Techniques for predicting total phosphorus in urban stormwater runoff at unmonitored catchments

D. May* M. Sivakumar†

(Received 8 August 2003; revised 31 March 2004)

Abstract

This paper investigates the applicability of using artificial neural network (ANN) and multilinear regression models to predict urban stormwater quality at unmonitored catchments. Models were constructed using logarithmically transformed environmental data. Violation of the assumption of data independence lead to the inclusion of insignificant variables when a straightforward stepwise regression was applied. To overcome this problem, cross validation was used to determine when to stop adding variables. Regression models calibrated using event mean concentration (EMC) as the dependent variable were

*Sustainable Earth Research Centre, Environmental Engineering, University of Wollongong, NSW, AUSTRALIA. mailto:dbm05@uow.edu.au

†Sustainable Earth Research Centre, Environmental Engineering, University of Wollongong, NSW, AUSTRALIA. mailto:siva@uow.edu.au

See http://anziamj.austms.org.au/V45/CTAC2003/Mayd for this article, Published May 15, 2004. ISSN 1446-8735

more accurate than those using event load. Regression models developed on a regional subset of data were more accurate than the models developed on the entire data set. Even though regression and ANN models yielded similar predictions, regression modelling was considered to be a more applicable approach. Compared to ANN models, regression models were faster to construct and apply, more transparent and less likely to overfit the limited data.

Contents

1 Introduction

The adverse impacts of urbanisation on the aquatic environment have been recognised for many decades. In order to design appropriate measures to control polluted stormwater runoff, the extent of the problem must be known. The high costs associated with the collection and analysis of stormwater quality sampling data has created a demand for models capable of predicting urban stormwater quality at unmonitored catchments [1, 2, 3, 4]. Simple estimates of pollutant loads at unmonitored sites are often obtained using event mean concentration (EMC) values from sampling programs in similar regions. The high variability of water quality data observed at and between

sites has contributed to the limited success of these models [5]. Large national and regional databases have also been used to regress water quality variables against general geographic and climatic data. In a study by Driver and Tasker [3] U.S. data from the National Urban Runoff Program (NURP) was separated into three regions based on the mean annual rainfall of the sites. Three separate models for each water quality variable were then generated by linearly regressing the logarithmic transforms of the dependent and independent variables.

Artificial neural network (ANN) models are capable of modelling complex, nonlinear systems without prior knowledge of the exact relationships between variables [6]. The ability of ANN models to replicate nonlinear relationships makes them suitable for modelling environmental systems [7]. ANN models have recently been used in many water resources applications, including surface water quality forecasting and the prediction of chemical dosage in water treatment plants [8]. In this paper, the applicability of using ANN and regression models to predict urban stormwater quality at unmonitored sites was assessed.

2 Methodology

The data used in this study consisted of water quality, climatic and geographic data collected by the USEPA and USGS in the 1970's and 1980's [9]. The dependent variable analysed was total phosphorus (TP); measured as either a load or a concentration. The independent variables used in the following analyses that had values for every storm event are presented in Table 1. Maximum 24 hour precipitation intensity that has a 2 year recurrence interval (INT), measured in millimetres, was the only independent variable that had missing values.

The main objective of the study was to model storm events at typical urban watersheds. Catchments that contained uncharacteristic catchment

TABLE 1: Independent variables

Variable	Abbreviation	Unit
percentage of residential landuse	LUR	% drainage area
percentage of non urban landuse	LUN	% drainage area
percentage of commercial landuse	LUC	% drainage area
percentage of industrial landuse	LUI	% drainage area
impervious area	IA	% drainage area
drainage area	DA	ha
total event rainfall	TRN	mm
mean annual rainfall	MAR	mm

attributes beyond the scope of modelling were identified as potential outliers. Catchments with drainage areas greater than 3000 hectares, proportions of agricultural landuse greater than 50%, proportions of industrial landuse greater than 50%, population densities greater than 130 people per hectare or with detention basins upstream of the sampling point were removed. A total of 275 storm events were removed. A base ten logarithmic transformation was then applied to both the dependent variable and independent variables. A constant was added to variables that had zero values in order to scale the data into a suitable domain prior to logarithmic transformation. Numerous studies indicate that water quality variables follow lognormal distributions [3, 2, 5]. Logarithmic transformation of the data ensured that large, potentially outlying values did not bias the optimisation of calibration coefficients [3]. The other advantage of the logarithmic transformation was that it enabled the construction of nonlinear, nonadditive models using a simple multilinear regression procedure.

Error measure selection can influence the relative judgement of model performance. The standard error of estimate (SEE) and average absolute percentage error (AAPE) do not place considerable emphasis on the large

potentially outlying values, and allow the direct comparison between models constructed using load or concentration as the dependent variable when runoff volume is known. The standard error of estimate is a pseudo percentage error calculated from the mean square error in log (base 10) units:

$$\text{SEE} = 100 \times \sqrt{e^{[5.302\,\sigma^2]} - 1}\,. \tag{1}$$

The standard error of estimate places greater emphasis on the under prediction of large values than the average absolute percentage error. Both the standard error of estimate and average absolute percentage error were used to compare predictions from the constructed models.

Regression models were initially constructed using data from 754 storm events. This was to enable a direct comparison between load and concentration models on an equivalent domain. Stepwise multilinear regression models were developed using the logarithmically transformed data. If the p-value of a variable was greater than 0.05, the variable was entered into the model. Variables already in the model were removed if their p-value increased above 0.1 . The p-values represent the probability that the regression coefficient is not significantly greater than 0 . Regression models were created using either total phosphorus load or concentration as the dependent variable. Once a series of regression models had been developed (ranging from a simple one variable model to more complicated multivariable models) the standard error of estimate and average absolute percentage error were calculated for each model. The dependent variable producing the minimum error was analysed in more detail. Since multiple storm events were monitored at almost all of the catchments in the data set, the majority of independent variables did not satisfy the assumption of data independence. All analysed independent variables apart from total storm rainfall had constant values for a given catchment. This reduced the effective size of the data set, resulting in an underestimation of the p-value and the potential incorporation of spurious variables into the model. To overcome the problem, a cross validation approach was adopted. Ten disjoint data sets each containing approximately 10% of the data were created. Ten analyses were undertaken, with a different set being used for

validation purposes each time. The remainder of the data (90%) was used for model calibration. Inputs were sequentially entered into the models in the same order as a stepwise regression model calibrated on all the available data. The average absolute percentage error and standard error of estimate were calculated for each validation set and averaged. Variables leading to an increase in either of the error measures were typically considered to be insignificant and removed from the model.

A second regression analysis was undertaken on a regional subset of the data. Catchments with mean annual rainfalls between 500 and 1000 millimeters were separated from the total data set, in accordance with the study by Driver and Tasker [3]. The variables found to be significant in the study by Driver and Tasker were analysed along with the variables found significant in the cross validated, regression analysis of the larger data set. The variables were entered into the regression model in order of their anticipated significance. Cross validation using 10% of the data for validation was used to verify the significance of the independent variables. Variables not reducing the validation set errors were typically considered to be insignificant and removed from the model.

ANN models were constructed using the dependent and independent variables found significant in the regression analysis. The "pruning method" based upon the sensitivity analysis of constructed ANN models was perceived to be an excessively time consuming way to select ANN input variables. Feedforward, backpropagation neural networks were optimised using the normalised cumulative delta rule learning algorithm. The equation for the update of network weights is

$$\Delta w_{ji}(t) = \frac{1}{\sqrt{\varepsilon}} \sum_{s=1}^{\varepsilon} [\eta(d_j - y_j) f'(.) y_i] + \mu \Delta w_{ji}(t-1)\,, \tag{2}$$

where Δw_{ji} = weight update between nodes i and j at time t, η = learning rate, d_j = the actual output value, y_j = the predicted output value, $f'(.)$ = the derivative of the transfer function with respect to its input, ε = epoch

size, μ = momentum and s = the training sample presented to the network. An epoch equal to the training set size was selected. The logarithmically transformed data was scaled within the bounds of the hyperbolic tangent transfer function. The input variables were scaled between -1 and 1, and the output variable between -0.8 and 0.8. The learning rate for the weights connecting the input layer to the hidden layer and the hidden layer to the output layer were set at 0.04 and 0.02 respectively. Momentum was set at 0.01. Only one hidden layer was used, due to the limited amount of data available. The number of hidden nodes was determined by trial and error. The effect of learning rates and momentum on model accuracy was also analysed. Cross validation was used to determine when to stop training the network. The data was separated into ten disjoint sets, equivalent to those defined during the regression analysis. Ten ANN models were created, using a different 10% of the data as a test set each time. For each of the ten test sets, the mean square error was calculated for each weight update. An average of the mean square errors for the ten test sets was calculated for each weight update. The number of weight updates corresponding to the lowest average test set error was defined as the stopping point.

3 Results and Discussion

The results from the multilinear stepwise regression models constructed on the 754 data point set are presented in Table 2. Load models had errors more than 50% larger than the concentration models. Therefore concentration was used as the dependent variable during subsequent analyses. Stepwise regression models were then constructed using the 965 data point set in order to maximise the quantity of data used to construct the concentration model. Table 3 compares the results from the regression analyses using 100% and 90% of the data for calibration. The errors from the cross validation analysis using 90% of the data for calibration are prediction errors, whereas the errors from the analysis using all the data for calibration are calibration errors. The re-

TABLE 2: Comparison between concentration and load stepwise regression models developed on a 754 data point set

Concentration model			Load model		
Variable added	AAPE (%)	SEE (%)	Variable added	AAPE (%)	SEE (%)
MAR	89	107	DA	297	284
LUR	79	96	TRN	190	189
LUN	78	95	MAR	165	166
DA	77	94	LUN	150	159
TRN	77	93	LUC	140	153
LUI	76	92	LUI	134	152
			LUR	132	151

sults from the cross validation analysis showed that only mean annual rainfall and the percentage of residential landuse lead to improvements in both the standard error of estimate and absolute average percentage error. Therefore, only these two variables were deemed to be significant on the 965 data point set.

Regression equations were then developed on the regional subset consisting of 374 storm events. Results from the analysis of the larger data set justified the use of total phosphorus concentration as the dependent variable. The variables found to be significant in the study by Driver and Tasker [3] were total storm rainfall, total contributing drainage area, impervious area and maximum 24 hour precipitation intensity that has a 2 year recurrence interval. These variables were combined with mean annual rainfall and the percentage of residential landuse. Theoretical considerations combined with information extracted from stepwise regression models determined the order of variable entry into the final regression model. The results from the analyses using 100% and 90% of the data for calibration are presented in

TABLE 3: The effect of input addition on model error for regression models developed on the 965 data point set using total phosphorus concentration as the dependent variable

Calibration set	100%		90%	
Variable added	AAPE (%)	SEE (%)	AAPE (%)	SEE (%)
MAR	86.2	102.3	90.8	106.5
LUR	78.8	94.4	83.4	98.5
TRN	78.5	93.8	83.8	98.4
DA	77.9	93.4	83.9	98.7
LUN	77.5	92.5	83.9	98.2

Table 4. The cross validation analysis isolated drainage area as the only variable that did not improve either the average absolute percentage error or standard error of estimate. Therefore, drainage area was removed from the model. Impervious area and total event rainfall only reduced one error measure. However, when impervious area and total event rainfall were added together, both error measures reduced. Total event rainfall was also the only available variable in the data set capable of describing storm to storm variability at a site. Therefore, total event rainfall and impervious area were left in the model.

Artificial neural networks were developed to predict total phosphorus concentration. Two independent variables were not considered to be sufficient to construct ANN models on the 965 data point set. Therefore, ANN models were only constructed on the regional dataset using the significant inputs from the regional regression analysis. The optimum number of hidden nodes was shown to be 10. The standard error of estimate changed by less than two percent when the number of hidden nodes was varied between 5 and 15. Increasing the learning rates and momentum typically increased the speed

TABLE 4: The effect of input addition on model error for regression models developed on the regional subset using total phosphorus concentration as the dependent variable

Calibration set	100%		90%	
Variable added	AAPE (%)	SEE (%)	AAPE (%)	SEE (%)
INT	71.4	84.4	72.8	86.2
LUR	61.6	76.8	63.7	79.0
MAR	60.3	76.2	62.5	78.4
IA	59.3	75.5	61.4	78.6
TRN	59.3	75.2	61.6	78.2
DA	59.3	75.2	63.3	80.2

of the network convergence and the size of the error oscillations near the local minimum. Hidden layer learning rates greater or equal to 0.08 produced excessively large error oscillations around the local minimum. The standard error of estimate changed by less than 0.1% when the hidden layer learning rate was less than 0.06 . Varying the momentum also only changed the standard error of estimate by about 0.1%. In general, the selection of the number of hidden nodes influenced model accuracy more than the selection of learning rates or momentum.

Regression and ANN models were compared on the regional subset. The results presented in Table 5 suggest that regression and ANN models constructed on the regional subset had very similar accuracies. The regression model constructed using regional data was more accurate than the model constructed using all the available data. It was anticipated that a more complicated combination of relationships between variables was present within the larger data set. The lack of significant inputs restricted the ability of the regression model to replicate the complicated relationships.

TABLE 5: Validation results from concentration models compared on the regional subset

Model Type	Calibration Domain	AAPE (%)	SEE (%)	Input Variables
ANN	374	61	78	INT,LUR,MAR,IA,TRN
Regression	374	62	78	INT,LUR,MAR,IA,TRN
Regression	965	65	90	MAR,LUR

Data limitations in the current study were exacerbated by violations of data independence for the bulk of variables. Instead of analysing a large dataset equal to the number of storm events, a smaller subset equal to the number of catchments was effectively analysed. The effective size of the data set was approximately an order of magnitude smaller than the actual data set size. This made the modelled relationships tenuous, thereby decreasing the likelihood that ANN and regression models would accurately predict water quality at unmonitored sites. Inaccurate predictions are inevitable without the inclusion of a significant descriptor of storm to storm variability at a single site. Total event rainfall was not able to accurately define such variability. The comparable accuracies of the regression and ANN models constructed on the regional dataset inferred that the ANN model was not more adept at defining storm to storm variability at a site. This inferred that ANN models constructed on the total dataset would probably require additional storm descriptors, which were generally unavailable at a large proportion of the studied catchments. The inclusion of additional storm descriptors would have further reduced the size of the dataset, thereby limiting the applicability of applying ANN. The construction of an ANN model on the entire dataset using of all available variables might produce more accurate results. However, the potential inclusion of superfluous variables was perceived to reduce the accuracy of the final models and make it difficult to isolate significant variables. The identification of additional synergistic relationships between

the existing variables was considered to be overly time consuming compared to the benefit extracted from the identification of such relationships.

4 Conclusions

It was found that models using concentration as the dependent variable were more accurate than those using load. This was an important finding considering that the majority of current computer simulation models require estimates of concentration rather than load. When load was used as the dependent variable, the regression models were forced to simulate the known relationship existing between load and runoff volume, leading to an unnecessary increase in the complexity of the models. However, if the volume of runoff is not accurately known, load models might provide better estimates of the total load than the concentration models. Regression models constructed using the total data set were less accurate than those constructed on the regional subset of data. The reduced data complexity combined with the use of additional variables contributed to the increased accuracy of regression models constructed on the regional subset.

Violation of the assumption of data independence significantly reduced the applicability of constructing models on the larger data set. Total event rainfall was the only variable capable of describing storm to storm variation at a single catchment. However, total event rainfall was deemed to be insignificant on the larger data set. This meant that the effective size of the larger data set was too small to successfully apply ANN. In general, the regression equations were shown to be a more applicable approach on the regional subset. The simple form of regression models made them quick to construct and less likely to overfit the data.

Acknowledgments: The authors express their gratitude to Pam Davy for her assistance during the course of the research.

References

[1] C. S. Barks. Adjustment of regional regression equations for urban storm-runoff quality using at-site data. *Transportation Research Record*, vol. 42, pages 141–146, 1996. C297

[2] P. L. Brezonik and T. H. Stadelmann. Analysis and predictive models of stormwater runoff volumes, loads, and pollutant concentrations from watersheds in the Twin Cities metropolitan area, Minnesota, USA. *Water Research*, vol. 36, pages 1743–1757, 2002. C297, C299

[3] N. Driver and G. Tasker. Techniques for estimation of storm-runoff loads, volumes, and selected constituent concentrations in urban watersheds in the United States. *United States Geological Survey Water-Supply Paper 2363*, pages 177–246. United States Government Printing Office, Washington, 1990. C297, C298, C299, C301, C303

[4] L. Sliva and D. D. Williams. Buffer zone versus whole catchment approaches to studying land use impact on river water quality. *Water Research*, vol. 35, no. 14, p. 3462–3472, 2001. C297

[5] J. T. Smullen, A. L. Shallcross and K. A. Cave. Updating the U.S. nationwide urban runoff quality data base. *Water Science and Technology*, vol. 39, no. 12, pages 9–16, 1999. C298, C299

[6] S. Lek, M. Guiresse and J. Giraudel. Predicting stream nitrogen concentration from watershed features using neural networks. *Water Research*, vol. 33, no. 16, pages 3469–3478, 1999. C298

[7] H. R. Maier and G. C. Dandy. The effect of internal parameters and geometry on the performance of back-propagation neural networks: and empirical study. *Environmental Modelling & Software*, vol. 13, pages 193–209, 1998. C298

[8] H. R. Maier and G. C. Dandy. Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. *Environmental Modelling & Software*, vol. 15, pages 101–124, 2000. C298

[9] Cahaba/Warrior Student Chapter of the American Water Resources Association (University of Alabama) NURP data. (updated 5th March 1998). [Online] http://www.eng.ua.edu/~awra/download.htm C298