18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Predicting Biochemical Oxygen Demand As Indicator Of River Pollution Using Artificial Neural Networks

Talib, A.[1], Y. Abu Hasan [2] and N.N. Abdul Rahman[3]

[1] 1 School of Distance Education, Universiti Sains Malaysia, USM, Penang, Malaysia, 11800
[2] School of Mathematical Sciences, Universiti Sains Malaysia, USM, Penang, Malaysia, 11800
[3] School of Industrial Technology, Universiti Sains Malaysia, USM, Penang, Malaysia, 11800
Email: anita@usm.my

Abstract:

Artificial Neural Networks (ANNs) are frequently used to predict various ecological processes and phenomenon related to water resources. Various ANN applications involve the prediction of water quality using various environmental parameters. This paper discusses the use of two environmental data sets from two sampling sites, merged together and applied to forecasting a dependent variable, namely biological oxygen in demand (BOD) using ANN modelling. ANN was applied to map the relationships between physical, chemical and biological time-series data of Sungai Air Itam, Pulau Pinang, Malaysia. This river is part of the Sungai Pinang river basin and is considered as highly eutrophic. Sungai Pinang and its tributaries are the main rivers flowing through the state of Penang, Malaysia. The water quality has been increasingly deteriorated by both natural and anthropogenic effects. The purpose of this study is to investigate the application of ANN to predict the biochemical oxygen demand as a measure of eutrophication status of rivers. The results of the study show that it is possible to forecast 1 month ahead BOD for Sungai Air Itam using a simple ANN with 16-4-1 architecture. The most important input for this predictive model is phosphate, and the sensitivity of the ANN models to the inputs is also dependent on the training datasets.

Keywords: *Artificial Neural Networks, River Pollution, Biochemical Oxygen Demand*

1. INTRODUCTION

Water quality in rivers need to be maintained as "*clean*" in terms of their physical, chemical and biological conditions. In Malaysia, the Interim Water Quality Standard set for the country is adopted by the Department of Environment (DOE) to assess and classify river water quality. These classes can be referred to for determining their uses based on the Water quality Index (WQI), in relation to a number of water quality parameters (DOE, 1999). The management of water quality in Malaysia is a challenging task because river basins are situated in densely populated areas. Several Malaysian government bodies are responsible for the management of river water quality including DOE, Department of Irrigation (DID) and in some cases private companies commissioned to maintain the river water quality , such as Alam Sekitar Malaysia Sdn. Bhd. (ASMA). Several water quality models based on a traditional mechanistic approach were developed as a management measure to conserve water quality. Current efforts include the use of modern modelling approaches towards managing the river water quality for the future.

Different modelling approaches that are generally applied to analysing water resources include differential equation, statistical and computational methods (Recknagel, 2003). The use of various approaches for finding patterns and to applying them for forecasting is also known as data mining. Ecological data mining refers to the search and discovery of patterns for forecasting parametric values of the processes in ecological datasets. Artificial Neural Networks (ANNs) are a branch of Artificial Intelligence (AI), in which 'connectionist' paradigms are used to extract and store implicit knowledge embedded in the data. Unlike traditional statistical and differential equation approaches, ANN is considered to be a powerful data-modelling tool as it is able to capture and implicitly represent complex relationships within various variables, such as the input/output variables.

Basically, the advantages of neural networks are that they are able to represent both linear and non linear relationships and are able to learn the relationships directly from data used for training the network. Many researchers compared statistical methods with that of ANN based modelling (Lek *et al*, 1996; Maier and Dandy, 1996) and concluded that in their studies ANNs performed better when compared to traditional Multiple Regression and other classes of statistical modelling techniques. ANNs are more flexible, and are found to be more suitable for prediction purposes as they produce more accurate results that could be replicated. ANNs have been successfully applied in temporal studies of salinity in rivers (Maier and Dandy, 1996) benthic communities (Chon *et al*, 2000) in streams, eutrophication and algal blooms in lakes (Karul *et al*, 1999; Recknagel *et al*, 1997; Talib, 2006). A distinct advantage of ANNs compared to that of statistical and differential equation approaches is that the former can predict the timing and magnitude of species succession (Recknagel and Wilson, 2000). Various studies on various water bodies have shown that ANNs could predict algal blooms with abundance and succession patterns of blue-green algae species (Recknagel *et al*, 1997; Talib, 2006).

In this case study, we use ANN for 1 month ahead forecasting of water quality using BOD. BOD is an important parameter as it measures the amount of biodegradable organic matter in water. By definition, it is the amount of oxygen required for aerobic microorganism to oxidize the organic matter to a stable organic form (Chapman, 1992). Testing for BOD is a time consuming task as it takes five days from data collection to analysing with lengthy incubation of samples (referred to as BOD_5 reading). For the prediction of BOD, ANN is trained and tested using datasets from Sungai Air Itam or Air Itam River. This river is part of the Pinang River basin and is considered as highly eutrophic. Sungai Pinang and its tributaries are the main rivers flowing through the state of Penang, Malaysia. The water quality has been increasingly deteriorated by both natural and anthropogenic effects. A forecasting model based on BOD can assist decision makers in managing river water quality.

2. SITES AND DATASETS

Sungai Air Itam is part of the Pinang River Basin, located in the northern part of Penang Island (Figure 1). Pinang River basin is the biggest river basin made up of tributaries from other rivers including Air Terjun, Pinang, Dondang and Jelutong Rivers. It is one of the most polluted river basins in Malaysia (DOE 1999) mainly due to effluent from point source (natural rubber processing factories and pig farms), and non point sources. Various raw domestic and industrial solid and liquid wastes are dumped into the river adding to effluent from urban runoff. Hence, the name of the river is "Air Itam" which means "black water".

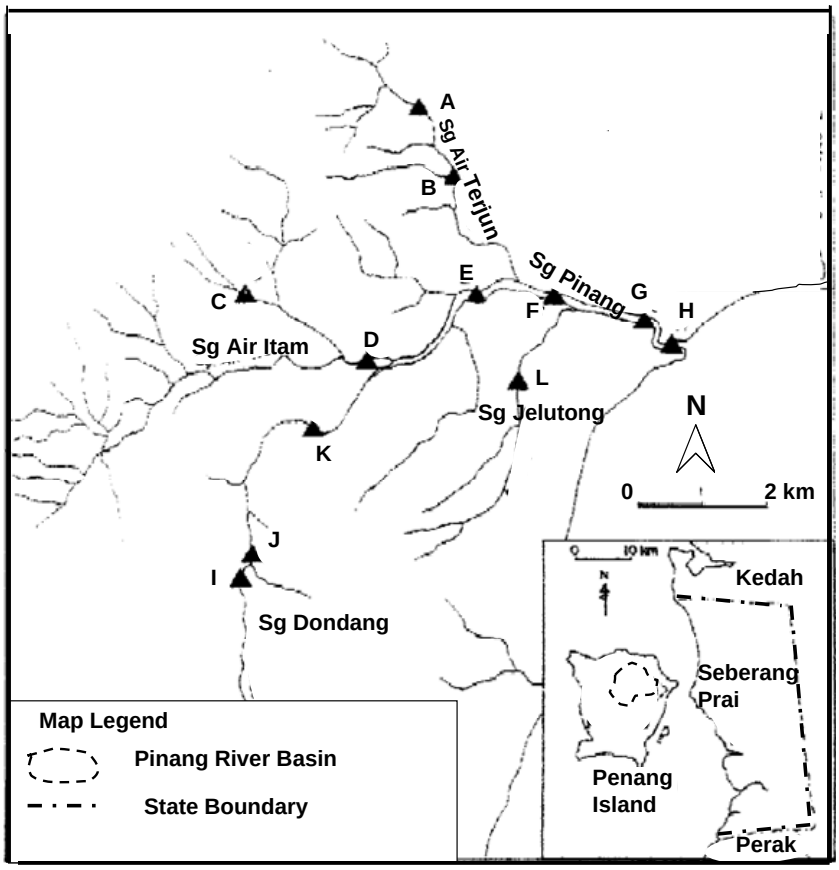

Figure 1 Location of Air Itam River and other river tributaries in Pinang River basin.

This study was conducted using a sample of River Air Itam datasets collected from 2001 to 2008. The monthly datasets are taken from station 1 and 2 (labelled as D and E, respectively in Figure 1). Three datasets were applied for training and testing using ANN:

1. Data from station 1

2. Data from station 2

3. Merged data from station 1 and 2.

The datasets were preprocessed as follows:

i. scaling

ii. removal of missing variables and outliers

iii. analysing data with descriptive statistical analysis.

3. ANN ARCHITECTURE AND TRAINING

An ANN architecture suitable for this study was selected manually, i.e. 16-4-1 architecture .The hyperbolic tangent was selected for input and output, with sum of squares as an output error function. Eight variables were selected for inputs, including temperature, pH, salinity, NO_3, PO_4, turbidity, dissolved solids and *E.coli* with an output of 1 month lag BOD.

This study involves training and testing of the ANN model by "trial and error" approach. ANN model with the 16-4-1 architecture, with conjugate gradient descent as training algorithm is chosen and applied. ANNs are computationally intensive and many parameters have to be determined with a few guidelines and no standard procedures to define the architecture (Lek and Guegan, 1999). Training was stopped based on the number of iterations, i.e 5000 and as the error on the validation set increases.

4. RESULTS AND DISCUSSION

4.1 ANN model in forecasting

Forecasting with ecological time series data sets using an ANN model is a complex task. Basically, the four main steps taken in this forecasting study are as follows:

Model Design: to choose a suitable model representation based on expert-knowledge.

Training: to estimate the parameters of the model.

Validation: to test the model on unseen data sets to determine its validity.

Interpretation: investigate into insights gained, causal explanations the model produced based on the hypothesis being tested.

Although ANNs are increasingly being used for the prediction of dependent variables in water resources, the problem of assessing the optimality of the results still exists. Apart from the importance of preprocessing , specific mapping of ANN depends on the architecture of the network, training techniques and modelling parameters.

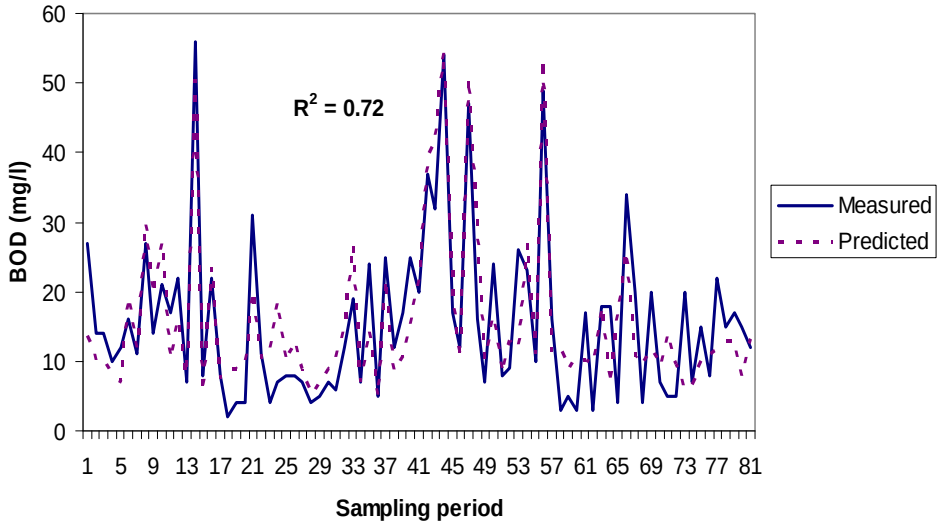

Figure 2 Results of 1 month ahead forecast of BOD using dataset from Station 1.

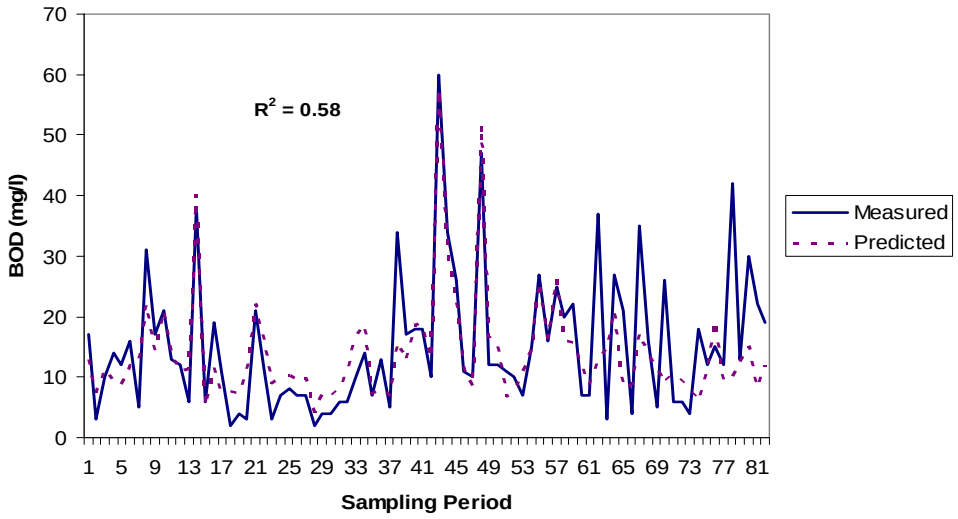

Figure 3 Results of 1 month ahead forecast of BOD using dataset from Station 2.

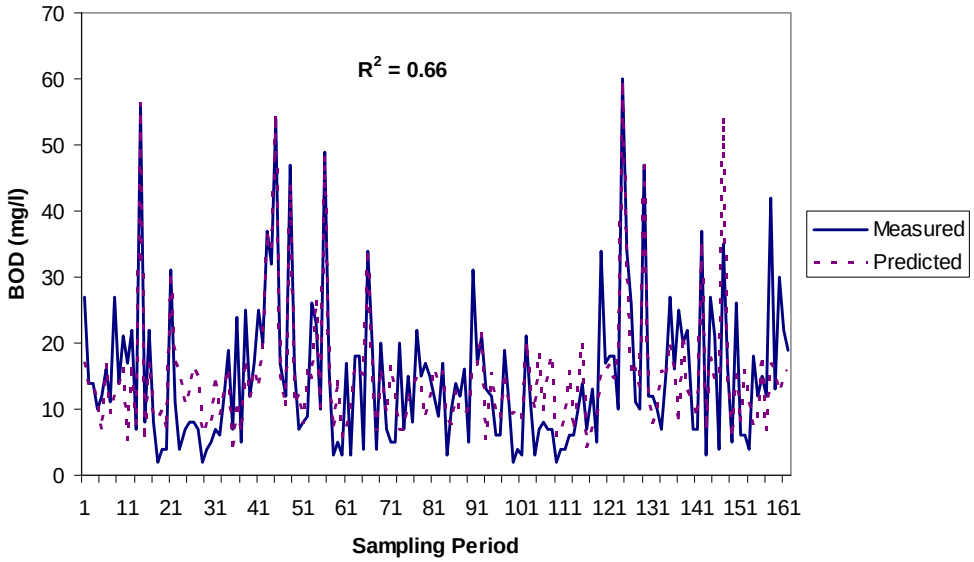

Figure 4 Results of 1 month ahead forecast of BOD using datasets from Station 1 and 2.

4.2 Modelling parameters

ANNs are sensitive to the composition of the training data set and to the initial network parameters. For this dataset the split-set validation is chosen for testing the models, with sequential data partitioning i.e. training (68%), validation (16%) and test (16%). Another validation technique, the leave-one-out bootstrapping and leave-*k*-out cross-validation have been attempted by (Wilson, 2004) whereby no user decisions are required regarding the division of data into training and testing sets. Another problem with ANN modelling is that it is difficult to predict an unknown event that has not occurred in the training data. The values of training data should therefore cover as wide a range as possible (Aoki *et al* 1999).

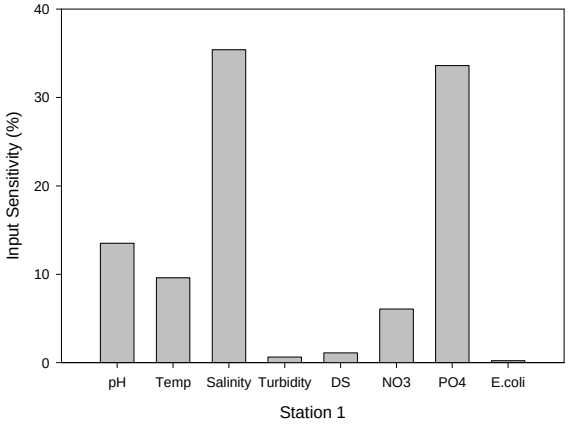

Figure 5 Results of input sensitivity on the ANN model trained on dataset from Station 1.

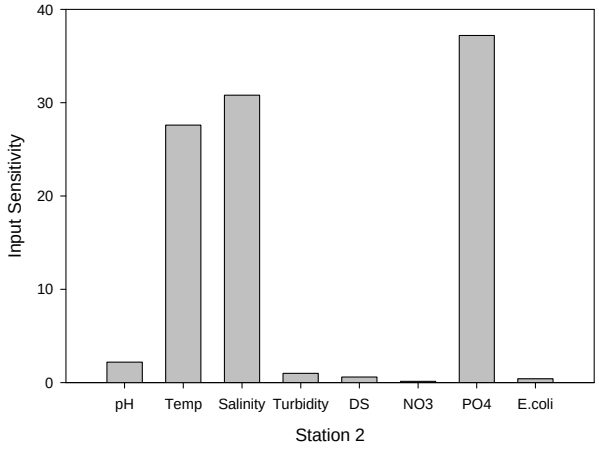

Figure 6 Results of input sensitivity on the ANN model trained on dataset from Station 2.

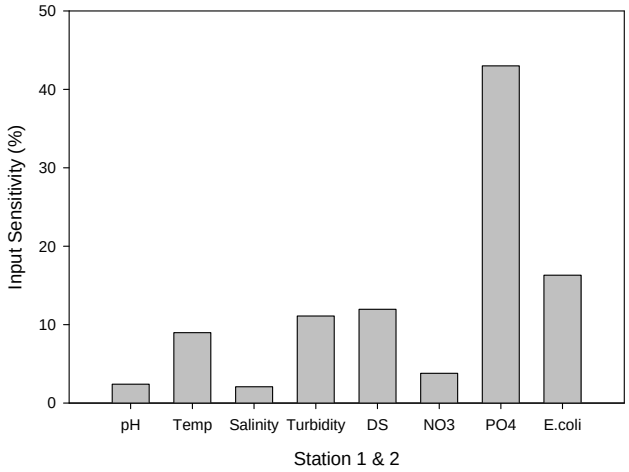

Figure 7 Results of input sensitivity on the ANN model trained on datasets from Station 1 and 2.

The results for 1 month ahead BOD forecast model from Station 1, showed R^2 values of 0.72 (see Figure 2). The results of the 1 month ahead BOD forecast for Station 2 is as shown in Figure 3 with R^2 values of 0.58, while for the merged datasets, it is as shown in Figure 4. The forecast for the merged model is quite good with R^2 values of 0.66. The results for ANN model from Station 1 and merged ANN model show good timing and magnitudes of forecast for high BOD periods. The result of the merging of datasets from two sampling stations is an improved prediction, due to increased training data.

Based on the results of the input importance (Figures 5, 6 and 7), we conclude that the composition of training data will also influence the results of the input sensitivity analysis. Results from the model show that when trained separately, the models revealed that the most important inputs are phosphate, temperature , salinity and pH, but when datasets are merged together, the ANN model is sensitive to inputs of phosphate, *E.coli*, turbidity, dissolved solids. BOD can increase when there is an increase in organic load to the rivers. Increase in dissolved organic matter consumes large amounts of oxygen. A scenario with reduced pH and an increasing *E.coli* abundance, is also associated with an increase in BOD. Organic water in polluted river is mainly urban wastewater that consists mainly of carbohydrates, proteins and lipids which will undergo anaerobic fermentation forming ammonia and organic acids. It is the hydrolysis of the acidic materials that also contribute to the decrease in pH values. Our results indicate a successful 1 month ahead prediction of BOD as an indicator of water quality for Sungai Air Itam. Hence, ANNs can unravel the complex non-linear relationship between the input variables. Good prediction models are needed for decision-making and it is an

important prerequisite for the management of the rivers as the BOD levels indicate the often unprecedented anthropogenic effects of pollution.

5. CONCLUSIONS

Various computational techniques can be applied to data mining ecological datasets. The results of the study show that it is possible to forecast 1 month ahead BOD for Sungai Air Itam using a simple ANN with 16-4-1 architecture that could reasonably forecast in terms of timing and magnitude. The most important input for this predictive model is phosphate, and the sensitivity of the ANN models to the inputs is also dependent on the training datasets. Apart from the ANN architecture and the training algorithm, good forecast is also dependent on ANN modelling parameters. Merging of datasets increase data for training, and improve generalization.

ACKNOWLEDGMENTS

The authors would like to acknowledge funding from the USM RU Grant, for the contribution to this research and conference participation and presentation.

REFERENCES

Aoki, I., Komatsu, T., and Hwang, K. (1999), Prediction of response of zooplankton biomass to climatic and oceanic changes. *Ecological Modelling* 120, 261-270.

Chapman, D (1992). Water quality Assessments, 1st Edition, Chapman and Hall Ltd., London , 80-81.

Chon, T.-S., Park, Y.-S., and Park, J.H.(2000), Determining temporal pattern of community dynamics by using unsupervised learning algorithms, *Ecological Modelling*, 132, 151-166.

DOE (Department of Environment, Malaysia), 1999. Classification of Malaysian Rivers. Volume 9: Pinang River. Draft Final Report: Project On Water Pollution Control: A Study to Classify Rivers in Malaysia (Phase V). (Mansor, M., Eng, L. P., Eng, S. C., Chan, N. W., Rawi, S. M., Rainis, R., Kadir, O. H. & Tan, E. (eds).Universiti Sains Malaysia, Pulau Pinang, Malaysia.

Karul, C., Soyupak, S., and Yurteri, C.(1999), Neural network models as a management tool in lakes, *Hydrobiologia*, 408/409,139-144.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., and Aulagnier, S.(1996), Application of neural networks to modelling nonlinear relationships in ecology, *Ecological Modelling*, 90,39-52.

Lek, S., Guegan J.F. (1999), Artificial neural networks as a tool in ecological modelling, An introduction, *Ecological Modelling* 120, 65-73.

Maier, H.R., and Dandy, G.C.(1996), The use of Artificial Neural Network for the prediction of water quality parameters, *Water Resources Research*, 32,1013-1022.

Recknagel , F. (ed) (2003), *Ecological Informatics: Understanding Ecology By Biologically Inspired Computation*, Springer Verlag.

Recknagel, F., French, M., Harkonen, P., and Yabunaka, K.-I. (1997), Artificial neural network approach for modelling and prediction of algal blooms, *Ecological Modelling*, 96,1-28.

Recknagel, F., and Wilson, H.(2000), Elucidation and prediction of aquatic ecosystems by artificial neuronal networks, In *Artificial Neuronal Networks: Application to Ecology and Evolution*, Springer Verlag, 143-156.

Talib, A., Abu Hasan, Y, and Varis O. (2008), Application of machine learning techniques in data mining of ecological datasets. *Proceedings of International Conference on Environmental Research and Technology ICERT 2008*, 677-681

Talib A. (2006), Comparative Ecological Study of Two Dutch Lakes Using Computational Modelling. PhD Thesis. University of Adelaide, Australia.

Wilson, H. (2004) Short-term forecasting of algal blooms in drinking water reservoirs using artificial neural networks. In *Environmental Biology*: The University of Adelaide. 299.