Water Science and Technology **Vol 48 No 10 pp 225–232** © IWA Publishing 2003

Predicting near-shore coliform bacteria concentrations using ANNS

B. Lin, S.M. Kashefipour* and R.A. Falconer

School of Engineering, Cardiff University, PO Box 925, Cardiff CF24 0YF, UK
(E-mail: *linbl@cardiff.ac.uk*; *falconerra@cardiff.ac.uk*)

* Irrigation Department, Shahid Chamran University, Ahwaz, Iran

Abstract Details are given of the application of Artificial Neural Networks (ANNs) to predicting the compliance of bathing waters along the coastline of the Firth of Clyde, situated in the south west of Scotland, UK. Water quality data collected at 7 locations during 1990–2000 were used to set up the neural networks. In this study faecal coliforms were used as a water quality indicator, i.e. output, and rainfall, river discharge, sunlight and tidal condition were used as input of these networks. In general, river discharge and tidal ranges were found to be the most important parameters that affect the coliform concentration levels. For compliance points close to the meteorological station, the influence of rainfall was found to be relatively significant to the concentration levels.

Keywords ANNs; bathing waters; coastal basin; faecal indicators; prediction

Introduction

Bathing water quality has increasingly become one of the main concerns of coastal environmental managers in the U.K. Pathogenic bacteria in bathing waters have been always the most important way of distributing dangerous and epidemic diseases. Individual pathogens are generally difficult and expensive to measure and therefore in water quality studies it is a common practice to measure and/or model the levels of related indicator organisms (Thomann and Muller, 1987). The faecal coliform (FC) bacterial indicator is widely used in assessing bathing water quality. The European Union (EU) has published several standards regarding guideline and mandatory concentration levels of pathogen indicators for bathing water (Council of the European Communities, 1976).

Currently numerical models based on solving solute transport and kinetic equations have been the main tool used to predict pathogen bacterial concentrations in coastal waters. The main advantage of using numerical models is that they are able to predict distributions of bacterial concentration at any time during the simulation. However, they need time series of hydrodynamic and water quality variables to drive and calibrate, and also require detailed bathymetric data, which are usually expensive and time consuming to obtain. Applying these models for predicting flow and water quality may also be restricted by the boundary conditions around the model domain (Falconer *et al.*, 2000).

In recent years artificial neural networks (ANNs) have been increasingly applied to a wide range of problems in various scientific fields (Widrow *et al.*, 1994). In this paper ANNs are applied to predict FC concentrations at several compliance sites along a coastline located at the south west of Scotland, UK, from Girvan in the south to Ardrossan in the north. The effect of the individual variables on the FC concentrations at those sites is discussed later in this study, and the failure of the bacterial concentrations in complying with the EU standards is also investigated.

Artificial neural networks

ANNs are a powerful tool in simulating dependent variables for a wide range of scientific

and engineering problems, in particular where non-linear relationships do exist between the variables (Zhang *et al*., 1998). Fundamental concepts of ANNs may be found in Anderson (1996), Landau and Taylor (1998) and Pham and Liu (1999).

In recent years ANNs have been increasingly used for modelling hydrological and water engineering problems, such as wastewater treatment works, river and coastal water quality, sediment transport, flood routing, waves and tides, ground waters, etc. (Hung and Foo, 2002). There are several types of neural network in the literature for simulating different variables and for different fields of studies, with the feed-forward networks (FFN, see Figure 1) being the simplest and the most popular procedures in ANNs (Pham and Liu, 1999). In the current study this type of neural network was used to model and estimate the FC concentration at the compliance points using the other measured or recorded variables.

The most important aspect of an ANN is its capacity to determine the interconnection weights; this is called learning or training. A supervised learning algorithm adjusts the strengths or weights of the inter-neuron connections according to the difference between the desired and actual network outputs, corresponding to a given input. After each iteration, the weights are adjusted using the back propagation (BP) algorithm, which is based upon the generalised delta rule proposed by Rumelhart and McClelland (see Pham and Liu, 1999). In this study a commercial network package, namely Qnet-2000, was used to model FC concentrations at the bathing water compliance points (Qnet-2000, 1999).

Faecal coliform modelling

The faecal coliform (FC) bacteria group is indicative of organisms from the intestinal tract of humans and other animals. The total coliform (TC) bacteria group is a large group of bacteria that has been isolated from both polluted and non-polluted soil samples as well as the faeces of humans and other warm blooded animals. This indicator was widely used in the past as a measure of health hazards and continues to be used in some areas. In recent years due to some difficulties in tests, such as the occurrence of non-faecal bacteria, the TC test is being gradually replaced by the FC test (Thomann and Mueller, 1987).

In mathematical modelling the fate of FC is generally expressed as first order decay. Recent research has shown that the fate of FC bacteria can be affected by several environmental factors such as light intensity, turbidity, temperature, salinity and pH level, etc. (Chapra, 1997). Several other factors such as rainfall and river flow are also significant in transporting FC from catchment areas and inland waters to estuarine and coastal waters.

The European Community (EC) guideline and imperative standard values for faecal coliform bacteria are 100 cfu/100 ml and 2,000 cfu/100 ml respectively (Council of the European Communities, 1976). If a sample has a faecal coliform concentration more than the imperative value, then it is said that the sample failed to comply with the standard values.

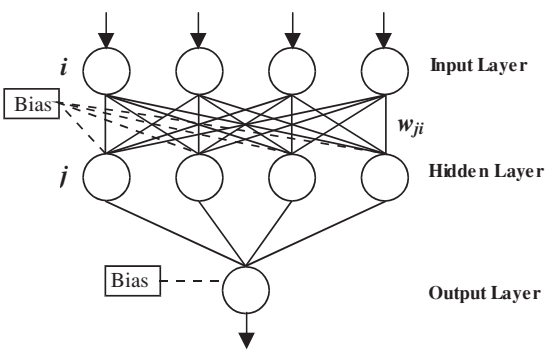

Figure 1 A typical feed-forward neural network

Model area and specifications

In this study the selected coastal zone comprised about 55km of the coastline, namely the Firth of Clyde located along the south west coast of Scotland, U.K., from Girvan in the south to Ardrossan in the north. Comprehensive data were available at seven compliance points, including: Girvan (Site 1), Turnberry (Site 2), Ayr South (Site 3), Prestwick (Site 4), Troon South (Site 5), Irvine (Site 6) and Saltcoats (Site 7), see Figure 2. Network simulations were carried out for all of these compliance points, but results at two sites will be presented and discussed herein.

Several hydrological parameters were recorded and/or measured on this coastline for 11 years (1990–2000). Faecal coliform concentrations were measured at the compliance points about 20 times each year. During a sampling exercise, the average daily flow rates from rivers discharging into the study area were also available for three days, including the sampling day and the previous two days. Rainfall at two meteorological sites including Girvan and Auchincruive (see Figure 2) and sunshine hours at Girvan were also collected for three days, including: the same day, one day and two days before sampling. The specifications of the rivers situated within the model domain and also the other variables are summarised in Tables 1 and 2.

The signs "0", "–1" and "–2" refer to the day of sampling, one day and two days before sampling day respectively. For example the sign "QIV-2" was applied for the Irvine River discharge measured two days before sampling day, and the sign "QGN-01" means the average river flow for the day of sampling and one day before for the Girvan River.

Methodology

In total 227 sets of measured and/or recorded data were available at each site (except Site 3), with 207 sets of data (1990–1999) being used to establish the ANN model and 20 sets of them (2000) being applied to verify the model. Only 153 data sets were available at Site 3 (Ayr South) for ANN training. Three main indexes were considered for the establishment and verification of the ANN model for each site. These indexes for training and verifying patterns were: (i) correlation coefficients, (ii) RMS (Root Mean Square) error, and (iii) number of the samples which failed the standard values.

In this study the networks for the considered sites were produced only with one hidden layer and the number of nodes made on this layer for each network was chosen according to the number and characteristics of the variables used for the network.

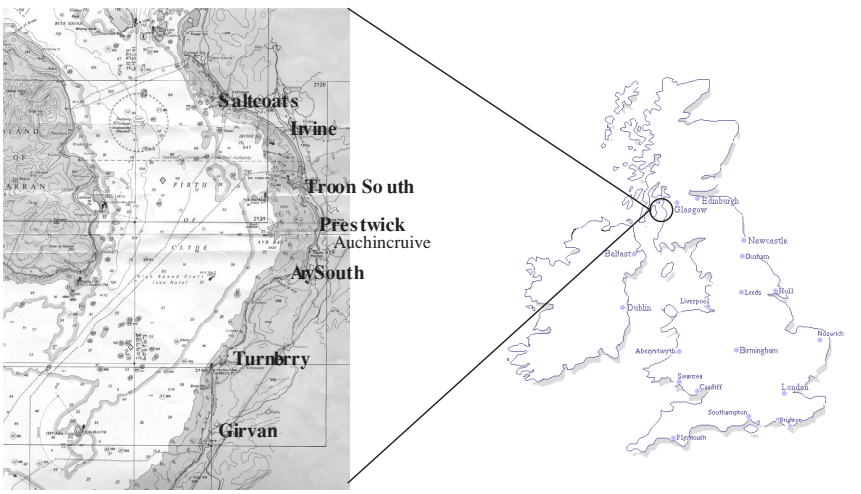

Figure 2 Firth of Clyde coastline in south west of Scotland, U.K.

Table 1 Variable specifications used for ANNs

Variable	Symbol	Explanations	Measured Unit
River Girvan	QGN	Close to Sites 1 and 2	m^3/s
River Irvine	QIV	Close to Sites 5, 6 and 7	m^3/s
River Doon	QDN	Close to Sites 2 and 3	m^3/s
River Lugar	QLG	Upstream of Ayr River (inland)	m^3/s
River Ayre	QAY	Close to Sites 2 and 3	m^3/s
River Lugton	QLT	Close to Sites 5, 6 and 7	m^3/s
River Garnock	QGK	Close to Sites 5, 6 and 7	m^3/s
Sum of rivers Girvan, Lugton and Garnock	QIVLTGK	Close to Sites 5,6 and 7	
Sunshine	SUN	Girvan	hr
Rainfall	GNRF	Girvan	mm
Rainfall	AURF	Auchincruive	mm
Tide Ht	THt	Height of tide at sampling day	m
High Tide	HTHt	High tide height	m
Relative time	RT	Sampling time relative to the nearest high tide[1]	–

[1] Normalised form $(T_1 - T_2)$/(tidal period)

Table 2 Variables used for the ANN models

Site	Variables	NOV[1]	NOHN[2]
1	QGN, QGN-1, HTHt, RT, SUN, SUN-1, GNRF, GNRF-1	8	8
3	THt, RT, HTHt, QDN-12, QAY-12, SUN-012, AURF-01	7	7

[1] Number of variables [2] Number of hidden nodes

For each site the variables were selected according to the possible relationships of the coliform bacterial concentrations to those variables. For example sunshine hours, rainfall and tide measurements were used as the variables for the first trial of all ANN networks for all sites involved. For each site only the discharges of a river that was close to the site were used in the ANN network, since it was thought that the rivers far from the compliance point would have little effect on the coliform bacterial concentrations at that site. For the first trial of each ANN network the considered variables for all three days were used and then they were subsequently changed to a proper number of variables due to the three aforementioned indexes. For each site several runs were carried out and by trial and error procedure, i.e. by comparing the simulation results with measured values for both training and verifying patterns, the most suitable ANN network was produced with the proper variables.

For each run a few random values of weights were first inserted and then subsequently adjusted automatically by the network by several iterations due to the pre-regulated RMS error. The number of iterations was optional and up to the users and for this ANN network a number of 100,000 was applied for each run.

It should be noted that the measured faecal coliform concentrations ranged from less than 10 cfu/100 ml to more than 50,000 cfu/100 ml. Therefore, for output targets, i.e. measured bacterial concentrations, the logarithms of the measured values were used instead of the absolute values to avoid the negative amounts for the model outputs and also to valorise the small changes in the bacterial concentrations, when the absolute concentrations are low. Since no significant differences existed between the minimum and maximum measured data (i.e. covering one or two logarithmic scales), the other variables were applied in the ANN models with their original measured values.

Results and discussion

Site 1: Girvan

QGN, SUN, GNRF for the sampling day, one day and two days before sampling, THt, HTHt and RT were initially selected as the input variables. In total 12 variables were entered for the first trial. A number of runs were carried out and the variables were subsequently changed to give the best results and the final selected variables were QGN, QGN-1, HTHt, RT, SUN, SUN-1, GNRF and GNRF-1. Also for the best results the number of nodes for the hidden layer was found to be 8, which was the same as the number of variables.

A summary of the variables used for the ANN model at this site is listed in Table 2. It can be noted that at Site 1 the recorded parameters two days before the sampling day were not significant to the FC concentration on the sampling day. This is thought to be firstly due to the short distance of the discharge measuring point relative to the compliance point (Site 1) and secondly due to the low T_{90} value (i.e. high decay rate) of faecal coliform bacteria in this area (Kashefipour *et al.*, 2001).

Table 3 is a summary of the simulation results, including: correlation coefficients between training and verifying patterns, the RMS error, and the number of predicted and measured failures of the samples with the standard values. As can be seen from Table 3 the correlation coefficients for both the training and verifying patterns were relatively good. From this table it also can be concluded that for training and verifying patterns the model was able to predict about 65.6% of failure samples, with a percentage of about 56.3% happening on the same day.

Figure 3 compares the predicted and measured faecal coliform concentrations (logarithm values) at Site 1 for both training and verifying results. Figure 4 shows the contribution of all input variables in the ANN model. The maximum contributions with more than 18% were related to the discharge of Girvan River for the sampling day (QGN) and rainfall for one day before sampling day (GNRF-1). Correlation between rainfall and river discharge is clear. However, it can be seen that due to the wideness of the catchment area the

Table 3 Statistical analysis of the results obtained from ANN models

	Correlation coefficient			Number of failed samples					
	Training pattern	Verifying pattern	RMS error	Training pattern			Verifying pattern		
				Measured	Model	SM[1]	Measured	Model	SM
1	0.748	0.503	0.089	31	20	17	1	1	1
3	0.815	0.570	0.089	34	31	24	1	1	0

[1] Happened on same day

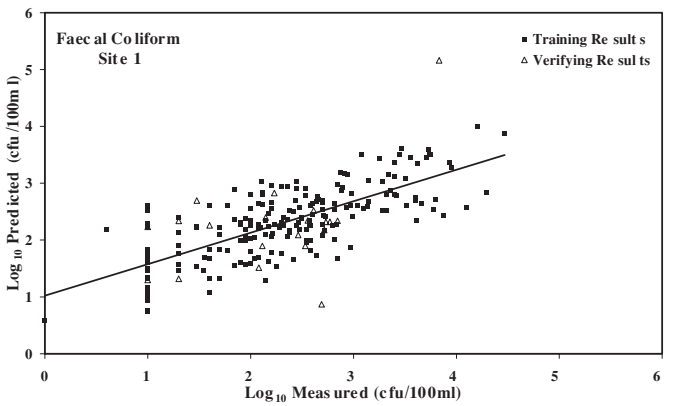

Figure 3 Comparison of predicted and measured faecal coliform concentrations for Site 1

effect of rainfall has lasted for two days. After these two variables the height of high tide (HTHt) was found to be significant. Although the percentage contribution of sunshine was small, applying this variable increased the correlation coefficients for training and verifying of the ANN model.

Site 3: Ayr South

Input variables used for producing the ANN model at this site are appended to Table 2. The statistical results relating the training and verifying patterns are summarised and added to Table 3. This table shows an excellent agreement between predicted and measured faecal coliform concentration at this compliance point. The model was able to predict more than 90% of the failed samples with a percentage of about 71% happening on the same day.

Comparison of the predicted faecal coliform concentration with the corresponding measured values and the percentage contribution of the variables used for the ANN model are shown in Figures 5 and 6 respectively. Figure 6 shows again the river flow discharges were the most important variables, with contributions of more than 18% for two rivers close to this compliance point. The best correlation coefficients were obtained when the average river discharges for two days, including one day and two days before sampling, were used as the input variables for the model. A rainfall station, i.e. Auchincruive, was close to this compliance point and as a result the contribution of this variable was high (see Figure 6). This figure also shows that the tidal conditions are relatively important to bacterial concentrations.

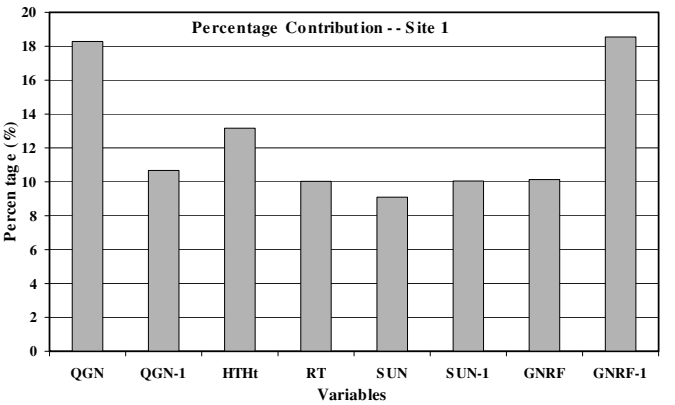

Figure 4 Percentage contribution of variables in ANN model for Site 1

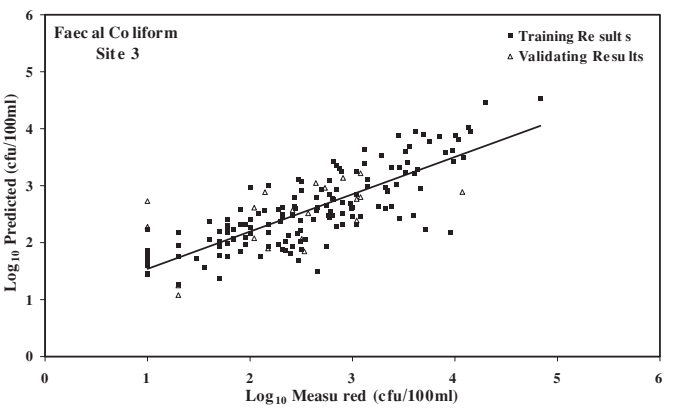

Figure 5 Comparison of predicted and measured faecal coliform concentrations for Site 3

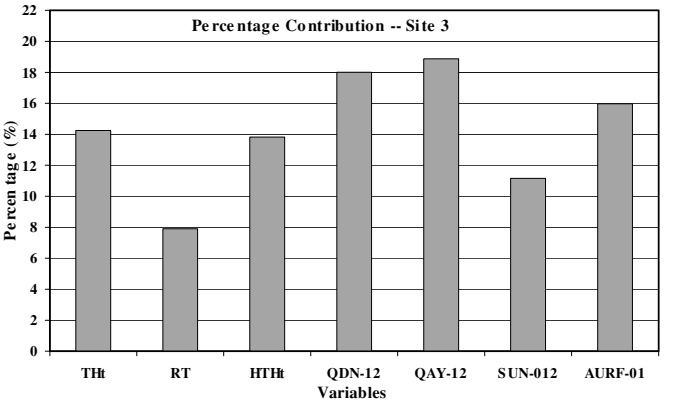

Figure 6 Percentage contribution of variables in ANN model for Site 3

Conclusions

In this study artificial neural networks were used to predict the faecal coliform concentration at the compliance points along a coastline situated in the south west of Scotland. Variables used in the neural networks include river discharges, sunshine hours, rainfall and tidal conditions, such as time of sampling relative to the high tide, water elevation at the sampling time and the height of high tide. Data related to 1990–1999 were used for the learning process and establishing the ANN networks and the data of 2000 were used to verify the model. The main conclusions can be summarised as follows:

- For all of the ANN networks, river discharge is found to be the most significant input variable to the bacterial concentration at the compliance points.
- For the compliance points located far away from the river outlets the number of samples failing to comply with the standards value was significantly low.
- For all of the ANN networks the height of the high tide was found to be relatively significant.
- Since the tidal range in this area is moderate and as a result the current speed was generally low, the effect of wind speed and direction may also be important. However, these data were not available at the time of this study.

Acknowledgement

The study reported herein was funded through a number of grants including those from: the Engineering and Physical Sciences Research Council WITE programme (grant GR/M99774) and project GR/N03662. The authors would also like to acknowledge the support and assistance given by Professor D. Kay, Dr N. Humphrey, Dr M. Wyer and Dr C. Stapleton.

References

Anderson, J.A. (1996). *An introduction to neural network*. Massachusetts Institute of Technology, 650 pp.

Chapra, S.C. (1997). *Surface water quality modelling*, McGraw-Hill Book Companies, Inc., USA, 844 pp.

Council of the European Communities, 1976, Council Directive of 8th December 1975 concerning the quality of bathing water (76/160/EEC), Official Journal of the European Communities No. L31, pp. 1–7.

Falconer, R.A., Lin, B., Harris, E. and Kashefipour, S.M. (2000). DIVAST Model: Reference manual, Cardiff University, Environmental Water Management Research Centre, 21 pp.

Hung, W. and Foo, S. (2002). Neural network modelling of salinity variation in Apalachicola River. Research Note, *Water Research*, **36**, 356–362.

Kashefipour, S.M., Lin, B. and Falconer, R.A. (2001). Modelling the fate of faecal indicator in a coastal basin. *Journal of Environmental Engineering*, ASCE (submitted).

Landau, L.J. and Taylor, J.G. (eds) (1998). *Concepts for Neural Networks*: *A Survey*, Springer-Verlag London Limited, 307 pp.

Pham, D.T. and Liu, X. (1999). *Neural Networks for Identification, Prediction and Control*. 4th printing, Springer-Verlag London Limited, 238 pp.

Qnet 2000 (1999). Qnet 2000 Neural Network Modelling for Windows 95/98/NT, QnetToll User's Guide and Datapro User's Guide, Vesta Services, Inc. USA.

Thomann, R.V. and Muller, J.A. (1987). *Principles of Surface Water Quality Modelling*, Harper Collins Publishers Inc., New York, 644 pp.

Widrow, B., Rumelhart, D.E. and Lehr, M.A. (1994). Neural network: applications in industry, business and science. *Communications of ACM*, **37**, 93–105.

Zhang, G., Patuwo, B.E. and Hu, M.Y. (1998). Forecasting with artificial networks: The state of art. *International Journal of Forecasting*, **14**, 35–62.