Malays. Appl. Biol. (2020) 49(1): 69–74

PREDICTION OF THE LEVEL OF WATER QUALITY INDEX USING ARTIFICIAL NEURAL NETWORK TECHNIQUES IN MELAKA RIVER BASIN

ANG KEAN HUA[1,2*]

[1]School of Biological Sciences, Faculty of Science and Technology,
Quest International University Perak (QIUP),
No. 227, Plaza Teh Teng Seng (Level 2), Jalan Raja Permaisuri Bainun,
30250 Ipoh, Perak Darul Ridzuan, Malaysia
[2]Department of Environmental Sciences, Faculty of Environmental Studies,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[*]E-mail: keanhua.ang@qiup.edu.my; angkeanhua@yahoo.com

Accepted 18 February 2020, Published online 30 June 2020

ABSTRACT

Artificial Neural Network (ANN) techniques were used to develop and validate water quality by predicting the Water Quality Index (WQI) in Melaka River Basin, Malaysia. Nine sampling stations were monitored in total. ANN techniques were applied for testing and developing the water quality prediction based on two sets of data. In the first data set, the independent water quality of six variables was used as input into ANN for trained, test and validated samples. In the second data set, a combination between Multiple Linear Regression (MLR) and ANN indicating only Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BOD), Suspended Solid (SS), and Ammoniacal-Nitrogen (AN) are accounted for training, testing and validating in modeling the water quality. Generally, MLR is used to exclude the lowest value invariance of independent variables, while rejecting the Dissolved Oxygen (DO) and pH. Based on the result of the correlation coefficient, the second set data (0.89) is marginally better than the first set data (0.87). These circumstances stated that predictions for WQI using ANN are acceptable, and the result is better when the variables of DO and pH are eliminated.

Key words: Artificial neural network; multiple linear regression; water quality index; water quality prediction

INTRODUCTION

Melaka River basin is considered as one of the most urbanizaed river basins in Malaysia, with a high density population due to attractive circumstances of availability of fertile lands, water supply for multipurpose usage (e.g. industrial, irrigation, and drinking) as well as transportation purposes. For the past decade, the Melaka state had experienced transformation of the agriculture landscape to industrial-commercial landscape due to extensive urbanization. The increasing population led to the increased urban activities thus expansion of industrial zones and housing estates. According to Hua (2017a; 2017b), the changes in landscape from vegetation into urban landscapes could indirectly affect water quality due to enhanced contamination of the river. Department ot Environment (DOE) report in 2012, explains the same while enlisting the discharge of municipal wastewater and industrial effluents as major pollutant sources being detected in Melaka River. Therefore, water quality management is considered as a major challenge especially determining the point and non-point sources of pollutants in the Melaka river.

In environmental perspective, water quality is affected by the quality of each water body; hence, it is better to integrate data of water quality to perform overall index which known as water quality index (Horton, 1965; Maier & Dandy, 1996; Xu & Liu, 2013). Therefore, WQI is a successful tool in water quality evaluation and has been applied in various studies by researchers and academicians (Tokatli, 2019; Stambuk-Giljanovic, 1999; Aliyu *et al.,* 2020). Nevertheless, reducing the cost and time is the main challenge in studying water quality and therefore, evaluation of the water quality through computer-aided tools play important roles,

* To whom correspondence should be addressed.

involving the urban water of non-linear behavior in the past, for prediction in future circumstances. ANN is most popular and reliable to apply for prediction of environmental with non-linear relationship data (Jain & Indurthy, 2003; Jiang *et al*., 2013; Sarkar & Pandey, 2015; Xu & Liu, 2013; Zhang & Stanley, 1997). This study aims to investigate and determine the level of WQI using ANN in the Melaka River basin.

MATERIALS AND METHODS

In this study, nine (9) sampling stations (Figure 1) and six (6) parameters were used to calculate the WQI of Melaka River, namely Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Suspended Solid (SS), Ammoniacal-Nitrogen (AN), and pH. The data used in this study was obtained from DOE representing five (5) years, which is 2001 to 2005 and it consists of 270 data sets (6 variables x 9 stations x 5 years).

Multiple Linear Regressions

Multiple linear regression (MLR) is used in the study to identify the relationship within water quality parameters as well as the impact on WQI. The MLR can be expressed in Eq. (1):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1}\beta_{p-1} + \varepsilon \quad (1)$$

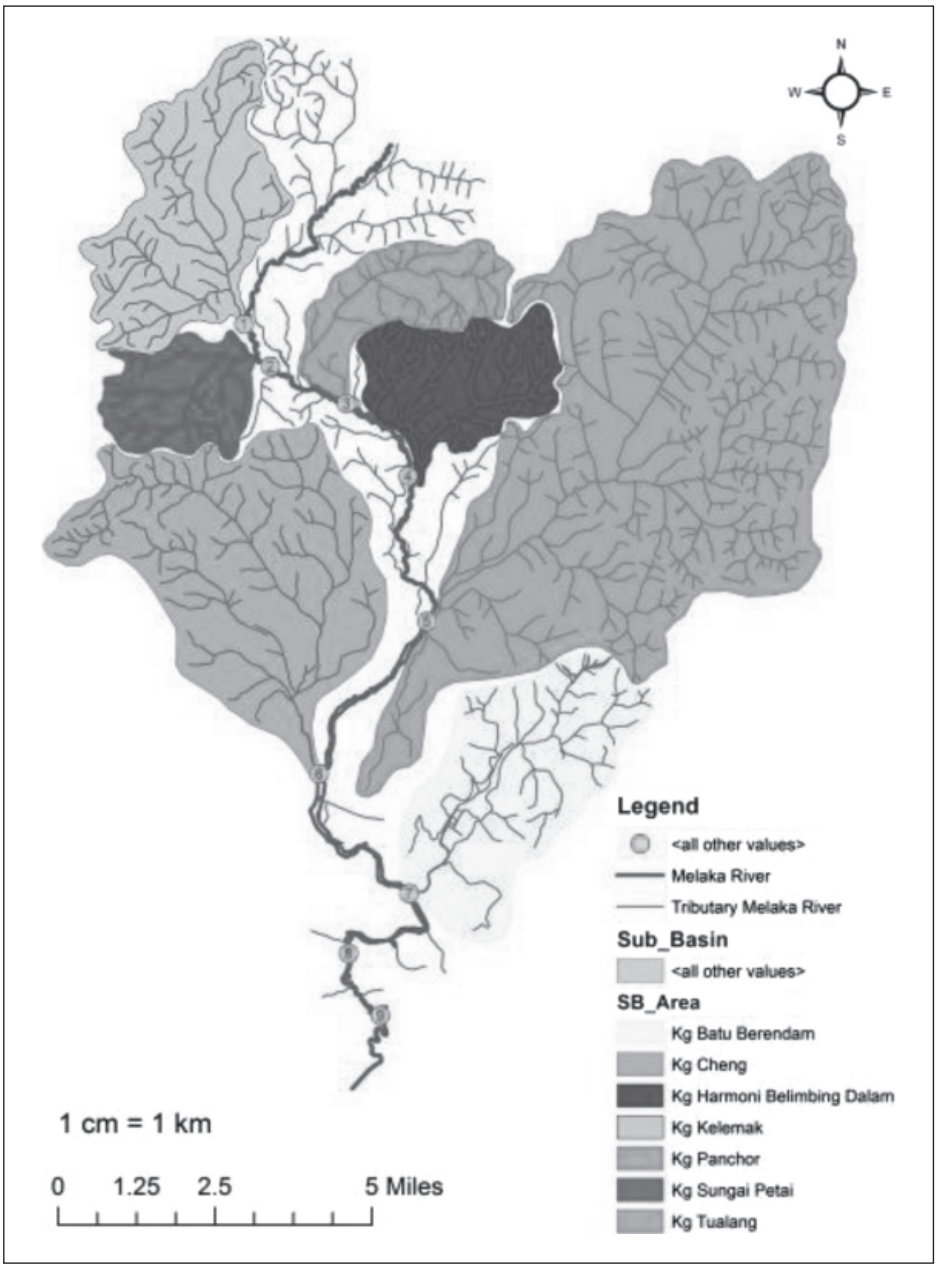

Fig. 1. 9 sampling stations along Melaka River basin.

where, Y refers to a responsive variable, while $p - 1$ are explanatory variables; x_1, x_2, …, x_{p-1}, and p refers to parameters (regression coefficient) of β_0, β_1, β_2, ..., β_{p-1}. Specifically, R^2 and adjusted R^2 value will be included in this study; whereby R^2 is the variance in Y, which is calculated to result in the model of regression, and adjusted R^2 is the variance in Y, based on the sample that was considered to perform the regression model. Based on Stein's formula, the R^2 model cross-validates using Eq. (2):

$$adjusted\ R^2 = 1 - [\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)](1 - R^2) \quad (2)$$

where, R^2 is the outcome value for the adjusted method, n is the number of subjects, and k is the number of predictors in the model.

Artificial Neural Network

Artificial neural network (ANN) models are characterized by node, network, and training (or learning) rules. The outcome of this model is based on the knowledge that consists of an interconnected set of weights. ANN comprises simple processing units that exist in large numbers, which are connected through excitatory or inhibitory interaction between each other (Figure 2). The processing units of the large numbers that were interconnected between each unit can be represented in three different layers, namely the input layer, hidden layer, and output layer. Input layer is referred to the six parameters (DO, BOD, COD, SS, AN and pH). The information in the network is characterixed as nodes, hidden layer (one or more) referring to intermediate computational layer consisting of multi-layer feed-forward network produced by individual hidden layer, as well as output layer refering to outcome that is produced during analysis which in this study is WQI. Training process during the input of the model is important to measure the error and is adjusted for the internal configurations (the weight for processing element that connected each other), which is required to bring down the error for the whole process involved (Abyaneh, 2014; Faruk, 2010; Chatterjee *et al*., 2017; Chebud *et al*., 2012; Sarkar & Pandey, 2015; Sudheer *et al*., 2003; Xu & Liu, 2013). In other words, the calibration of the model is determined only based on the training process (Abyaneh, 2014; Rojas, 1955; Sarkar & Pandey, 2015; Xu & Liu, 2013). In this study, two inputs were trained in ANN, randomly initialized on the 20 networks differently. Log-sigmoidal (also known as logsig) and the transfer function in linear form (can also be called as purelin) are applied in the network to functions as two activatiors in training process during the initial stopping approach, as well as the set of training data are applied to determine weights and biases.

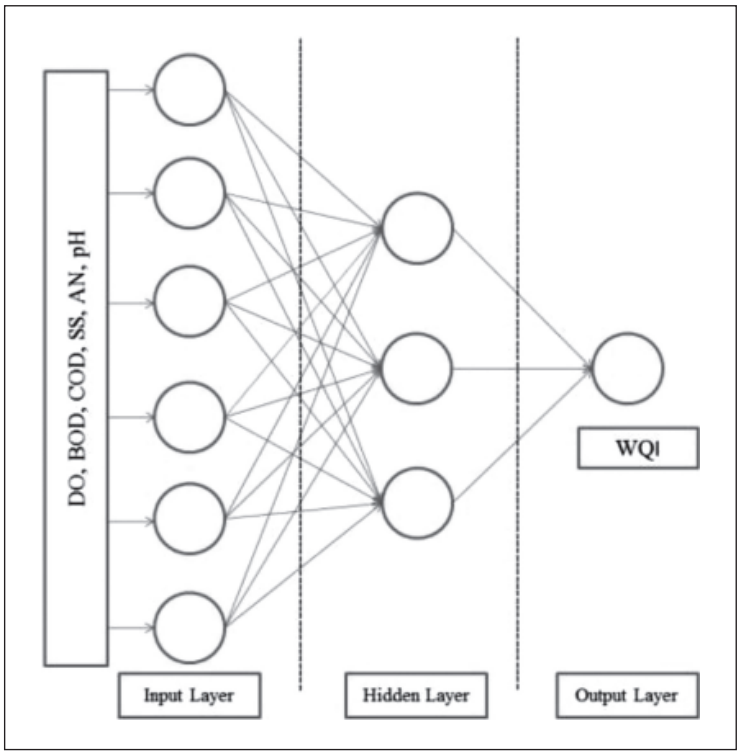

Fig. 2. Example of ANN configuration.

RESULTS AND DISCUSSION

Multiple Linear Regressions

The main purpose of using MLR in this model is to exclude the parameters in prediction of water quality with missing information. MLR is applied to identify the relationship within water quality parameters towards the WQI as a dependent variable. Through the analysis, the result indicates COD, BOD, SS, and AN meeting the requirement to be included in the equation, where all these variables have the variance in WQI with approximately 70%. DO (6%) and pH (3%), were excluded because they did not achieve the minimum requirement. The significant correlation were observed based on the selected four variables in WQI, namely COD (r=0.744), BOD (r=-0.551), SS (r=-0.270), and AN (r=-0.427) with the p-value more than 0.001 respectively.

Multiple correlation coefficient between the independent variables (or predictors) with the dependent variables (or outcome) are shown in Table 1, where the result between COD, BOD, SS and AN with WQI produces the R-value of 0.827 with $p<0.01$. Meanwhile, R^2 technique is used to measure possible variability of predictors outcome, which indicates that the result for variance of COD in WQI increses from 55.4% (R^2=0.554) to 68.4% when including the BOD, SS and AN. The result for the cross-validity of model-based on adjusted R^2 is 0.680. Hence, adjusted R^2 is likely the most ideal model generated as compared to the model based on R^2.

Table 2 indicates the regression model, of influences of input variables referring to the β values. Based on the influences of the variables to WQI, COD is considered to have the most influence on WQI with β = 0.415,as compared to BOD (β = -0.281), SS (β = -0.216) and AN (β = -0.138). In other words, β values describes the relationship for predictor on the WQI in the provided model. For instance, when β value is positive, the relationship between both variable are positive. The result of β values stated only COD as positive, while other variables remain negative (Table 2), as shown in the model below:

$$WQI = 47.933 + 4.348COD - 0.417BOD - 0.009SS - 1.276AN \quad (3)$$

Artificial Neural Network

MLR techniques provide only four variables that are considered for further analysis into ANN, which is COD, BOD, SS, and AN. Apart from using four variables, the models involving six, five, three or two variables are also for training. In other words, six different models were used to assess the performance of WQI prediction. The performance of the model is shown in Table 3. In training phases, the model of neural network [5,1,9,1] involved with predictors of five, hidden layers with one unit, hidden neurons of nine and the output with one neuron showing the result of 2.91 is better than 16.99 of the model of neural network [6,1,8,1] in predictors of six, hidden layer with one unit, hidden neurons of eight, and the output with one neuron.

Table 1. Summary of the regression model

Model	R	R^2	Adjusted R^2	Std. Error of Estimate
1	0.744[a]	0.554	0.553	13.11493
2	0.783[b]	0.613	0.611	11.99067
3	0.816[c]	0.666	0.664	11.40149
4	0.827[d]	0.684	0.680	11.12261

a. Predictors: (Constant), COD.
b. Predictors: (Constant), COD, BOD.
c. Predictors: (Constant), COD, BOD, SS.
d. Predictors: ((Constant), COD, BOD, SS, AN.

Table 2. Coefficient of regression model

Predictors	Unstandardized Coefficients		Standardized Coefficients	t	Significant level (p)
	B	Std. Error	Beta		
(Constant)	47.933	1.817		23.962	<0.000
COD	4.348	0.306	0.415	14.796	<0.000
BOD	-0.417	0.055	-0.281	-8.412	<0.000
SS	-0.009	0.004	-0.216	-6.448	<0.000
AN	-1.276	0.227	-0.138	-4.315	<0.000

Table 3. Performance of ANN model

Model	No. of hidden neurons	Sum of Square Error			Correlation Coefficient
		Training	Testing	Validation	
NN[4,1,15,1]	15	3.746	11.387	8.851	0.887
NN[5,1,9,1]	9	2.912	12.192	12.438	0.878
NN[2,1,7,1]	7	7.123	14.446	14.418	0.873
NN[6,1,8,1]	8	16.994	10.036	15.533	0.870
NN[3,1,10,1]	10	5.675	19.194	17.943	0.862

Meanwhile, testing phases indicate the model of neural network [6,1,8,1] of six predictors, hidden layer with one unit, hidden neurons of eight and the output with one neuron (10.04) is greater than the model of neural network [3,1,10,1] of three predictors, hidden layer with one unit, hidden neurons of ten and the output with one neuron (19.19). Lastly, the model of neural network [4,1,15,1] for predictors of four, hidden layer with one unit, hidden neurons of fifteen and the output with one neuron (8.85; 0.89) indicate best prediction of WQI with the lowest validation and correlation coefficient than the model of neural network [3,1,10,1] for predictor of three, hidden layer with one unit, hidden neurons of ten and the output with only one neuron (17.94; 0.86). In other words, the predictors model with four parameters (neural network [4,1,15,1]) is considered better as compared with other predictors model of artificial neural network. Therefore, the combination of ANN with MLR approach could provide the highest value of variance and best model in the prediction of the WQI data set.

CONCLUSION

This study predicts the water quality by reducing the parameters without having any loss of information. By applying MLR, the DO and pH are left out in the ANN analysis due to less variance of WQI and are excluded while prediction of WQI. Comparing the original with the excluded model, the inputs variable into ANN shows four parameters; COD, BOD, SS, and ANN that have the outcome performance with the best prediction among others. Simultaneously, the performance of the was optimum with 15 hidden neurons. This study is positively shows the performance of ANN, excluding the predictor DO and pH through MLR analysis to contribute an appropriate model of WQI prediction.

ACKNOWLEDGEMENT

The author would like to acknowledge the Faculty of Science and Technology, Quest International University Perak and the Faculty of Environmental Studies, Universiti Putra Malaysia for funding this grant to carry out the research.

REFERENCES

Abyaneh, H.Z. 2014. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*, **12(1)**: 40.

Aliyu, A.G., Jamil, N.R.B., bin Adam, M.B. & Zulkeflee, Z. 2020. Spatial and seasonal changes in monitoring water quality of Savanna River system. *Arabian Journal of Geosciences*, **13(2)**: 1-13.

Chatterjee, S., Sarkar, S., Dey, N., Ashour, A.S., Sen, S. & Hassanien, A.E. 2017. Application of cuckoo search in water quality prediction using artificial neural network. *International Journal of Computational Intelligence Studies*, **6(2-3)**: 229-244.

Chebud, Y., Naja, G.M., Rivero, R.G. & Melesse, A.M. 2012. Water quality monitoring using remote sensing and an artificial neural network. *Water, Air, & Soil Pollution*, **223(8)**: 4875-4887.

Department of Environment (DOE) Malaysia. 2012. *Malaysia Environmental Quality Report 2012*. Department of Environment, Ministry of Natural Resources and Environment Kuala Lumpur, Malaysia.

Faruk, D.Ö. 2010. A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence*, **23(4)**: 586-594.

Horton, R.K. 1965. An index number system for rating water quality. *Journal of Water Pollution Control Fed*., **37(3)**: 300-305.

Hua, A. 2017a. Application of Ca-Markov model and land use/land cover changes in Malacca River Watershed, Malaysia. *Applied Ecology and Environmental Research*, **15(4)**: 605-622.

Hua, A. 2017b. Analytical and Detection Sources of Pollution Based Environmetric Techniques in Malacca River, Malaysia. *Applied Ecology and Environmental Research*, **15(1)**: 485-499.

Jain, A. & Prasad Indurthy, S.K.V. 2003. Comparative analysis of event-based rainfall-runoff modeling techniques deterministic, statistical and artificial neural network. *Journal of Hydrological Engineering*, **8**: 93-98.

Jiang, Y., Nan, Z. & Yang, S. 2013. Risk assessment of water quality using Monte Carlo simulation and artificial neural network method. *Journal of Environmental Management*, **122**: 130-136.

Maier, H.R. & Dandy, G.C. 1996. The use of artificial neural networks for the prediction of water quality parameters. *Water Resources Research*, **32(4)**: 1013-1022.

Rojas, R. 1955. *Neural networks: A Systematic introduction*, Springer-Verlag, Berlin, 27-448.

Sarkar, A. & Pandey, P. 2015. River water quality modelling using artificial neural network technique. *Aquatic Procedia*, **4**: 1070-1077.

Stambuk-Giljanovic, N. 1999. Water quality evaluation by index in Dalmatia. *Water Resources*, **33**: 3423-3440.

Sudheer, K.P., Nayak, P.C. & Ramasastri, K.S. 2003. Improving peak flow estimates in Artificial Neural Network riverflow models. *Hydrological Processes*, **17**: 677-687.

Tokatli, C. 2019. Drinking Water Quality Assessment of Ergene River Basin (Turkey) by Water Quality Index: Essential and Toxic Elements. *Sains Malaysiana*, **48(10)**: 2071-2081.

Xu, L. & Liu, S. 2013. Study of short-term water quality prediction model based on wavelet neural network. *Mathematical and Computer Modelling*, **58(3-4)**: 807-813.

Zhang, Q. & Stanley, S.J. 1997. Forecasting raw-water quality parameters for North Saskatchewan River by neural network modeling. *Water Resources*, **31**: 2340-2350.