Int. J. Environ. Res., 2(2): 183-188, Spring 2008
ISSN: 1735-6865

Prediction of Water Quality Indices by Regression Analysis and Artificial Neural Networks

Rene, E R.[1*] and Saidutta, M. B. [2]

[1]Department of Chemical Engineering, University of La Coruna, E-15071, Spain
[2]Department of Chemical Engineering, National Institute of Technology Karnataka, Surathkal – 575025, Mangalore, Karnataka, India

Received 12 Oct. 2007; Revised 20 Sep. 2007; Accepted 10 Jan. 2008

ABSTRACT:The quality of wastewater generated in any process industry is generally indicated by performance indices namely BOD, COD and TOC, expressed in mg/L. The use of TOC as an analytical parameter has become more common in recent years especially for the treatment of industrial wastewater. In this study, several empirical relationships were established between BOD and COD with TOC using regression analysis, so that TOC can be used to estimate the accompanying BOD or COD. A new, the use of Artificial Neural Networks has been explored in this study to predict the concentrations of BOD and COD, well in advance using some easily measurable water quality indices. The total data points obtained from a refinery wastewater (143) were divided into a training set consisting of 103 data points, while the remaining 40 were used as the test data. A total of 12 different models (A1-A12) were tested using different combinations of network architecture. These models were evaluated using the % Average Relative Error values of the test set. It was observed that three models gave accurate and reliable results, indicating the versatility of the developed models.

Key words: Neural Networks, Regression Analysis, BOD, COD, Prediction, Average Relative Error

INTRODUCTION

The rapid growth and proliferation of industrial sector have contributed to severe deleterious effect on the environment. Good engineering practice dictates that waste materials can be discharged into receiving water in such a way that nature's ability to assimilate these wastes is utilized without any deleterious effects on the water quality. It is however necessary to analyse the industrial wastewater to determine its suitability for reuse, the degree of treatment required prior to its disposal or to device suitable measures for the recovery of useful products. It is of great importance in water quality control that the amount of organic matter present in the system be known and that the quantity of oxygen required for its stabilisation be determined. Over the past few years, a number of different tests have been developed to determine the organic content of wastewater (Sawyer, *et al.*, 1994; Metcalf and Eddy, 1995). In general, these tests may be divided into those used to measure gross concentrations of organic matter greater than 1 mg/L and those used to measure trace concentrations in the range of 10^{-6} to 10^{-3} mg/L. Laboratory methods commonly used today to measure the gross amount of organic matter (greater than 1 mg/L) in wastewater include, Biochemical Oxygen Demand (BOD_5), Chemical Oxygen Demand (COD) and Total Organic Carbon (TOC). Till now correlation between parameters like BOD, COD, TOC are few in the literature. On the other hand, the other indices like, Total Suspended Solids (TSS), Total Dissolved Solids (TDS), Phenol concentration, Ammoniacal Nitrogen (AMN) and Kjeldahl's Nitrogen (KJN) have not been related to BOD and COD. Probably this goes back to problems in interpretation of correlation between these parameters because they depend on the composition of the waste developed in a particular industry, their on-linear

*Corresponding author: Email-eldonrene@yahoo.com

relationship which could vary from process to process and no generalisation method was followed till now. Due to these reasons an information processing system is required which is not based on physical or chemical modelling but gets its knowledge about the process from information supplied through current data and the system dependent measuring data. At this articulate moment, there is a strong need to establish suitable empirical relationships between these variables, so that monitoring and prediction of water quality indices in industrial facilities would be simple and reliable. This paper reports the results of regression analysis (relations between COD, BOD and TOC) and describes a neural network approach to predict the BOD and COD concentration of a refinery effluent from easily measurable water quality parameters.

Neural Networks are able to learn non-linear static or dynamic behaviour among variables based on the given set of data. Since the knowledge of internal procedure is not necessary, the modelling can take place with minimum previous knowledge about the process through proper training of the network (Hawkin, 1994). The propagation of the data through the network starts with the presentation of an input stimulus at the input layer. The data then flows through and are operated on by the network until an output stimulus is produced at the output layer (Maier and Dandy, 2001). For each neuron or processing element (PE), each of its inputs is multiplied by its corresponding weight, i.e., the weight assigned to each input connection path to that neuron and the total sum of these products plus a constant term yields the neuron input I_j, which can be represented by,

$$Ij = \Sigma^{N}_{i=1} Wij\, Xi + \theta j \qquad (1)$$

Where, N- total number of neurons in the preceding layer, X_i-the neuron input received from the i^{th} neuron in the preceding layer, W_{ij}-the connection weight assigned to the path linking the neuron to that i^{th} neuron and θ_j- the neuron threshold value.

The neuron threshold value, θ_j provides the means of adding a constant value to the sum term which enhance the flexibility of the network by allowing an additional degree of freedom when trying to minimize the error between observed and predicted values (Pai, *et al.*, 2007). The inclusion of the term θ_j in the equation is analogous to that of considering an intercept in the context of linear regression (Hawkin, 1994; Bose, 1998). The neuron input I_j is transformed to the neuron output Y_j by the application of the most popular transfer function used in neural network studies for the neurons in the hidden and output layers, the logistic function (also called as the Sigmoid Function) which takes the form (Jacek, 1992),

$$Yj = f(Ij) = (1/1 + e^{-Ij}). \qquad (2)$$

MATERIALS & METHODS

The various wastewater parameters such as TSS, BOD, COD, TOC, and phenol concentration, AMN, KJN and TDS were obtained from the quality control laboratory of a refinery located in Mangalore, India. Water samples collected from the effluent treatment plant after tertiary treatment were analyzed for the above mentioned parameters, which were later divided into training set (103) and test set (40). The ranges of various values of different parameters used for training and testing are shown in Table 1 and 2 respectively.

Table 1. Range of water quality parameters used for training (mg/L)

No.	Parameters	Range	Mean
1	BOD	2 – 34	13.52
2	COD	12 – 160	61.64
3	TOC	3.1 – 18.5	8.21
4	TSS	4 – 71	18.602
5	TDS	343 – 1851	858.62
6	AMN	1.4 – 92	19.04
7	KJN	1.8 – 93.4	20.8335
8	Phenol	0.08 – 0.8	0.286

Regression analysis for the given data set was carried out using Microsoft Excel and their performance was indicated by the correlation coefficient, R^2 values and % Average Relative Error (% ARE) values, while ANN based models were handled using the software NN MODEL and their performance were evaluated solely using % ARE.The percentage Average Relative Error (%ARE) was estimated from this relation (van der Walt *et al.*, 1993),

$$\%ARE = \frac{1}{N}\sum \frac{\left|A_{Expt} - A_{Pred}\right|}{A_{Expt}} * 100 \qquad (3)$$

Table 2. Range of water quality parameters used for testing(mg/L)

No.	Parameter	Range	Mean	Std. Deviation
1	TOC	4.04 – 18.5	9.67	3.864
2	TSS	6 - 41	18.125	7.408
3	TDS	480 - 1720	973.725	300.49
4	Phenol	0.1 – 0.8	0.3063	0.159
5	AMN	9.5 - 94	31.9	23.307
6	KJN	10.3 – 96.8	34.48	23.68
7	BOD	6.1 - 34	15.61	6.92
8	COD	38 - 114	72.075	24.79

RESULTS & DISCUSSION

Regression analysis was carried out to relate TOC with BOD and TOC with COD using the training data set. For the BOD: TOC relationship the R^2 varied from 0.5426 to 0.6082, while for the COD: TOC relations the value of R^2 ranged from 0.3136 to 0.4033, indicating a poor correlation between these parameters. More ever, in practice situations may arise when the wastewater contains no organic carbon i.e., when TOC=0 mg/L, and contains only inorganic pollutants. In this case there will be no measurable BOD in the wastewater because BOD is a measure of only the biodegradable organics present in wastewater. The relation between BOD: TOC, COD: TOC and COD: BOD with their respective R^2 values and the type of regression analysis carried out is shown in Table 3.

The equations obtained were then tested with 40 test data to evaluate the predictability of the developed empirical relations. The % ARE value was found for all the regression equations using these test data and the best relation was selected. It was observed that the relation $Y=1.722\ X^{0.9643}$ (% ARE=12.873) can be used to estimate BOD, while the relation Y=5.2537 X + 18.691 can be used to find COD, which gave a % ARE of 11.398.

The input to the network was selected keeping in mind the parameters which affects the output. In the present work TOC, phenol, TSS, TDS, AMN and KJN were used as the inputs in different combinations to predict the BOD and COD values. The standard back propagation algorithm (BEP) was used by the software NN Model. A total of 12 models (A1 to A12) were developed out which 5 models were trained to predict BOD, 5 models for predicting COD and 2 models to predict both BOD and COD (Table 4).

Table 3. Correlation of BOD and COD with TOC using regression analysis

Parameters in Y, X	Type	Relation	R^2
BOD,TOC	Linear through origin	$Y=1.6236\ X$	0.5507
	Linear , with intercept	$Y=1.4228\ X + 1.8954$	0.5636
	Logarithmic	$Y=11.294\ \ln(X) –9.3242$	0.5693
	Power	$Y=1.722\ X^{0..9643}$	0.6082
	Exponential	$Y=26.888\ e^{0.1153X}$	0.5426
	Polynomial, 2nd order	$Y=-0.0299X^2 + 1.9403X$	0.5723
	Polynomial,3rd order	$Y=-0.0012X^3 –0.0032X^2+1.8055X$	0.5730
COD, TOC	Linear through origin	$Y=7.234\ X$	0.3136
	Linear , with intercept	$Y=5.2537\ X + 18.691$	0.3748
	Logarithmic	$Y=39.787\ \ln(X) –18.853$	0.3447
	Power	$Y=13.286\ X^{0.7105}$	0.4033
	Exponential	$Y=26.888\ e^{0.0896X}$	0.4001
	Polynomial, 2nd order	$Y=-0.0122X^2 + 5.4776X +17.797$	0.3748
	Polynomial,3rd order	$Y=-0.0723X^3 + 2.141X^2-13.673X+67.58$	0.3971

Table 4. Various models developed using neural networks

Model No.	Input Parameters	Output
A1	TOC, phenol, TSS, TDS.	BOD
A2	TOC, phenol, TSS, TDS.	COD
A3	TOC, Phenol, TSS, AMN	BOD
A4	TOC, Phenol, TSS, AMN	COD
A5	TOC, Phenol, TSS, TDS, KJN	BOD and COD
A6	TOC, Phenol, TSS	BOD
A7	TOC, Phenol, TSS	COD
A8	TOC, Phenol, TSS, TDS,	COD and BOD
A9	TOC, Phenol, TDS	BOD
A10	TOC, Phenol, TDS	COD
A11	TOC	BOD
A12	TOC	COD

The training of these models were started with the default values of NN Model software with a training count of 1000 and 4 hidden neurons in the hidden layer. From the next trail, the optimum training count for the network was decided. This was done by trial and error by checking the % ARE after each cycle of training .The optimum training count was the one which gave a lower % ARE for the test data. After deciding the maximum training count for these models the number of hidden neurons in the hidden layer were varied by small increments by maintaining constant training count till the desired % ARE for the test data was obtained. The training was also done for these models by varying the learning rates (LR) of the network (LR = 0.35 – 0.75), and it was observed that there was no refinement in the % ARE values when the learning rate was changed. After suitable training and testing with the data matrix, it was inferred that Model A3 with TOC, Phenol, TSS and AMN as the input parameters was best suited to predict BOD. The % ARE value was 11.6614 when tested with the test data at a training count of 5000 and 7 hidden neurons in the hidden layer. All the other models showed comparatively poorerresults than model A3 while both training and testing. Similarly for predicting COD, Model A10 with TOC, Phenol and TDS as the input parameters produced better results at a training count of 1500 and 8 hidden neurons in the hidden layer (% ARE = 6.9729). On the other hand, the models developed to give both BOD and COD as outputs, and with TOC, Phenol, TSS and TDS as the input parameters were able to predict good results for both BOD and COD compared to model A5. Model A8 produced better results at a training count of 5000 and 6 hidden neurons in the hidden layer. This model, when tested with the test data gave %ARE of 8.201 for BOD and 11.0835 for COD. The measured and predicted values of BOD and COD for models A3 and A10 are given in Figs. 1 and 2, while predictions from model A8 are shown collectively in Fig. 3 (A, B).

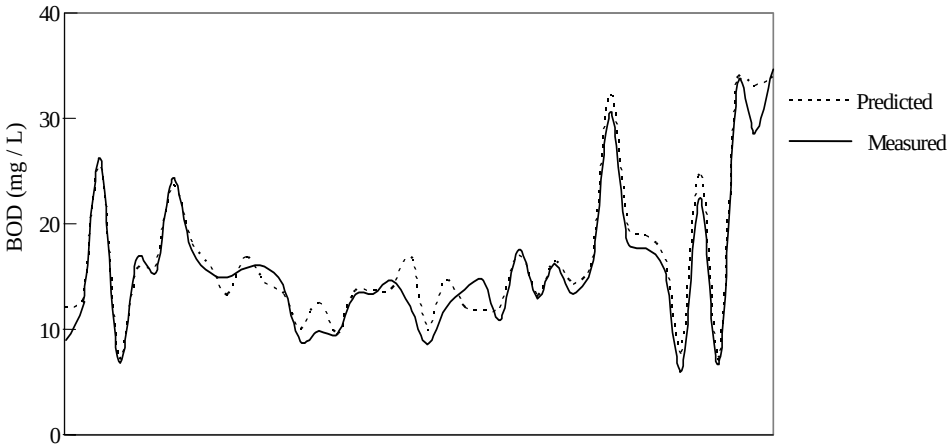

Fig. 1. ANN predictions of BOD concentration for the test data (Model A3)

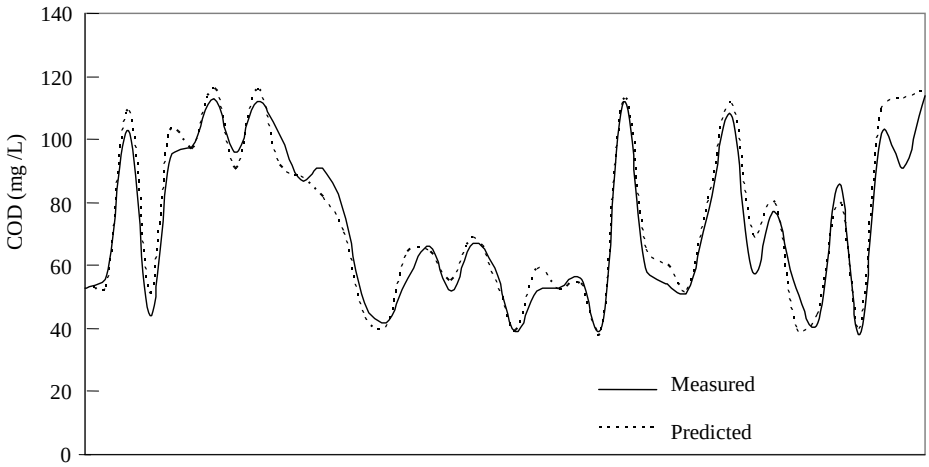

Fig. 2. ANN predictions of COD concentration for the test data (Model A10)

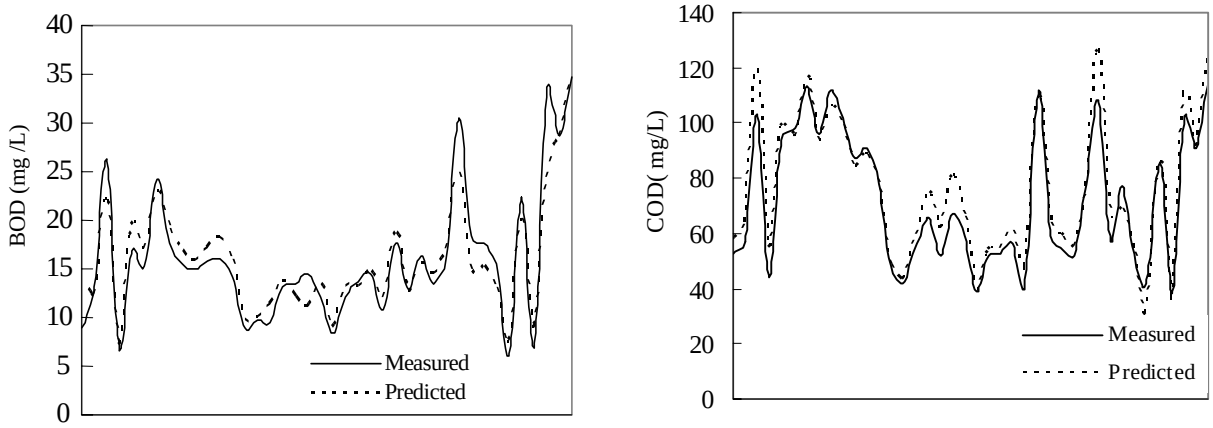

Fig. 3. ANN predictions of concentrations of (A) BOD and (B) COD for the test data (Model A8)

CONCLUSION

This study aimed in mapping the relations between various water quality parameters using regression analysis and artificial neural networks. The results from regression analysis provided different empirical equations for BOD and COD in terms of TOC. The results of models obtained from NN Model showed good significance at the 10 % level for the test data. Model A3 showed good results for BOD at a training count of 5000 and 7 hidden neurons in the hidden layer, while Model A10 gave good results for COD at a training count of 1500 and 8 hidden neurons. On the other hand, model A8 was developed to predict both BOD and COD simultaneously. Collectively, all these models gave very less % ARE values, indicating that the predictions are highly significant when tested with the test data. The empirical relations developed in this study and the developed ANN based models can be applied with high degree of confidence for refinery wastewaters.

REFERENCES

Bose, N. K., (1998). Neural Networks Fundamentals with Graphs, Algorithms and Applications, McGraw Hill publications.

Hawkin, S., (1994). Neural networks - A comprehensive foundation, Macmillan college publishing company, NY.

Jacek, Z. M., (1992). Introduction to Artificial Neural Networks, IInd Edition, Prentice Hall, NJ.

Maier, H. R. and Dandy, G. C., (2001). Neural network based modelling of environmental variables: A systematic approach, Math and Comp. Mod., **33**, 669-682.

Metcalf. Eddy., (1995). Wastewater Engineering, Treatment, Disposal and Reuse, Vth Edition, McGraw Hill, NY.

Pai, T. Y., Tsai, Y. P., Lo, H. M., Tsai, C. H. and Lin C. Y., (2007). Grey and neural network prediction of suspended solids and chemical oxygen demand in hospital wastewater treatment plant effluent, Comp. Chem. Eng., **31**, 1272-1281.

Sawyer, C. N., McCarty P. L. and Parkin G. F., (1994). Chemistry for Environmental Engineering; 4$^{th.}$ Ed., McGraw-Hill International Editions.

Van der Walt, T. J., Van Deventer, J. S. J. and Barnard, K., (1993). The estimation of kinetic viscosity of petroleum crude oil and fractions with a neural network, Chem. Eng. J., **51**, 151-158.