Use of Artificial Neural Networks for Modelling Chlorine Residuals in Water Distribution Systems

M. S. Gibbs[a], N. Morgan[a], H. R. Maier[a], G. C. Dandy[a], M. Holmes[b] and J. B. Nixon[b]

[a]Centre for Applied Modelling in Water Engineering, School of Civil and Environmental Engineering, The University of Adelaide. E-mail: mgibbs@civeng.adelaide.edu.au

[b]United Water International Pty Ltd, Adelaide, Australia

Abstract: Drinking water contaminated by microorganisms can be a major risk to public health. Disinfection is used to destroy microorganisms that are potentially dangerous to humans. In order to prevent bacterial regrowth, it is also desirable to maintain a disinfectant residual in the water distribution system. The most commonly used disinfectant is chlorine. If the dosing rate of chlorine is too low, there may be insufficient residual at the end of the distribution system, resulting in bacterial regrowth. On the other hand, the addition of too much chlorine can lead to customer complaints about taste and odour, corrosion of the pipe network and the formation of potentially carcinogenic by-products. Consequently, in order to determine the optimal chlorine dosing rate, it is necessary to be able to predict chlorine decay in the network. In this paper, two data-driven techniques, namely linear regression models and multi layer perceptron artificial neural networks, are used to predict chlorine concentrations at two key locations in the Hope Valley water distribution system, which is located to the north of Adelaide, South Australia. A 5-year data set containing routinely measured parameters is used for model development and validation. The results obtained indicate that both techniques are relatively successful in predicting chlorine concentrations in the distribution system. This is despite the fact that there is no hydraulic model of the system and that only data that were collected on a routine basis are used for model development. Overall, the performance of the multi layer perceptron is slightly better than that of the regression model, suggesting the presence of some non-linearities in the underlying physical processes governing chlorine decay.

Keywords: *Water distribution system; Chlorine; Artificial neural network; Modelling*

1. INTRODUCTION

Drinking water contaminated by microorganisms can be a major risk to public health. Disinfection is the water treatment process carried out to destroy any harmful microorganisms that are contained in the drinking water. In order to prevent bacterial regrowth, it is desirable to maintain a disinfectant residual in the network.

Chlorine is the most commonly used disinfectant due to its ease of application and monitoring, its low cost and its effectiveness in killing bacteria (Hua et al., 1999). However, as a result of chlorine reacting with various substances in the water and on the pipe wall, its concentration can decrease as it travels through the distribution system. This is known as chlorine decay. The amount of chlorine added to the water is very important. If the dosing rate is too low, there may not be a residual left at the end of the distribution system to protect against recontamination. If the dosing rate is too high, it can lead to customer complaints, corrosion of the pipe network or the formation of byproducts, including trihalomethanes (THMs), which are suspected carcinogens.

Generally, the chlorine dosing rate at a water treatment plant is determined from operator knowledge and by monitoring residual chlorine concentrations and coliform levels in the distribution network. This is a sub-optimal method of operating, as the dosing rate is adjusted only after the chlorine concentration in the field is detected to lie outside a desirable range. It would be beneficial to accurately predict the chlorine dosing rate that is required to achieve a balance between sufficient chlorination to ensure bacteriological quality and providing customers with water that they find pleasant to drink.

Traditionally, chlorine decay has been predicted using process-based models that assume chlorine decay follows a first order equation. The main advantage of process-based models is that they are based on the underlying physical processes, so

the results obtained generally have a wide range of applicability. To develop a process-based model, a good understanding of the system is required along with extensive, accurate data to produce the hydraulic model used to determine travel times of water in the system. Flows in individual pipes, as well as the values of constants required for the chlorine decay model (which can be dependent on many factors, including temperature, initial chlorine concentration, source water quality and biofilm presence), must all be determined.

An alternative modelling procedure consists of data-driven statistical models. Statistical chlorine decay models can be used to predict residual chlorine based on empirical relationships between a number of dependent and independent variables. The main difference between a statistical model and a process-based model is that statistical models are driven by observed relationships in the data, rather than an assumed knowledge of the actual process occurring in the system. The development of statistically based models for disinfection control purposes is justifiable in cases where parameter estimation within a process-based model is imprecise or difficult to obtain (Rodriguez et al., 1997) or where the data required for the development of process based models are not available. This approach offers the advantage of not requiring extensive *a priori* knowledge of the laws of chemistry and mathematics governing the behaviour of residual chlorine (Sérodes et al., 2001) or the distribution system being studied. However some knowledge of the factors that will influence the chlorine decay can help identify which data are relevant for the analysis.

Data-driven modelling approaches are becoming more popular due to the increasing availability of data in the water industry. Water utilities possess large quantities of data derived from control and monitoring facilities. Rather than devising data collection schemes to collect the large amounts of data required to develop a process-based model, statistical techniques can be applied to extract useful relationships from existing data sets, thus making maximum use of the data that are already available.

The objective of this research is to assess the feasibility of using data-driven models for determining the chlorine decay in a water distribution network to aid in optimising chlorine dosing rate at the Hope Valley water treatment plant, South Australia. Two approaches will be considered, namely linear regression and artificial neural networks (ANNs). ANNs are used due to their ability to handle nonlinearity and large amounts of data, as well as their fault and noise tolerance and their learning and generalisation capabilities (Lawrence, 1994). Regression is used as a benchmark against which the performance of the ANN can be compared, as there has been a notable lack of research comparing the performance of ANNs with more conventional statistical approaches (Dawson et al., 2001). Linear regression was used in preference to non-linear regression as the functional form of the process to be modelled must still be assumed when the latter modelling technique is used. Although the functional form of chlorine decay with time is resonably well known, the relationship between chlorine other input parameters, such as organic content, is largely unknown.

2. PREDICTIVE MODELLING THEORY

2.1. Linear Regression Analysis

To perform a linear regression, the chlorine concentration, Y, is assumed to be a linear function of the inputs, X. The unknown parameters to be determined, a_i are the coefficients, as given in (1):

$$Y=a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n \quad (1)$$

Where n is the number of inputs used. The coefficients are chosen so as to minimise the sum of the squared differences between the predicted and actual values of Y.

2.2. Multi Layer Perceptron

The multi layer perceptron (MLP) using the backpropagation training algorithm is the most widely used neural network for forecasting and prediction applications (Maier et al., 2000). MLPs generally consist of three layers: an input layer, a hidden layer and an output layer, as shown in Figure 1. However, MLPs may contain more than one hidden layer.

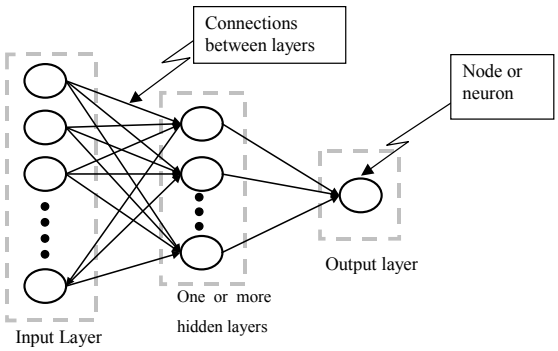

Figure 1. General Structure of MLPs.

Each layer consists of nodes or neurons, which are connected to nodes in the previous and following layers by connections. The strength of each connection, referred to as its connection weight, can be adjusted. The connection weight from the i^{th} node to the j^{th} node is denoted by w_{ij}.

Input data are presented to the network through the input layer, the values of which are denoted by x_i. Data are passed from the input layer to the hidden layer. Each node in the hidden layer receives the weighted outputs ($w_{ij}x_i$) of the nodes in the preceding layer. These outputs are then summed and added to a threshold value, θ_j, to produce the node input, I_j, as shown in (2).

$$I_j = \sum_i w_{ij} x_i + \theta_j \qquad (2)$$

The node input is then passed through an activation function, $f(I_j)$, to produce the node output, y_j. This node output is then used to compute the inputs for nodes in the following layer, until the final output is calculated.

3. CASE STUDY

3.1. Hope Valley Water Distribution System

The case study considers chlorine residuals and chlorine consumption in the Hope Valley water distribution system, Adelaide, Australia. Hope Valley was Adelaide's first water treatment plant (WTP). It has a design capacity of 273 ML/day and serves a population of approximately 180,000, making it the third largest WTP in Adelaide.

A 5-year data set collected by United Water International was used for this project. Chlorine concentrations were predicted at two locations in the water distribution system, sampling points 1064 at the Queen Elizabeth Hospital on Woodville Road and 1066 at the Port Adelaide Primary School on Portland Road, as shown in Figure 2. Each dot in the figure represents a sampling point in the network.

The network parameters available included the water temperatures at the WTP and at the sites of interest, flow from the WTP, chlorine concentrations throughout the network, the dissolved organic carbon (DOC) content, UV light absorbance, the time at which each measurement was taken and the previous week's chlorine concentration at the site. The measurements at points 1064 and 1066 were taken weekly, which dictated the frequency of the patterns in the data sets. Some parameters were measured daily, such as the WTP flow and temperature, which resulted in a week of flow measurements available as input parameters for each chlorine concentration at the site being predicted. Other measurements, such as DOC, were measured fortnightly or monthly.

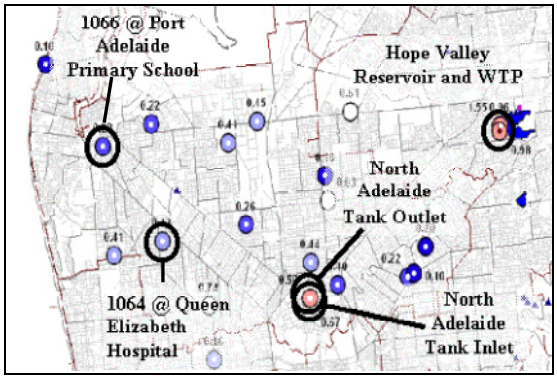

Figure 2. Location of chlorine sampling points.

3.2. Predictive Model Development

The systematic modelling procedure implemented in this research can be seen in Figure 3. The main steps involve data preparation, input selection, data division, model selection, model calibration and performance evaluation. The analytical techniques used to help with the input determination process were the coefficients of correlation (CC), sensitivity analysis (SA) and partial mutual information (PMI). The division of data for use in the model calibration and validation steps was implemented using a self organising map (SOM). The two types of models used were linear regression and multi layer perceptrons (MLPs). Model calibration was performed using the "Least Squares" method for the linear regression, and the backpropagation algorithm for the MLP. Performance evaluation was then used to test the accuracy of each calibrated model, which included the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and the maximum error produced by a prediction from the model (Max). Each step of the process is outlined in the following section.

3.3. Input Determination

If redundant input parameters can be identified and removed then model size can be reduced significantly. Using fewer input parameters will also decrease the amount of measured data required, which will effectively reduce the amount of noise (measurement errors) introduced into the model. This can lead to a more efficient and accurate model for predicting chlorine levels. There are many different methods that can be used to determine which inputs should be used in a predictive model. *A priori* knowledge of a

system can be used to identify significant inputs if the system being studied is well understood. Inspection of time series plots can also be used to identify input/output relationships. However, if the system is not well understood, analytical techniques can be used (Maier and Dandy, 2000).

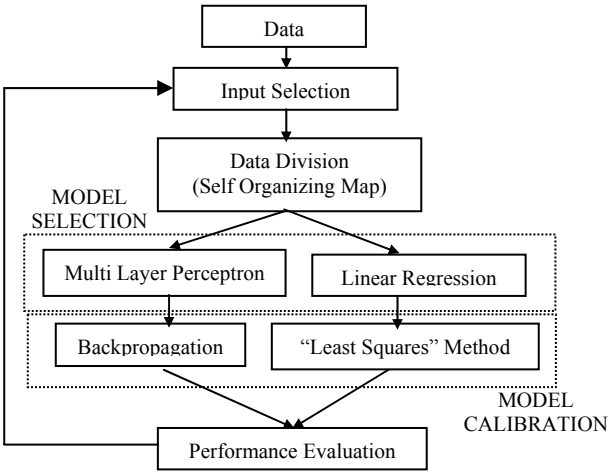

Figure 3. Modelling methodology used.

Three different techniques were adopted in this research to gain an appreciation of the inputs required for this study. They were correlation coefficients, sensitivity analysis and partial mutual information (Sharma, 2000). By using a combination of *a priori* knowledge and the results of the analytical procedures, the most significant inputs for the Hope Valley water distribution network were selected for model development. These inputs included the water temperature at the sites of interest and at the WTP, the WTP flow 5 days before the chlorine measurement being predicted, and the chorine concentrations at the WTP, at the North Adelaide tank inlet (see Figure 2) and at the site under prediction the previous week, resulting in a total of 6 model inputs.

3.4. Data Division

The development of any model requires the partitioning of the parent database into statistically similar subsets in order to calibrate and validate models. The method of data division used in this project was based on the method proposed by Bowden et al. (2002), using a self organizing map (SOM). The Kohonen Self Organising Map (SOM) is a type of unsupervised neural network, which produces a topologically ordered output based on statistical similarities (such as the mean and variance) between input patterns. Using the SOM the parent dataset was divided into 80% calibration and 20% validation. The validation set was set aside in the calibration process, as it was only used for the comparison of results, to provide an unbiased assessment. This method has advantages over conventional methods as it ensures the subsets are representative of the same population (Bowden et al., 2002). To ensure the MLP does not become overtrained the calibration set was further divided into two subsets using the SOM technique, 80% for training and 20% for testing. The testing set was used to decide when to stop training in order to avoid overfitting and to decide which model geometry and internal parameters are optimal.

3.5. Model Selection

Linear Regression Analysis

Regression analysis was performed using the Data Analysis Tool-pack in Microsoft Excel. A 95% confidence level was used to calculate the regression coefficients. The regression coefficients were determined using the calibration data set (testing and training), and the coefficients found were implemented in the linear regression equation to predict the measured values in the validation data set.

Multi Layer Perceptron

The commercially available neural network software package Neuframe 4.0 was used for the development of the MLP. The network uses the backpropagation algorithm to optimise connection weights. The input layer transfer function was determined to be linear, while sigmoidal logistic functions were used for the hidden and output layers. Cross-validation was used to determine when training should be stopped to prevent overfitting.

A scaling interval of 0.1 – 0.9 was used. Optimal learning rates, momentum values and network geometries were determined using trial and error. Learning rates and momentum values ranging from 0.05 to 0.95 were tried. The optimum values were found to be 0.25 for the learning rate and 0.75 for the momentum value for all models. The optimal number hidden nodes was found to be 3 for the models predicting chlorine demand and chlorine consumption at location 1064 and 8 for the models developed for location 1066. Using different random number seeds to initialise the connection weights was found to have a minimal impact on predictive accuracy.

4. RESULTS

Figure 4 shows the MLP consistently outperformed the linear regression model in all the error measurements used when predicting the

chlorine consumed from the WTP to point 1064. A similar result was found for the chlorine residual predictions at this point. The MLP also outperformed the linear regression at point 1066 for both chlorine residual and consumption predictions, however the result was not as pronounced. Model performance was evaluated using the root mean square error (RMSE), the mean absolute error (MAE) and the maximum absolute error (Max). When comparing the prediction accuracy between two models the RMSE was considered the best indicator, as this error measurement penalises larger prediction errors more harshly than the MAE. The maximum absolute error for the models developed were also determined, as this gives an indication of the worst case prediction made by that model.

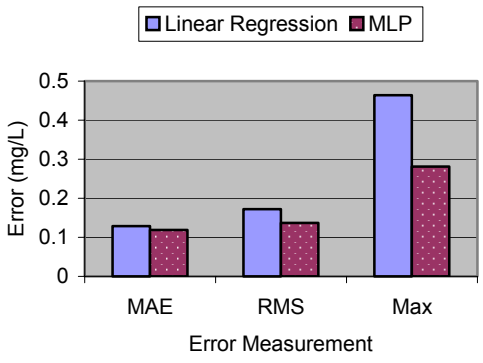

Figure 4. Summary of chlorine consumption prediction errors at point 1064.

All models produced a lower minimum RMSE at point 1066 than at point 1064. This may be due to the presence of complex processes not well explained at point 1064, such as moving mixing zones. For point 1066, the two models performed similarly based on the evaluation criteria used for both residual and consumption predictions. This may be due to a relatively simple relationship in the available data between the input parameters used and the chlorine concentrations at point 1066.

The scatter plot for predicting chlorine consumption at 1064 with the MLP is shown in Figure 5. The solid line shown on the plot is a linear trendline fitted to the predicted values, the dashed lines indicate ± 0.2mg/L errors from the line of perfect predictions, indicated by the black line. It can be seen that there does not appear to be any general trend to the errors in the predictions, whereas the regression model tended to over-predict the low values and under-predict the high values. These results imply there is a significant non-linear relationship governing the processes, which, unlike the linear regression analysis, the MLP is able to predict.

Regression, being the simpler and easier model to implement, provided a good method to obtain quick predictions of chlorine concentration trends, but struggled to predict the extreme measurements, as indicated by a large average maximum error at both points.

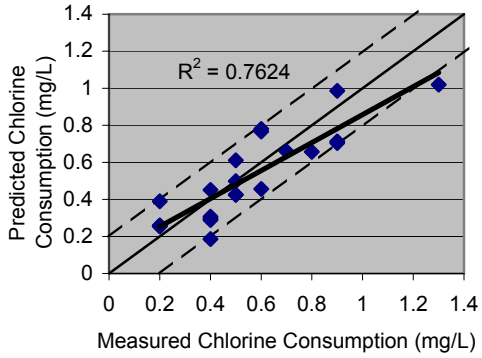

Figure 5. Scatter plot of the chlorine consumption for 1064 using the MLP model.

5. DISCUSSION

Chlorine measurements in the network range from 0.7mg/L to <0.1mg/L at point 1064 with measurements occurring at 0.1mg/L intervals. This corresponds to an average measurement accuracy of 14%. Model performance is penalised as the model output is continuous, whereas the measured values are discrete. For example, a prediction of 0.45mg/L will be penalised regardless of the measured value of chlorine concentration, which is a multiple of 0.1. Therefore, the accuracy of the chlorine measurements may be restricting the predictive accuracy of the models, as in the calibration process the model parameters will be adjusted, possible incorrectly, to conform to the measured values. A possible solution to this problem could be to allow a tolerance before any error is calculated.

It is likely that there is a significant daily variation in chlorine concentrations at the sampling locations. The diurnal variation in demand will result in varying travel times. The variation in flow during a day may also have an effect on the dosing rate of chlorine. When the demand changes rapidly, the response time for adjusting the chlorine dosing rate may cause plugs of water, containing either high or low concentrations of chlorine, to move through the system. Changing demands will also affect the path the water may take from the WTP to the sampling location. The variation in factors such as these are most likely poorly represented by the available data. Hence the models developed

assume these occurrences are constant and do not affect the chlorine concentration, which is most likely an invalid assumption.

In terms of assessing the applicability of the different modelling approaches for use in controlling disinfection, a number of outcomes were achieved. The linear regression model showed that it had the potential to predict the general trends of chlorine evolution in a distribution system. The principal advantages of the linear regression analysis are the speed of calibration and the small number of parameters that required optimisation. Based on prediction accuracy, the MLP is the more appropriate model for assisting in disinfection control. However, the difficulties associated with the calibration of this more complex model may outweigh the benefits of prediction accuracy when compared with linear regression.

The results obtained indicate that it is quite possible to predict the chlorine consumed in a water distribution network using intelligent data driven methods such as neural networks. Rodriguez et al. (1997) implemented a single smoothing factor general regression neural network to predict the chlorine residual in a simple trunk main without offtakes using hourly chlorine measurements. The Hope Valley system that has been considered in this case study is much more complex and not well understood, but even with the weekly time step between chlorine measurements, the models implemented managed to predict the general trends of chlorine concentrations. The results indicate the potential to implement these methods to assist Water Treatment Plant management in determining the optimal chlorine dosing rate.

6. CONCLUSIONS

The findings of this research suggest that the factors that are important for the prediction of chlorine concentrations for the case study considered are the North Adelaide tank inlet chlorine concentration, temperature at the sampling location, WTP chlorine concentration, WTP temperature, WTP flow and the previous chlorine concentration measurement at the sampling point.

It has been shown that a data-driven modelling approach is a suitable method for estimating the disinfectant concentrations in a water distribution network, especially in the case where the network is not well understood and the necessary data for a process based model are not available.

The multi layer perceptron model was found to consistently outperform traditional linear regression for this case study. The MLP has displayed the potential to be implemented as an online tool to aid in the determination of chlorine dosing rates in a water distribution network.

7. ACKNOWLEDGMENTS

Dr Chris Chow (The Australian Water Quality Centre, Adelaide, Australia) for the kind collaboration and assistance. The authors would also like to thank the CRCWQT for funding this research.

8. REFERENCES

Bowden, G. J., Maier, H. R. and Dandy, G. C., Optimal division of data for neural network models in water resources applications, *Water Resources Research*, 38(2), 2.1-2.11, 2002.

Dawson, C. W. and Wilby, R. L., Hydrological modelling using artificial neural networks, *Progress in Physical Geography*, 25(1), 80-108, 2001.

Hua, F., West, J. R., Barker, R. A. and Forster, C. F., Modelling of chlorine decay in municipal water supplies, *Water Research*, 33(12), 2735-2746, 1999.

Lawrence, J., *Introduction to Neural Networks. Design, Theory, and Applications*, California Scientific Software Press, Nevada City, 1994.

Maier, H. R. and Dandy, G. C., Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, *Environmental Modelling & Software*, 15(1), 101-124, 2000.

Rodriguez, M. J., West, J. R., Powell, J. and Sérodes, J. B., Application of two approaches to model chlorine residuals in Severn Trent Water Ltd (STW) distribution systems, *Water Science and Technology*, 36(5), 317-324, 1997.

Sérodes, J. B., Rodriguez, M. J. and Ponton, A., Chlorcast (c): a methodology for developing decision-making tools for chlorine disinfection control, *Environmental Modelling & Software*, 16(1), 53-62, 2001.

Sharma, A., Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - A strategy for system predictor identification, *Journal of Hydrology*, 239(1-4), 232-239, 2000.