm.

VI. ACKNOWLEDGEMENTS

The authors would like to thank Dr. William G. Booty, Jane Kerby and Kevin Hopkins for their helpful suggestions.

VII. REFERENCES

Schmuller, J. 1990, "Neural Networks and Environmental Applications", in J.M. Hushon (ed.) *Expert Systems for Environmental Applications*, ACS Symposium Series 431, pp. 52-68.

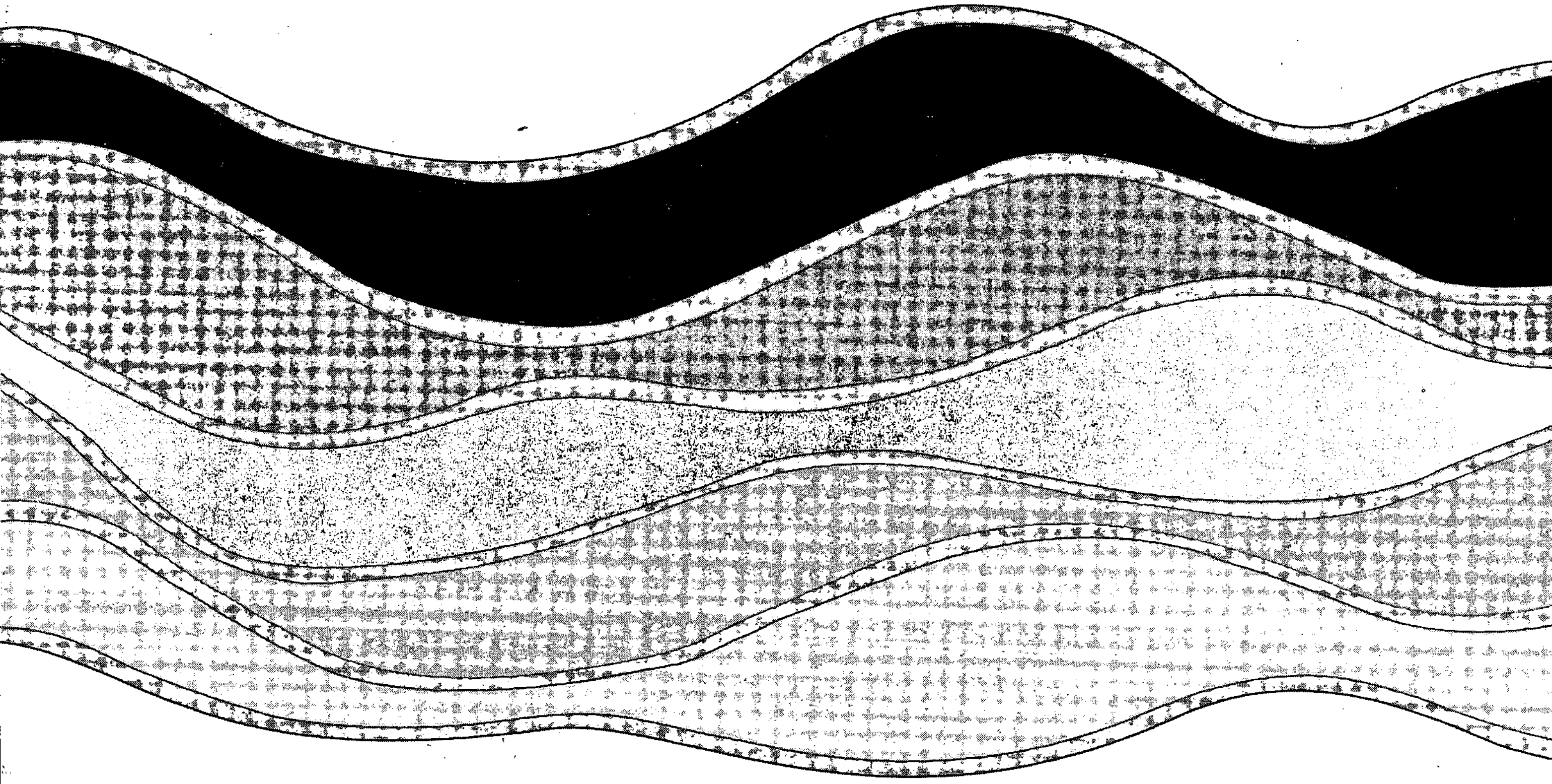
Lam, D.C.L. and D.A. Swayne, 1992. "Some experiences in applying the RAISON expert system to environmental problems. In (Eds. J.W. Brahan and G.E. Lasker) Advances in Artificial Intelligence - Theory and Applications, IIAS Publication, Windsor, Canada. pp. 59-64.

Rumelhart, D.E. and J.L. McCelland, 1986. "Parallel distributed processing, explorations in the microstructure of cognition. Vol. 1: Foundations", MIT Press, Cambridge, MA.

Fahlman, S.E., 1989, "Faster-learning variations on back-propagation: an empirical study", in Proceedings of the 1988 Connectionist Models Summer School, Morgan Kaufmann.

Massart, D.L. and L. Kaufman, 1983, "The interpretation of analytical chemical data by the use of cluster analysis", Wiley, pp. 147-149.

Maren, A.J., Harston, C.T. and R. M. Pap, 1990. Handbook of neural computing applications, Academic Press, pp.238-243.



NATIONAL WATER RESEARCH INSTITUTE
P.O. BOX 5050, BURLINGTON, ONTARIO L7R 4A6

Environment Canada Environnement Canada

Canadä

INSTITUT NATIONAL DE RECHERCHE SUR LES EAUX
C.P. 5050, BURLINGTON (ONTARIO) L7R 4A6