Water Quality and Sediment Behaviour of the Future: Predictions for the 21st Century (Proceedings of Symposium HS2005 at IUGG2007, Perugia, July 2007). IAHS Publ. 314, 2007.

Kohonen self-organising map (KSOM) extracted features for enhancing MLP-ANN prediction models of BOD_5

RABEE RUSTUM[1], ADEBAYO ADELOYE[1] & AURORE SIMALA[2]

1 *School of the Built Environment, Heriot-Watt University, Edinburgh EH14 4AS, UK*
a.j.adeloye@hw.ac.uk

2 *15 bis, allée Marie, F-93360 Neuilly-Plaisance, France*

Abstract This paper presents the results of developing a model to predict the concentrations of biological oxygen demand (BOD_5), in the effluent of the primary clarifier of an activated sludge wastewater treatment plant, using other easily measurable water quality parameters. The model is based on the Kohonen self-organising map (KSOM) and multi-layered perceptron artificial neural networks (MLP-ANN). The KSOM was used to extract the features of the measured data and to deal with the effects of noise and missing values. The best map units of the measurement vectors over the KSOM were used as inputs to the MLP-ANN to reduce the effects of noise and uncertainty in the measurement data, and to replace the missing elements in these measurements. The results of the KSOM-ANN modelling strategy were found to be better than those obtained by the MLP-ANN trained using the raw measurement data.

Key words wastewater treatment plant; primary clarifier modelling; neural networks; Kohonen self-organising map

INTRODUCTION

With tighter regulations on river water quality, it is important to limit point source pollution by improving the performance of wastewater treatment plants. Controlling treatment plants through modelling is technically the most feasible and maybe least costly way of achieving a sustainable improvement in performance. This is because modelling the wastewater treatment units can help the operator to test some corrective actions without the need to apply them to the real process and in this way identify the corrective actions that give better performance.

Modelling and controlling the primary clarifier (PC) are very important since its performance directly influences the subsequent biological and sludge treatment units and hence the overall performance of the treatment plant (Geraney *et al*., 2001). In addition, the biological load used for sizing the secondary treatment stage reactor and for estimating various process control parameters, e.g. the F/M ratio, is derived from the PC effluent BOD_5.

However, modelling the PC has many problems including the variability of influent characteristics, variability of particle sizes and corresponding settling velocities, scouring and re-suspension of settled particles, interactions between the different micro-organism populations, and the variability of desludging operations (Manfred *et al.*, 2002). All these problems give the PC its nonlinear characteristics and time-

varying parameters. Thus, most approaches to modelling the PC using mechanistic paradigms have relied on numerous simplifying assumptions in order to make the problem tractable (Lessard & Beck, 1998).

For this work, an alternative approach involving neural computing has been applied to model the PC. Artificial neural networks (ANNs) can be used to model any complex, nonlinear and dynamic systems without the need to specify the functional form of the governing relationship *a priori* (Pu & Hung, 1995). However, basic multi-layered perceptron (MLP)-ANNs are affected by the quality of the data such as noise and missing values, which can make effective training difficult. To solve this problem, the Kohonen self-organising map (KSOM), an unsupervised ANN, was further used to extract the features from the noisy data, which are then used to drive the MLP-ANN, as illustrated in Fig. 1. The results of the two approaches, i.e. MLP-ANN on noisy and on KSOM-features data, are presented and compared.

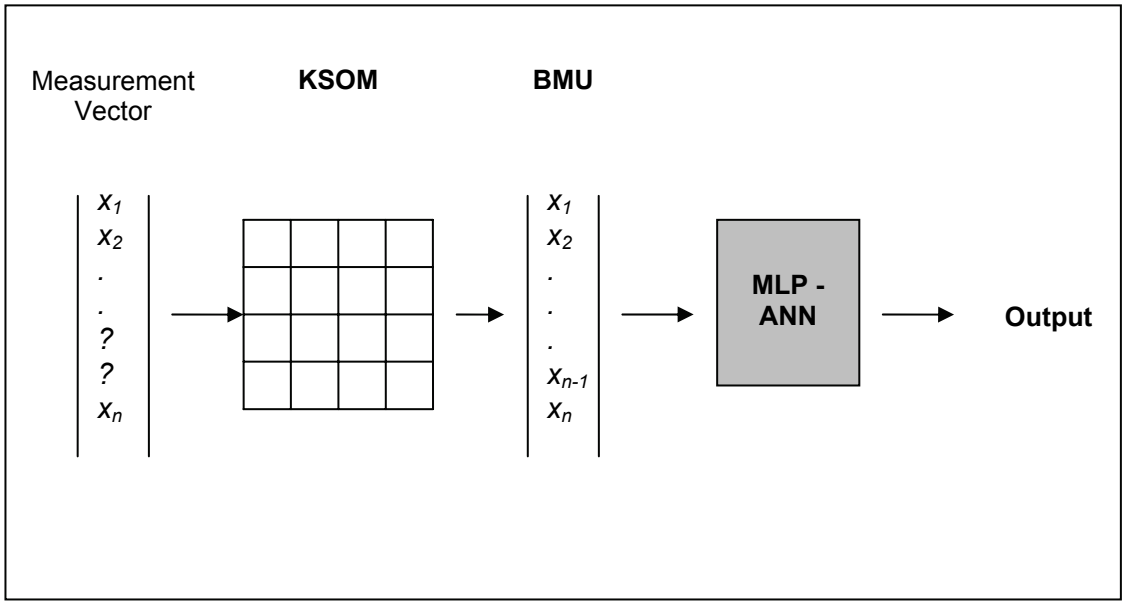

Fig. 1 Diagrammatic representation of the integrated KSOM-ANN modelling strategy.

METHODS

Basics of artificial neural networks

The theory and mathematical basis of artificial neural networks (ANNs) have been described excellently elsewhere (Bishop, 1995). ANNs consist of a set of artificial neurons which are called nodes, and they have connections between them, called weights. Optimal values for these weights are obtained by training the network. The most commonly used form of ANNs is the multi-layer perceptron (MLP). In general, a three-layered MLP can approximate any function with sufficient accuracy (Hagan *et al.*, 1996). ANNs offer a very powerful and very general framework for representing nonlinear mapping from several input variables to several output variables (Bishop, 1995; Demuth & Beale, 1998).

Kohonen self-organising map and features extraction

The KSOM (also called a feature map or Kohonen map) is an unsupervised ANN algorithm (Kohonen *et al*., 1996). It is usually presented as a dimensional grid or map whose units (nodes or neurons) become tuned to different input data patterns. The principal goal of the KSOM is to transform an incoming signal pattern of arbitrary dimension into a two-dimensional discrete map. This mapping roughly preserves the most important topological and metric relationship of the original data elements, implying that not much information is lost during the mapping.

To train the KSOM, the multi-dimensional input data are first standardised by deducting the mean and then dividing the result by the standard deviation. Then a standardised input vector is chosen at random and applied to each of the individual neurons, which are initially seeded with weights or code vectors. The elements of the code vectors also have a mean of zero and variance of unity. Comparison between the input vector and the code vectors is then made to identify the code vector most similar to the applied input vector. The identification uses the Euclidian distance, which is defined as:

$$D_i = \sqrt{\sum_{j=1}^{n} \left(x_j - w_{ij}\right)^2} \quad i = 1,2,...,M \tag{1}$$

where D_i is the Euclidian distance between the input vector and the weight vector i; x_j is the j-th element of the current input vector; w_{ij} is the j-th element of the weight vector I; M is the number of neurons in the KSOM; and n is the dimensionality of both the input and code vectors.

The neuron whose vector most closely matches the input data vector, i.e. for which the D_i is a minimum, is chosen as a winning node or the best matching unit (BMU). The vector weights of this winning neuron and those of its adjacent neurons are then adjusted to match the input vector using an appropriate algorithm, thus bringing the two vectors further into agreement as illustrated in Fig. 2. In this manner, each neuron

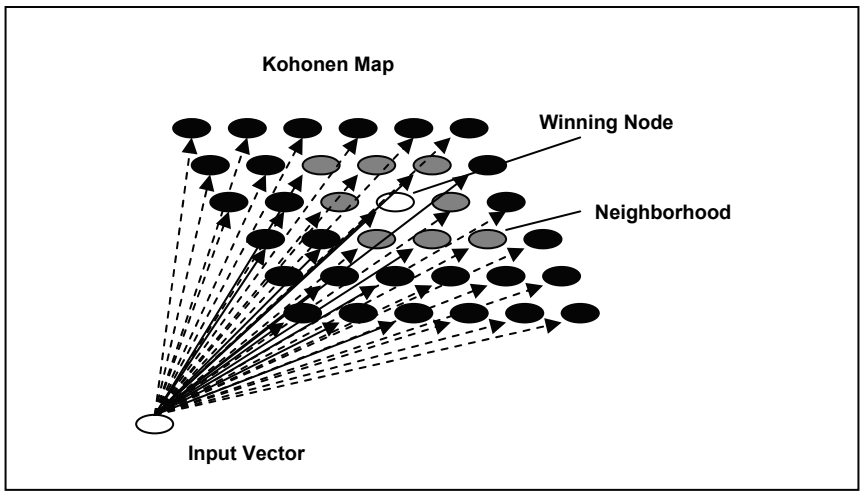

Fig. 2 Diagrammatic representation of the winning node and its neighbourhood in the KSOM.

in the map internally develops the ability to recognize input vectors similar to itself. This characteristic is referred to as self-organising, because no external information is supplied to lead to a classification (Penn, 2005). At the end of the training, the BMUs can be considered as the features of the measurement. Unlike the raw data, the extracted features do not have any outliers or missing values. More details about the use of the KSOM for extracting features are provided by Rustum & Adeloye (2006). Garcia & Gonzalez (2004) provide further details regarding the step-by-step procedure of the KSOM algorithm, including particular forms of the weight updating and neighbourhood functions.

CASE STUDY

Data

Daily record sheets describing the operation of the Seafield activated sludge treatment plant in Edinburgh (Scotland, UK) for a period of approximately three years (1 May 2002 to 31 March 2005) with a total of 1066 data vectors were obtained from the plant operators, Thames Water. A summary of the variables measured at the inlet to the works is shown in Table 1. An important feature of the data record is the large number of missing values.

The Seafield treatment plant relies on an activated sludge secondary treatment system and comprises six circular sedimentation tanks, six rectangular non-nitrifying aeration lanes and eight circular final settlement tanks. The main treatment is preceded by six screens (spacing: 6 mm) and four Detritor grit removal units. The works is located in the eastern part of Edinburgh and discharges its effluent directly to the sea. The maximum flow is 674 000 m^3/d, which equates to a BOD_5 population equivalent of 1 136 825. All incoming flow is treated by primary clarification and secondary biological aeration. However, only a fraction of the throughput is given tertiary treatment by ultra-violet light disinfection prior to discharge. The works currently performs to high standards and has not been violating its discharge standard of BOD_5 concentration of 25 mg/L maximum, according to an internal report produced by Thames Water.

Table 1 Water quality characteristics of the inflow to the Seafield sewage treatment works.

Symbol	Description	Unit	Min.	Max.	Mean	No. missing values
Q	Flow rate to the treatment plant	m^3/d	100 000	674 000	300 261	2
COD	Chemical oxygen demand	mg/L	74	880	350.58	143
SS	Suspended solids	mg/L	15	580	164.27	41
NH_4	Ammonia-nitrogen	mg/L	0.50	35.84	13.83	144
pH	pH	-	7.09	9.20	6.10	162
T	Temperature	°C	9	19	14	187
BOD	Five days at 20°C biological oxygen demand	mg/L	8	316	117	45

Numerical analysis and modelling

The SOM toolbox for Matlab 5 (Vesanto *et al*., 2000) was used for the KSOM training and to extract the features. The training of the MLP-ANN used the neural networks toolbox in Matlab. The neural networks have six input variables: influent flow, influent biological oxygen demand (BOD), influent chemical oxygen demand (COD), influent suspended solids (SS), influent ammonia, and the sludge blanket depth in the PC. These variables were chosen because of their relatively high correlation coefficients with the PC effluent BOD_5 (Rustum & Adeloye, 2006). To overcome the over-fitting problem, the early-stop rule was used which necessitated dividing the data into three subsets for training (500 data points), validation (200 data points) and testing (366 data points). The validation data set was used to stop the training when the errors in this set begin to increase during the training, following a sustained period in which the error fell. The testing set was used to assess the ability of the ANN to generalise. The models were evaluated using three criteria namely, correlation coefficient (R), mean square error (MSE) and average absolute error (AAE).

RESULTS AND DISCUSSION

To reach the suitable network architecture for the MLP-ANN, simulations were run for various assumed numbers of hidden neurons. Table 2 summarizes the errors during training, validation and testing for the MLP-ANN modelling of both data sets, i.e. raw data and features of the raw data. From Table 2, it is clear that the 10-node architecture can be taken as a compromise best structure since no significant improvement in all the three performance criteria occurs when the number of neurons is increased beyond 10. However, there were sustained, distinct improvements in the model performance until the 10-node structure was attained.

It is also evident in Table 2 that for each specific number of neurons in the hidden layer, the performance of the model is better using the features of the raw data than using the raw data itself. For example, the correlation coefficient between the measured and the predicted BOD during testing for the 10-neuron model was 96% for the features, compared to just 71% for raw data. The relative superiority of the features-derived models is also evident when both the MSE and AAE are considered. This is because the features have eliminated the noise in the raw data set, which affected the performance of the basic ANN.

Because the ANN model with 10 neurons in the hidden layer using the features method has the best performance, further analysis was only done with this model. Figure 3 shows the scatter plot of the measured and predicted BOD during training, validation and testing. Most of the data points are around the best linear fit line and also around the predicted equals observed line. The time series plots of the residuals are also shown in Fig. 3, from which it is clear that the residuals are random.

CONCLUSION

The current work used a new methodology based on a hybrid supervised-unsupervised artificial neural network to improve the performance of the basic backpropagation

Table 2 Results for the MLP-ANN models for predicting the PC effluent BOD_5.

Number of hidden neurons	Data type (Training/validation /testing)	Correlation %		MSE		AAE	
		Features	Measured	Features	Measured	Features	Measured
3	Training	93	71	155.0	561.87	9.88	18.95
	Validation	92	70	78.14	248.24	6.59	12.39
	Testing	94	71	73.11	270.55	6.81	13.15
5	Training	90	64	529.4	802.33	18.5	22.20
	Validation	91	68	133.3	248.12	9.30	12.42
	Testing	98	61	135.1	341.22	9.12	15.00
7	Training	96	77	145.2	476.17	9.42	17.45
	Validation	94	72	66.06	237.05	5.99	12.14
	Testing	96	70	89.57	273.22	7.66	13.27
10	Training	97	78	64.93	443.58	6.23	17.00
	Validation	95	71	44.23	262.42	5.08	12.64
	Testing	96	71	42.86	267.33	5.10	13.02
12	Training	97	79	68.49	417.76	6.42	16.22
	Validation	95	71	44.45	258.39	5.12	12.62
	Testing	96	71	49.04	261.77	5.55	12.89
14	Training	94	79	135.2	409.40	9.12	16.23
	Validation	91	72	74.76	275.03	6.75	12.69
	Testing	91	72	99.12	259.82	8.02	12.87
16	Training	92	80	659.5	406.66	20.5	16.23
	Validation	86	71	204.4	258.27	11.4	12.14
	Testing	84	71	183.0	264.11	10.8	12.82
18	Training	97	81	66.86	384.84	6.37	15.69
	Validation	95	73	42.47	249.33	4.94	11.85
	Testing	97	71	41.48	253.94	5.01	12.54
20	Training	94	80	138.4	394.62	9.47	15.70
	Validation	92	73	75.74	235.47	6.79	11.60
	Testing	95	73	67.71	251.23	6.40	12.73

neural network method in modelling the primary clarifier of a wastewater treatment plant. The method was applied to data taken from the Seafield wastewater treatment plant in Edinburgh, UK, during a period of about three years. Input variables were selected based on their correlation with the effluent BOD, which was the target prediction variable. Several ANN models with different numbers of neurons in the hidden layers were developed. For each model, two types of data were used, the first one is the raw data set and the second one is the extracted features of the raw data using the Kohonen self-organising map. The results showed that the models using the features were better than those using the raw data.

The findings prove the ability of KSOM to improve the performance of modelling using basic back-propagation neural networks, particularly when the available data are noisy, a common problem with the process data of wastewater treatment plants. Furthermore, the KSOM can readily deal with missing values in one or more of the input variables without significantly negative impacts on the accuracy of the model. The methodology is therefore applicable to other water and environmental engineering problems.

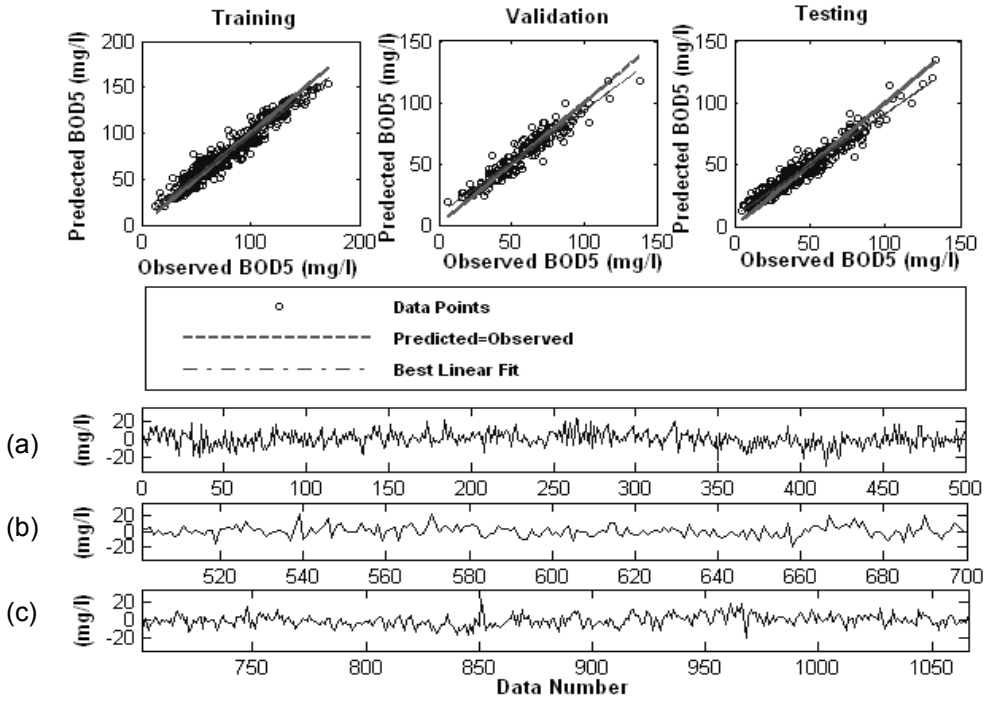

Fig. 3 Performance of the final ANN model in predicting the BOD_5. Plots (a), (b) and (c) are the residuals.

REFERENCES

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Demuth, H. & Beale, M. (1998) *Neural Network Toolbox, For Use with MATLAB*. The Math Works, Inc., User's Guide, Version 4.Natick, Massachusetts, USA

Geraney, K., Vanrolleghem, P. A. & Lessard, P. (2001) Modelling of a reactive primary clarifier. *Water Sci. Technol*. **43**(7), 73–81.

Garcia, H. & Gonzalez, L. (2004) Self-organizing map and clustering for wastewater treatment monitoring. *Eng. Appl. Artific. Intellig*. **17**(3), 215–225.

Hagan, M. T., Demuth, H. B. & Beale, M. (1996) *Neural Network Design*. PWS Publishing Company, Boston, Massachusetts, USA.

Kohonen, T., Oja, E., Simula, O., Visa, A. & Kangas, J. (1996) Engineering applications of the Self Organising Map. *Proc. IEEE* **84**(10), 1358–1384.

Lessard, P. & Beck, M. B. (1988) Dynamic Modelling of Primary Sedimentation. *J. Environ. Engng, ASCE* **114**(4), 753–769.

Manfred, S., Butler, D. & Beck, M. B. (2002) *Modelling, Simulation and Control of Urban Wastewater Systems*. Springer-Verlag, London, UK.

Penn, B. S. (2005) Using self-organizing maps to visualize high dimensional data. *Comp. Geosci.* **31**(5), 531–544.

Pu, H.-C. & Hung, Y.-T. (1995) Artificial neural networks for predicting activated sludge wastewater treatment plant performance. *Int. J. Environ. Studies* **48**, 97–116.

Rustum, R. & Adeloye, A. J. (2006) Features extraction from primary clarifier data using unsupervised neural networks (Kohonen Self Organising Map). In: *Seventh International Conference on Hydroinformatics* (HIC 2006), Nice, France.

Vesanto, J., Himberge, J., Alhoniemi, E. & Parhankangas, J. (2000) Self-organizing map (SOM) Toolbox for Matlab 5. Report no. A57, Laboratory of Computer and Information Science, Helsinki University of Technology, Helsinki, Finland.