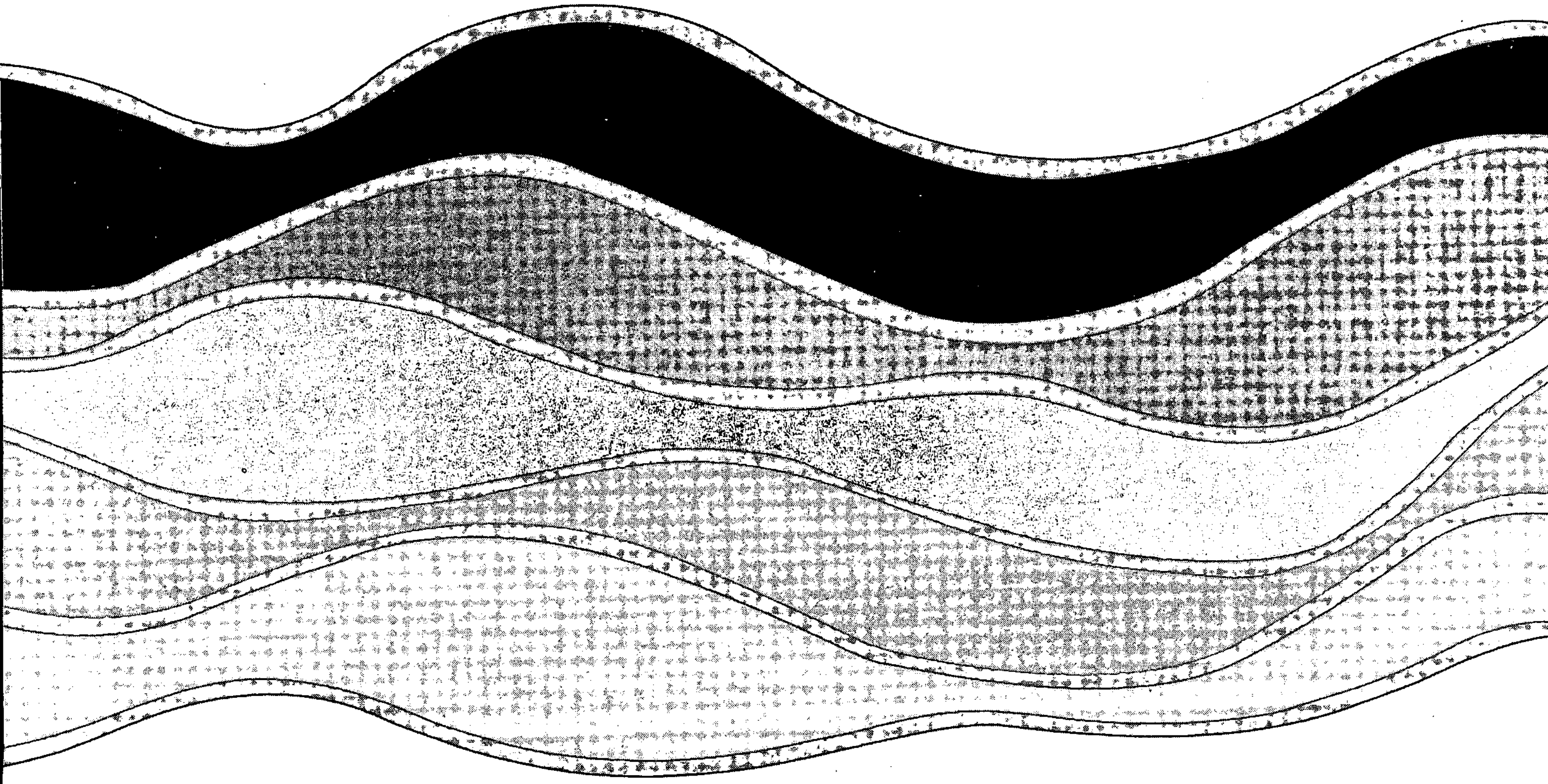
NATIONAL
WATER
RESEARCH
INSTITUTE

INSTITUT
NATIONAL
de RECHERCHE
sur les
EAUX



A HYBRID EXPERT SYSTEM AND NEURAL NETWORK APPROACH TO DATA MODELS:

AN EXAMPLE OF ENVIRONMENTAL APPLICATION

D.C.L. Lam[1], I. Wong[1], D.A. Swayne[2] and P. Fong[2]

NWRI Contribution No. 93-151

A HYBRID EXPERT SYSTEM AND NEURAL NETWORK APPROACH TO DATA MODELS:

AN EXAMPLE OF ENVIRONMENTAL APPLICATION

D.C.L. Lam[1], I. Wong[1], D.A. Swayne[2] and P. Fong[2]

NWRI Contribution No. 93-151

A HYBRID EXPERT SYSTEM AND NEURAL NETWORK APPROACH TO DATA MODELS: AN EXAMPLE OF ENVIRONMENTAL APPLICATION

D.C.L. Lam[1], I. Wong[1], D.A. Swayne[2] and P. Fong[2]

[1]National Water Research Institute
P.O. Box 5050, Burlington, Ontario, CANADA L7R 4A6

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

NWRI Contribution No. 93-151

MANAGEMENT PERSPECTIVE

This is an invited paper for the 4th International Symposium on Systems Research, Informatics and Cybernetics, to be held in Baden-Baden, Germany, September, 1993. It is a direct recognition of the work on the RAISON expert system developed at the National Water Research Institute. The significance of this research paper is on the ability to identify patterns, both in time and space, on environmental data and the capability to fill these gaps with accurate approximations to the degree possible by training the so-called neural network (a machine learning mechanism using artificial intelligence). Other applications of this technology include the normalization of environmental data measured by different analytical methods with varying degrees of accuracy. This study was funded partly by an ISTC AI R&D Fund and the application was used in a joint study between the National Water Research Institute and the State of Environment Reporting.

EXECUTIVE SUMMARY

This paper describes a new approach for developing a decision support system for aquatic environmental applications by using a hybrid combination of expert system and neural network techniques. Expert systems are based on available knowledge usually represented by a rule base. A neural network can be used, under supervised machine learning, to search for more knowledge, particularly to identify new temporal and spatial patterns hitherto not formally found in rule sets. A hybrid system, therefore, can utilize the best of these two methods. By demonstrating with two examples on identifying temporal and spatial data gaps, the importance of screening out irrelevant data is emphasized for building such a hybrid system.

ABSTRACT

A neural network is implemented to work with an existing expert system for environmental applications. Examples on filling spatial and temporal data gaps showed highly accurate estimates from the neural network if appropriate noise screening procedures are used. The feasibility of developing an integrative system with database, analysis, map, expert system and neural network is explored.

Keywords: Neural network, expert system, knowledge-based system, hybrid system.

INTRODUCTION

Environmental Science is multidisciplinary in scope. Solutions to environmental problems require not only the integration of knowledge from natural sciences (physics, chemistry and biology, etc.) but also information from social sciences (population dynamics, social psychology, etc.) Some environmental data are measured frequently or in real time, but other data are often incomplete and noisy. The ability to integrate multidisciplinary knowledge bases and to fully utilize information from incompatible databases has been the focus of many environmental studies.

Recently there have been several efforts to apply knowledge-based systems in the area of environmental science (Lam and Swayne, 1992). In this paper, we present a hybrid system in which a knowledge-based system called RAISON was used for data retrieval and knowledge acquisition, working in parallel with a neural net module. A screening procedure for selecting proper environmental factors was introduced by screening out irrelevant data before training the neural net. This step turns out to be very important in order to apply the neural network technologies in any practical sense for environmental problems. Such a step, if implemented intelligently, can reduce the network training time and can produce more accurate results. The purpose of this paper is to illustrate the hybrid approach using screening procedures with two examples of applications for filling spatial and temporal data gaps by using information from other data sets.

Methodology

We have developed on a microcomputer using the C++ language a knowledge-based system, RAISON, that has a database, a geographical information system (GIS) and a statistical analysis module (Lam and Swayne, 1992). Central to the architecture of this system is the expert system (inference engine) module that can be used to process rules acquired from experts. If required, the expert system can retrieve data from the database, analyze and display results. In this paper, we add a neural network capability (Fig. 1) that will work with other existing modules in the system. At the current stage of development, since we found that proper screening procedures (i.e. the method to select the appropriate data and/or parameters to train the network), are needed to work along with the network, the RAISON framework (Fig. 1) is a convenient one to facilitate direct linkages among database, statistical procedures and the neural network.

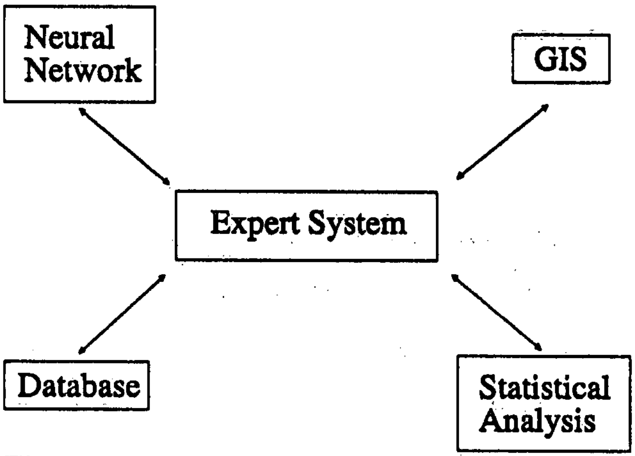


Figure 1 Hybrid system configuration

While a number of network paradigms can be used, we have chosen the multi-layer back propagation model (Rumelhart and McClelland, 1986) because of its relative simplicity and flexibility required to interface with other modules in RAISON. In this network model, input data are fed into one or more hidden layers and a set of connection 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 mean of sum of squared errors based on the square of the difference between the output value and some expected (observed) values. The backward error propagation is also implemented with momentum factors involving both current and previous corrections (Eberhart and Dobbins, 1990).

Example I: Filling Spatial Data Gaps

A common problem with environmental database is missing data. Often, a set of water or air quality parameters was measured in a number of stations, but occasionally one or more of the parameters were not measured, because of instrument failures and other reasons. One possible treatment is to use the neural network approach in which all or part of the parameters are used to train the network which is then used to estimate the missing values. As an example, we illustrate the approach by using one such data set, which is available on the RAISON System, on water quality in the Great Lakes Mixed Wood Plain Ecoregion. The parameters measured are sodium (Na), potassium (K), calcium (Ca), magnesium (Mg), sulphate (SO_4), alkalinity (Alk), dissolved organic carbon (DOC), aluminium (Al) and pH. While there are full records of all these parameters for most stations, the value for pH is missing for some. A straightforward application of the neural network is to make use of those stations with a full set of data and train the network to estimate pH values from the values of all the other parameters. When the network is deemed to be trained (i.e. satisfying error threshold criteria), the network can then be applied to those stations with missing data to predict the missing pH values by using the values from the other parameters. Since the data are interpolated across spatial stations, the approach amounts to filling spatial data gaps.

However, we propose a different strategy here. Instead of using all parameters, which may or may not be related to pH, we select only those that are related to be used in the neural network. There are many ways to establish whether one or more parameters are related to each other. The method we used is the cluster analysis method, which is available in the statistical module in RAISON. A cluster distance, e.g. one that is based on a linear regression coefficient, was assigned to each pair of parameters and the parameters were grouped into clusters. For example, in Fig.2, the cluster distance of Na and Ca and that of Mg and Alk were grouped into Na, Ca, Mg and Alk, and so on, until there appeared an abrupt breakpoint that clearly indicated two separate large clusters. One cluster contained pH and related parameters; the other cluster, DOC and Al. Therefore, DOC and Al were excluded from the network for this test case.

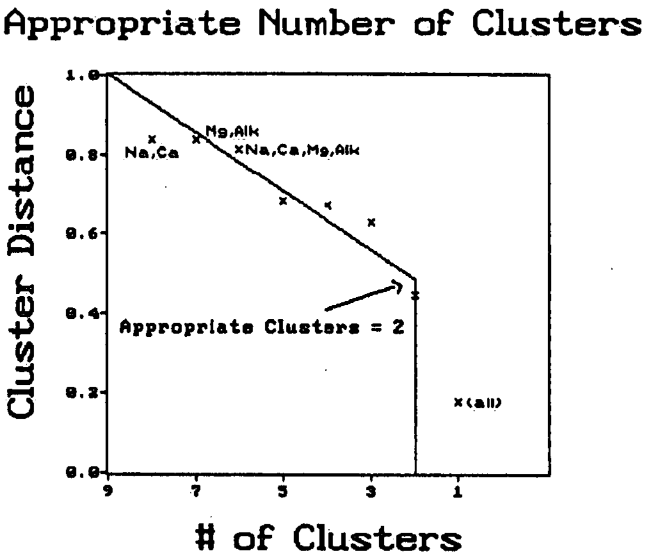


Figure 2 Selection of Clusters

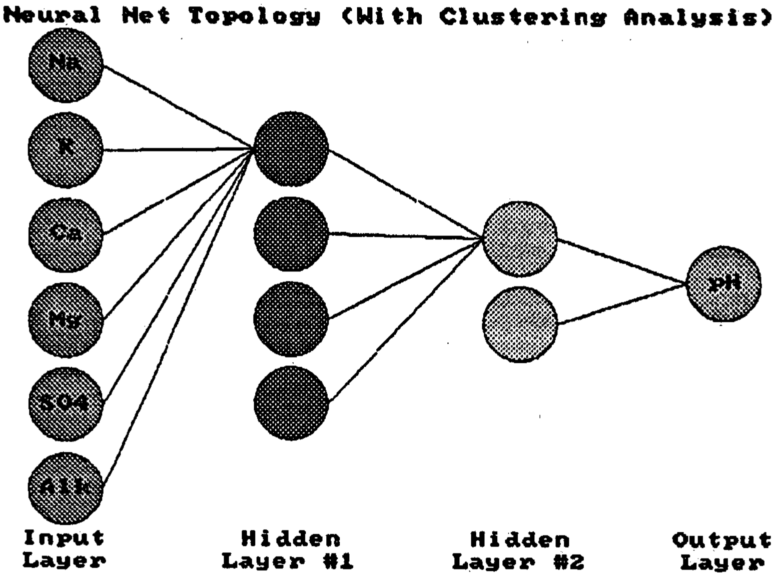


Figure 3 pH Neural Network Topology (with cluster analysis)

Figure 3 shows the network topology based on the parameters identified by the clustering analysis. The input layer consists of values of Na, K, Ca, Mg, SO_4 and Alk. The output is the estimated pH value. There are two hidden layers used for back propagation. For comparison purposes, we also trained the network using all parameters (i.e. including DOC and Al). For the same number (one million epochs) of training cycles, the network with clustering produced smaller errors (e.g. a 75th percentile error of 3.6% and a median error of 2.5%) than the network without clustering (a 75th percentile error of 5.5% and a median error of 2.4%).

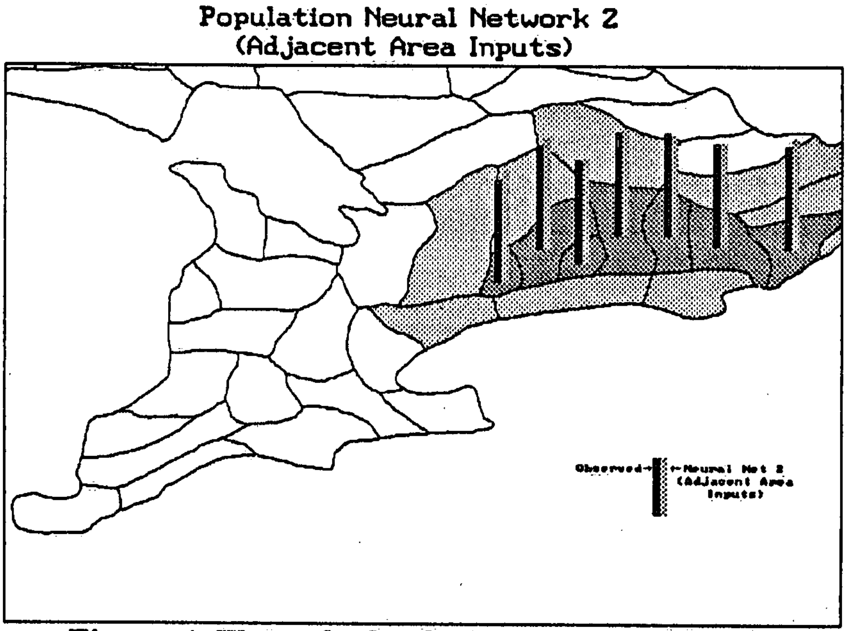


Figure 4 Watersheds (darker shade: with missing data; light shade: adjacent watersheds) (bars: observed and computed population)

Example II: Filling Temporal and Spatial Data Gaps

Another example is to apply the neural network approach to fill data gaps in space and time. For example, we have the population data for all the watersheds in southern Ontario (Fig. 4) for the census years, 1971, 1976, 1981 and 1986, except that, for the seven watersheds marked in Fig. 4, the 1981 data were not made available to us at the beginning of the study. Subsequently, they were provided to us and served as confirmation data for the neural network results. The aim of this exercise is to estimate the population for 1981 in the seven watersheds with missing data by using the population data from the other watersheds for 1971, 1976 and 1986. In other words, we have to interpolate the data in space and time.

Table 1 Results of the Population Neural Networks

Area	Observed	Neural Net 1	Relative Error	Neural Net 2	Relative Error
1	20924	22769	0.0882	21829	0.0433
2	43039	46275	0.0752	44727	0.0392
3	89959	91779	0.0202	91079	0.0125
4	55734	56794	0.0190	56598	0.0155
5	51205	50418	0.0154	50388	0.0160
6	115033	115431	0.0035	115097	0.0006
7	51783	56274	0.0867	53694	0.0369
Cumulative Relative Error			0.3082		0.1640

Instead of using information from other related parameters to train the network as in Example I, we use only the population data. However, we can choose the data from all or a subset of the watersheds (had we had more years of data, we could also have the choice of all or some of the years). We attempted two different ways of training the neural network: (1) the data from all the watersheds were used and (2) only the data from watersheds that are adjacent to the seven watersheds (Fig. 4) were used. In neural network 1, much more information was used but it could introduce more noises into the network. In neural network 2, less information was used but it might or might not be adequate for the training.

After several attempts, the final topology used for network 1 contains 26 input nodes, 2 hidden layers (16 and 10 nodes) and 7 output nodes for the 7 watersheds with missing data. For network 2, we used 11 input nodes, 1 hidden layer (9 nodes) and 7 output nodes. The results of the two networks are shown in Table 1, compared with the observed data that we subsequently obtained. Since we did not have the observed data when training the network, the relative errors (Table 1) between the predicted and observed values constitute a rigorous verification of the network results. Clearly, neural network 2 produces consistently much smaller errors (a cumulative relative error of 0.164) than network 1 (a cumulative relative error of 0.308).

Conclusions

For environmental data problems, the neural network approach offers a viable solution to filling data gaps. Some forms of screening or selection of input data are required to prevent irrelevant or noisy data from interfering with the performance of the network. As the selection procedures are based on statistical methods or knowledge-based information, an integrative system such as RAISON which already has these analytical and expert system tools can be conveniently developed into a hybrid neural network and knowledge-based system. While the system can be operated with a microcomputer, we are currently implementing the system on the workstation platform for more complex neural network.

ACKNOWLEDGEMENT

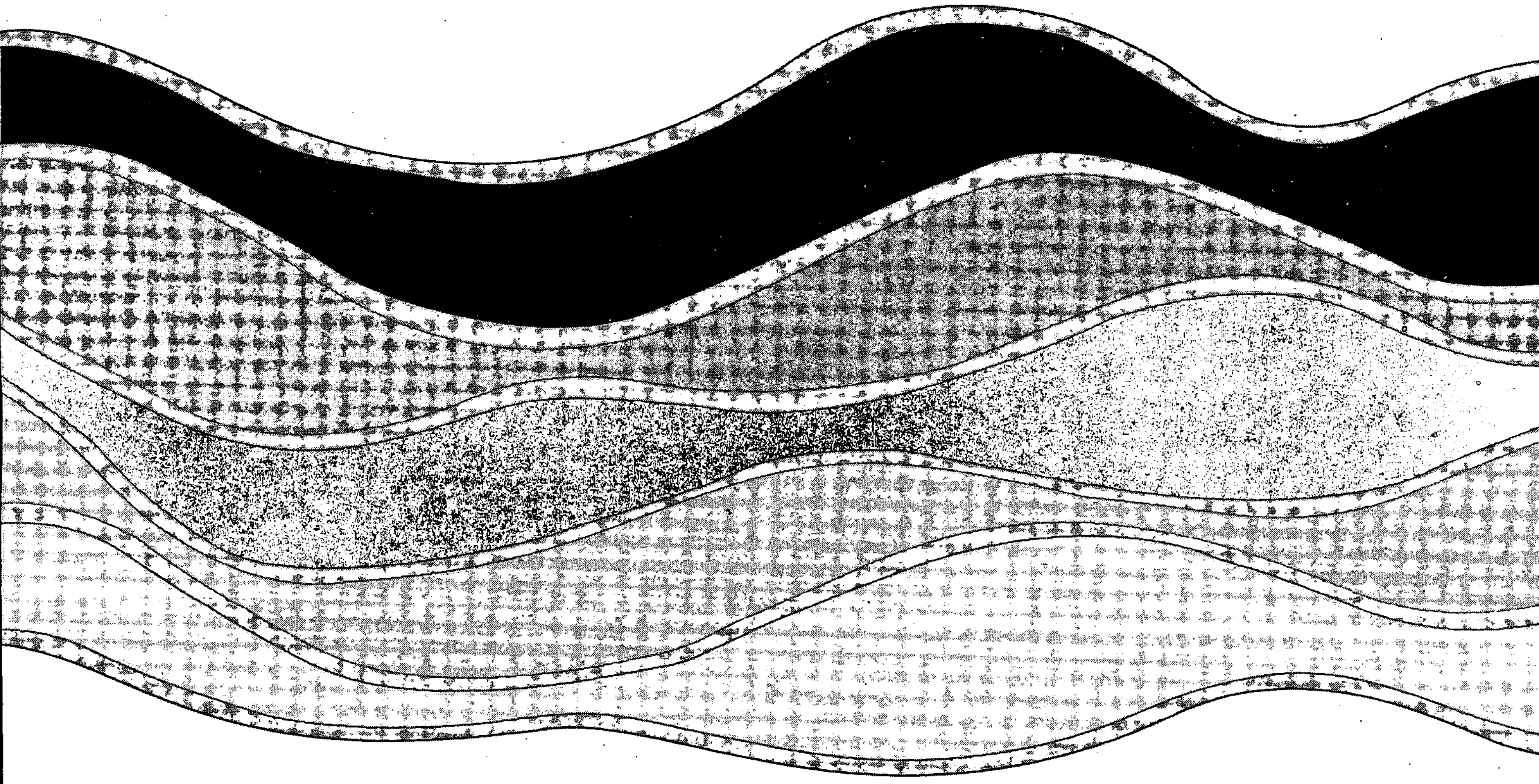
Statistics Canada provided the population data. Thanks are due to C. Pupp for his advice and to State of Environment Reporting, Environment Canada, for providing part of the study funding.

REFERENCES

Lam, D.C.L. and Swayne, D.A. 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 Application, IIAS Publication, Windsor, Canada. pp. 59-64.

Eberhart, R.C. and Dobbins, R.W. 1990. Neural Network PC Tools - a Practical Guide. Academic Press, N.Y. 414p.

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



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

Environment Canada Environnement Canada

Canada

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