Prediction of multi-components (chlorine, biomass and substrate concentrations) in water distribution systems using artificial neural network (ANN) models

Celia D. D’Souza and M. S. Mohan Kumar

ABSTRACT

Artificial Neural Networks (ANN) models are used to predict residual chlorine, substrate and biomass concentrations in a Water Distribution System (WDS). ANN models with different architectures are developed: a one output ANN model (predicting chlorine, substrate and biomass individually), a two output ANN model (predicting chlorine + substrate, chlorine + biomass or substrate + biomass) and a three output ANN model (chlorine + substrate + biomass). This study is carried out for the Bangalore City and North Marin WDSs. Data for these WDSs is obtained from the multi-component reaction transport model. The models are compared using the correlation coefficient (R) and the Mean Absolute Error (MAE). The models developed are able to predict, reasonably well, the temporal variations in the chlorine, substrate and biomass concentrations. Error analysis is carried out to determine the robustness of the models.

Key words | ANN, biomass and substrate concentrations, chlorine, Levenberg-Marquardt algorithm, water distribution systems, water quality

INTRODUCTION

Residual chlorine is a primary indicator of the water quality in a Water Distribution System (WDS). It must be maintained in sufficient quantity to guarantee microbiologically safe drinking water (Rodriguez et al. 1997). Residual chlorine attenuates due to its reaction with organic and inorganic nutrients and by its removal of some of the bacteriological species that may be present in water (Clark 1998). Hence at the periphery of a WDS there is low or zero residual chlorine and a high risk of bacterial regrowth (Falkinham et al. 2001).

Deterministic models have been developed to evaluate and generate different water quality scenarios in a WDS (Chungsying et al. 1994; Rauch et al. 1999; Munavalli & Mohan Kumar 2004a). They help to understand the processes involved in the attenuation of residual chlorine, the regrowth of microbial contaminants and substrate utilization. These models are based on the transport and reactions of the multi-components (chlorine, biomass and substrate). However calibration of these deterministic models is tedious and challenging due to the large number of parameters involved.

In this study artificial neural network (ANNs) are used to predict the multi-components in a WDS. ANNs are data-driven techniques that have remarkable information processing characteristics and can handle nonlinearity. They have fault and noise tolerance as well as learning and generalization capability (Basheer & Hajmeer 2000). The development of ANN models does not require a priori knowledge of the physical and chemical laws governing the parameters to be determined.

A few studies have focused on the role of ANNs in assessing water quality in a WDS. These include the prediction of tri-halo-methane (THM) formation (Rodriguez et al. 2003), the prediction of residual chlorine in storage tanks (Serodes & Rodriguez 1996) and the prediction of residual chlorine in WDS (Rodriguez et al. 1997;
The work of Rodriguez et al. (1997) in the prediction of the temporal variation of residual chlorine in a system has been further extended in this study to predict the temporal variations of also substrate and biomass concentrations. The prediction of multi-components and the use of different architectures of ANNs has so far not been reported in the literature. The objective of this paper is to predict the temporal variation of chlorine, substrate and biomass concentrations in the Bangalore City and North Marin WDSs. The training dataset for the development of different ANN models is obtained from the multi-component reaction transport model. Unseen data is used to test the robustness of the ANN models. The different ANN models are compared using the correlation coefficient (R) and the mean absolute error (MAE).

**DEVELOPMENT OF THE MODELS**

The development of a ANN model involves the selection of inputs and, a training algorithm, the determination of the size of the ANN and the testing of the developed model with unseen data.

**Selection of input data**

Most of the inputs are chosen based on their cause-effect relationship with the output parameters. These inputs are source chlorine, substrate and biomass concentrations, source inflow; water consumption and residence time. Source chlorine lag effects are chosen as inputs to assist in the convergence of the model. These lag effects quantify the impact on the current time period of the chlorine concentrations during prior time periods. The number of lag effects considered is determined by the average residence time of water.

Water quality data is not available for the two WDSs. Hence the dataset used to develop the ANN models is generated from the multi-component reaction transport model (Munavalli & Mohan Kumar 2004a). The dataset for the Bangalore City WDS is obtained by running the multi-component reaction transport model on actual hydraulic data and hypothetically assumed temporal variations of source chlorine, substrate and biomass concentrations. In the case of the North Marin WDS the dataset is obtained by running the multi-component reaction transport model on actual hydraulic data and source chlorine concentration while hypothetically assuming the temporal variations in substrate and biomass concentrations.

The multi-component reaction transport model was developed by Munavalli & Mohan Kumar (2004a) to determine chlorine, substrate and biomass concentrations spatially and temporally for WDSs. Separate mass balance equations for plug flow were developed for each of the three components in the bulk and wall zone of pipe sections. The equations describe processes such as bacterial growth and decay, attachment to and detachment from the pipe surface, substrate utilization for bacterial regrowth and disinfectant action on substrate and bacteria. Here substrate refers to the BDOC (Biodegradable Dissolved Organic Carbon). It is a portion of the TOC (Total Organic Carbon) that can be mineralized by micro-organisms. Bacterial growth refers to the multiplication of HPC or Coliform bacteria. In this model, all the living bacterial content is referred to as biomass. The mass balance equations are solved using the Lagrangian method; a time-driven approach. The equations and the numerical methods used to solve them are explained and discussed by Munavalli & Mohan Kumar (2004a,b). For each WDS the model was run with three different hydraulic loads and source residual chlorine, substrate and biomass concentrations. The objective was to increase the number of data-points in the dataset.

**Training algorithm**

Back propagation feed forward neural networks are widely used in forecasting and predicting applications (Maier & Dandy 2000). In this paper, the Levenberg-Marquardt algorithm is used. This algorithm provides a numerical solution to the problem of minimizing a function, generally nonlinear, over a space of parameters of the function. It is one of the most efficient training algorithms and has a high speed of convergence. The Levenberg-Marquardt algorithm is an approximation of Newton’s method. Its adaptation used in the training of neural networks is much more efficient than the usual gradient techniques (Costa et al. 2007).
The weight update equation of the Levenberg-Marquardt algorithm is

\[ w_{k+1} = w_k - \left[ J^T J + \mu I \right]^{-1} J^T e \]  

(1)

where \( w_{k+1} \) is the vector of the new weight values after an iteration, \( w_k \) is the vector of the current weight values, \( J \) is the Jacobian matrix containing the first derivative of the error with respect to the weights, \( \mu \) is a scalar controlling the learning process and \( e \) is the network output error vector for every input data. For large values of \( \mu \) the performance of the algorithm is comparable with the steepest gradient algorithm and for small values of \( \mu \) the performance of the algorithm is comparable to Newton’s method.

Pre-processing techniques

The number of inputs used in the prediction of chlorine, biomass and substrate concentrations is quite high. Hence the number of the inputs has been reduced by carrying out Principal Component Analysis (PCA) on the input vector. The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables (Jolliffe 2002).

Architecture of the ANN models

Three ANN architectures have been utilised in this study. In the first type of architecture, the output layer has one output with ANN models predicting individually the temporal variations in chlorine, substrate and biomass concentrations. In the second type of architecture, the output layer has two outputs. ANN models have been developed to predict chlorine and biomass; chlorine and substrate; and substrate and biomass concentrations. In the third type of architecture, the output layer has 3 three outputs. A single ANN model has been developed to simultaneously predict the temporal variations of all the three components—chlorine, substrate and biomass concentrations.
Calibration and testing of models developed

There are four parameters to be determined in the development of the ANN models. These are the number of hidden layers, the number of neurons in each layer, the learning rate and the momentum rate. There are certain rules of thumb mentioned in the literature \((\text{Basheer \\& Hajmeer 2000})\) which serve as guidelines in the initial determination of the number of hidden layers and the number of neurons in each of the hidden layers. The common approach is to start off with a single hidden layer with a minimum number of neurons. The number of hidden layers and the number of neurons in each layer is increased till the MSE (Mean Squared Error) of the training dataset is met.

The general values of learning and momentum rates given in the literature vary between 0.1–10 and 0.4–0.9 respectively \((\text{Wythoff 1993})\). The learning and momentum parameters are problem dependent and are normally determined by a trial and error procedure.

Each of the models developed is trained for approximately 75% of the entire dataset and then tested for the remaining 25% of the dataset. The code for the training and testing of all ANN models was written using MATLAB.

### CASE STUDY I—BANGALORE CITY WDS

#### Description of the system

The Bangalore City WDS considered has 87 nodes and 94 elements. Of the 87 nodes, 15 are reservoir nodes while of the 94 elements, eight are pump elements (see Figure 1). There are three water supply sources to the city Cauvery Stage I (node 1), Cauvery Stage II (node 2) and Arkavathi Scheme (node 3) which supply 1.577 m³/s, 1.682 m³/s and 1.577 m³/s respectively under normal operating conditions. The details of the system are given in \text{Datta \\& Sridharan (1994)}.

#### Preparation of inputs for the models

The Bangalore City WDS is divided into groups of nodes. From within these groups random selection is used to obtain 24 nodes for training and 6 nodes for testing of the models.(see Figure 1). The nodes were grouped prior to selection so that the entire WDS is fairly represented by the randomly selected nodes. Nodes representing reservoirs are not considered since the mechanisms governing the decay of residual chlorine, the utilization of substrate and biomass regrowth are different from the junction nodes. Hourly data of chlorine, substrate and biomass concentrations is obtained by running the multi-component reaction transport model for a simulation period of 24 hrs. Since three cases are considered, with hourly data obtained at every node, the total training dataset consists of 1728 (i.e. \(24 \times 24 \times 3\)) datapoints. The testing dataset consists of 432 (i.e. \(6 \times 24 \times 3\)) datapoints.

The total number of inputs to the model is 23 (inputs for source chlorine, biomass, substrate and inflow for each source node, 1 input each for water consumption and residence time and 3 lags of residual chlorine concentration for each of the source nodes). PCA is carried out on the entire dataset and the first 6 principal components are chosen. These principal components account for 99% of the variance of the entire input dataset.

### Table 1 | Architecture, R and MAE values for models of Bangalore city WDS

<table>
<thead>
<tr>
<th>Type</th>
<th>Architecture</th>
<th>R</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(C) (X) (S)</td>
<td>(C)</td>
<td>(X)</td>
</tr>
<tr>
<td>(C, S, X)</td>
<td>6-10-1</td>
<td>6-10-9-1</td>
<td>6-13-1</td>
</tr>
<tr>
<td>(C + X)</td>
<td>6-10-12-2</td>
<td>6-10-12-2</td>
<td>–</td>
</tr>
<tr>
<td>(C + S)</td>
<td>6-11-12-2</td>
<td>–</td>
<td>6-11-12-2</td>
</tr>
<tr>
<td>(X + S)</td>
<td>–</td>
<td>6-10-9-2</td>
<td>6-10-9-2</td>
</tr>
<tr>
<td>(C + S + X)</td>
<td>6-10-12-3</td>
<td>6-10-12-3</td>
<td>6-10-12-3</td>
</tr>
</tbody>
</table>

Note: C—Chlorine, S—Substrate, X—Biomass.
Results and discussions

The number of hidden layers and the number of neurons in each of the hidden layers for each model is determined by the performance function (MSE) of the model. The architecture, R and MAE values for all the models, using the testing dataset, are given in Table 1. For most of the ANN models the learning rate is found to be in the range of 0.45–0.5. The momentum rate is found to be in the range of 0.75–0.8.

The results in Table 1 indicate that all the models (one output, two output and three output) are able to predict the temporal variations of the multi-components reasonably well.

The three output model gives the best results followed by the two output model and the one output model. Since the three components are inter-dependent, the three output model is better suited to predict the temporal variations of the multi-components in a WDS.

Figure 2 | Variation of chlorine biomass and substrate for various models for Bangalore city WDS.
Error analysis

The robustness of the models is tested with noisy data. Noisy data is generated by introducing shocks to the inputs in the form of a random error of about 10%. This random error is introduced in one node of the training dataset and one node of the testing dataset. The error is introduced only in the inputs representing source chlorine, substrate and biomass concentrations. Plots of the temporal variations of the three components are shown in comparison with the error free dataset (Figure 2a–f). The results

<table>
<thead>
<tr>
<th>Type</th>
<th>Architecture</th>
<th>R</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C, S, X)</td>
<td>18-8-9-1</td>
<td>0.83</td>
<td>0.137</td>
</tr>
<tr>
<td>(C + X)</td>
<td>18-15-12-2</td>
<td>0.82</td>
<td>0.128</td>
</tr>
<tr>
<td>(C + S)</td>
<td>18-10-8-2</td>
<td>0.84</td>
<td>0.122</td>
</tr>
<tr>
<td>(X + S)</td>
<td>18-15-12-2</td>
<td>0.96</td>
<td>0.006</td>
</tr>
<tr>
<td>(C + S + X)</td>
<td>18-15-12-3</td>
<td>0.80</td>
<td>0.129</td>
</tr>
</tbody>
</table>

Figure 3 | Schematic of North Marin WDS.
obtained are also compared with the results obtained from the multi-component reaction transport model (deterministic model).

With the introduction of a 10% random error the models have still been able to predict reasonably well the temporal variations in chlorine, biomass and substrate concentrations.

**CASE STUDY II—NORTH MARIN WDS**

**Description of the system**

The North Marin WDS serves a suburban population of 53,000 people. It uses two sources of water: Stafford Lake and the North Marin Aqueduct. The aqueduct is a year-round source. Stafford Lake is used only during the warm
summer months when precipitation is low and the demand for water is very high. The Stafford Lake source operates only during the peak demand period from 6.00 am to 10.00 pm. The network consists of 117 pipes, 92 nodes, 2 pump elements, 3 tanks and 2 supply sources (see Figure 3). More details are given in (Clark et al. 1995).

Preparation of inputs for the model

As the North Marin WDS is relatively large only 50% of the nodes are considered. Using a procedure similar to the one used in the previous case study 26 and 6 nodes are chosen for training and testing respectively. Since three cases are considered, with hourly data obtained at every node, the total training dataset consists of 1872 (i.e. $26 \times 24 \times 3$) datapoints. The testing dataset consists of 432 (i.e. $6 \times 24 \times 3$) datapoints.

The total number of inputs to the model is 30 (inputs for source chlorine, biomass, substrate and inflow for each source node, 1 input each for consumption and residence time and 10 time lags of source chlorine for each source). PCA is carried out for the entire dataset. 18 principal components are obtained from the 30 inputs. These principal components account for 99% of the variance of the entire input dataset.

Results and discussions

The number of hidden layers and the number of neurons in each of the hidden layers is determined by trial and error using the approach from the earlier case study. The learning rate is found to be in the range of 0.45 – 0.5. The momentum rate is found to be in the range of 0.75 – 0.8. The architecture, R and MAE values for all the models, using the testing dataset, are given in Table 2.

The results in Table 2 indicate that all the models (one output, two output and three output) are able to predict the temporal variations of the multi-components reasonably well. No significant differences exist between the three models in their predication capability for this WDS.

Error analysis

The robustness of the models is tested with noisy data using the procedure explained in the previous case. Plots of the temporal variations of the three components are shown in comparison with the error free dataset (Figure 4a–f). The results obtained are also compared with the results obtained from the multi-component reaction transport model (deterministic model).

With the introduction of a 10% random error the models have still been able to predict well the temporal variations in biomass concentration.

CONCLUSIONS

This study uses ANN models to predict temporal variations of chlorine, substrate and biomass concentrations in a WDS. The results obtained are encouraging. All the models (one output, two output and three output) are able to predict the temporal variations of the multi-components. With the introduction of a 10% random error, the models, in most cases, have still been able to predict reasonably well the temporal variations of the multi-components.

REFERENCES


