VOL. 11, NO. 20, OCTOBER 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

FIRST-ORDER INTERACTION MULTIPLE REGRESSIONS MODEL ON WATER QUALITY INDEX IN MANJUNG RIVER AND ITS TRIBUTARIES

Aminatul Hawa Yahaya[1], Muhammad Aminurrasyid Sutarsono[2] and Farish Ahmad[3]
[1]Technical Foundation Section, Universiti Kuala Lumpur, Malaysian Institute of Marine Engineering Technology, Perak, Malaysia
[2]Maritime Management Section, Universiti Kuala Lumpur, Malaysian Institute of Marine Engineering Technology, Perak, Malaysia
[3]Pusat Pengajian Sains Nautika, Fakulti Pelaut, Kapal Diraja Sultan Idris 1, Tentera Laut Diraja Malaysia, Perak, Malaysia
E-Mail: aminatulhawa@unikl.edu.my

ABSTRACT

This research demonstrated the procedures in choosing the best model in forecasting the water quality index (WQI) in Manjung River and its tributaries using multiple regressions. Six independent variables which are the water quality parameters and WQI as the dependent variable were included in this data set. The Multiple Regression (MR) models were involved is the first-order interaction with 57 possible models were considered. In this research, the process of getting the best model from the total of 57 possible models had been shown. The backward elimination of variables with the highest p-value was engaged to get the selected model. The best model includes using the first order interaction with variables of (DO, COD, BOD, SS, AN and pH). The best model obtains then been verified by the Mean Absolute Percentage Error (MAPE) calculation to quantify the models' relative overall fit.

Keywords: first-order interaction, multiple regression, water quality index, eight selection criteria.

INTRODUCTION

Water pollution is one of the most critical environmental issues in Malaysia and the situation will tend to worsen in the future unless proper preventive measures are undertaken. Water and land based activities such maritime industrial activities, conventional types of farming and mining are affecting the quality of local rivers, lakes, streams and also groundwater. Pollutants may come from point and non-point sources, for example, farms waste product can increase the concentration of nutrients in water which may cause eutrophication such as algae bloom. Other industrial activities, marine or land based can increase the concentrations of metals and toxic chemicals, suspended solids and temperature. Highly polluted waters may lower the amount dissolved oxygen needed for the survival of the ecosystem and also making the water unusable to the public [13].

As study proved by Environmental Section (1977) 42 tributaries in Peninsular Malaysia has been categorized as very polluted. Until the year 1999, DOE confirmed there were about 13 polluted tributaries all over Malaysia with 36 polluted rivers due to human activities such as agricultural farming, construction projects and heavy industries at the tributaries [6]. [18] also concluded that there were only 48 clean rivers in 1990 compared to only 32 rivers in 1999 that could still be categorized as clean or not polluted..

There are various definitions of what is water quality index. As stated by [19] it is simply a single numeric expression that interprets complex information obtained from any body of water, mostly related to water quality. The type and number of parameters that can be included in a WQI model varied depending on the designated water uses and local government preferences. Some of the frequently used factors worldwide include but not restricted to are DO, pH, BOD, COD, total suspended solid (TSS), coli form bacteria, temperature, nutrients (nitrogen and phosphorus) and etc.[7].

The main objective of a water quality index is to compile complex data of the any body of water into useful and understandable information to the mass public and also to the government as the ultimate policy maker [22]. It is a lot similar to the air quality index which shows whether the air quality is healthy or unhealthy. The use of an index to "grade" water quality is a controversial issue which is argued among water quality scientists. [11] debated that it is impossible for a single number to explain water quality as there are many other water quality parameters that are excluded in the index. Unlike the air quality index which aims directly to the public health such as respiratory illnesses, the water quality index on the other hand gives general information on the conditions of the body of water in the particular area and the levels of pollutions it is facing which may affect the public in the long run [2].

The objectives of the study are:

a) To identify the main parameter that has a significant contribution in the WQI at Manjung River.
b) To determine the best first order interaction multiple regression model to predict WQI.

METHODOLOGY

Scope of study

The study on WQI is focused along the Manjung river basin and its main tributaries. The tributaries which are identified in the study are connected directly to the main river. A total of six sampling stations were proposed and identified to measure the water quality in that area using latitude and longitude points derived from Google maps. The area of study is selectively chosen to produce varying results, according to the activities of that particular

area. The latitude and longitude of these sampling stations are listed in Table-1.

Table-1. Location of water sample taken along the Manjung river basin and its tributaries.

Stations	Latitude	Longitude	Description
1	4°24′N	100°41′E	Towards palm oil factory at Changkat Kruing
2	4°19′N	100°40′E	At shrimp farm outlet
3	4°18′N	100°40′E	Near shrimp farm at Sg. Pasir
4	4°17′N	100°40′E	At the area of floating fish cages
5	4°16′N	100°40′E	Next to Jalan David Sung near fertilizer factory
6	4°16′N	100°39′E	At port near Lumut Maritime Terminal

Data collection

The data analysis of this study was taken as a secondary data from a water sampling research of Manjung River. It was taken along the Manjung river basin at 6 sampling stations with five times of frequency for both tides (study period is within July 2012 and November 2012). Each parameter was analyzed based on the Water Quality Standard and Regulation in Malaysia.

Among that information are the 6 variables that are taken as the independent variables. They are:

1. SI-Sub index of parameter
2. DO-Dissolve Oxygen
3. BOD-Biological Oxygen Demand
4. COD-Chemical Oxygen Demand
5. AN-Ammonical Nitrogen
6. TSS-Suspended Solid
7. pH-Salinity

All of these data are quantitative in nature. DO and pH were done in situ whereas BOD, COD, AN and SS were done in situ and laboratory tested in Universiti Teknologi PETRONAS (UTP) laboratories. Each of these independent variables is represented by X1 until X6 with 43 variables for each data. These data are summarized in the table below.

STATISTICAL ANALYSIS

Multiple regression (MR) models with interaction

Multiple Regression (MR) is utalized to represent (predict) the variance in an interval dependent, based on linear combinations of interval, dichotomous, or dummy independent variables [9]. MR can establish that a set of independent variables explains a proportion of the variance in a dependent variable at a significant level (significance test of R^2), and can establish the relative predictive importance of the independent variables (comparing beta weights). Interaction effects are sometimes called moderator effects because the interacting third variable which changes the relation between two original variables is a moderator variable which moderates the original relationship. The impact of one variable depends on the level of the other variable when an interaction is present. One adds interaction terms to the model as cross products of the standardized independents and/or dummy independents, typically placing them after the simple "main effects" independent variables. The idea of multiple effects should be studied in research rather than the isolated effects of single variables is one of the important contribution of Sir Ronald Fisher [15]. The interpretation of individual variables can be incomplete or misleading due to the presence of interaction effects. The specific MR model that has been explained by [12] can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 + X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_I \quad (1)$$

where X is the random variable representing the ith value of the DV, Y. Thus, X_{1i}, X_{2i}…X_{ki} are the i^{th} value of IV for i = 1, 2 ... n.

Table-2. Description of variables involved in the model.

Variable	Description	Unit
Y	Water Quality Index (WQI)	mg/l
X1	SI Dissolved Oxygen (DO)	mg/l
X2	SI Biological Oxygen Demand (BOD)	mg/l
X3	SI Chemical Oxygen Demand (COD)	mg/l
X4	SI Ammonia Nitrates (AN)	mg/l
X5	SI Suspended Solid (SS)	mg/l
X6	SI Salinity (pH)	pH

MODEL FINDINGS

All possible models

In the construction of the MR models for this dataset, WQI would be the DV noted by Y whereas DO (X_1), BOD (X_2), COD (X_3), AN (X_4), SS (X_5) and pH (X_6) would be the Independent Variables (IV). All possible models, N can be calculated by using the formula:

$$N = \sum_{j=1}^{q} j\left(C_j^q\right) \quad (2)$$

The number of possible models generated is denoted by N, q is the number of variables and j = 1, 2... q. For this research, q = 6. Therefore, the number possible models are:

$$N=1\left(C_1^6\right)+2\left(C_2^6\right)+3\left(C_3^6\right)+4\left(C_4^6\right)+5\left(C_5^6\right)+6\left(C_6^6\right)=192 \quad (3)$$

Table-3. Generated possible models.

No. of Variables	Single	Interaction					Total
		1st	2nd	3rd	4th	5th	
2	15	15					30
3	20	20	20				60
4	15	15	15	15			60
5	6	6	6	6	6		30
6	1	1	1	1	1	1	6
Total	63	57	42	22	7	1	192

Selected model

Phase 2 is the selected model. It consists of the multicollinearity test and the coefficient test. Multicollinearity is defined as the intercorrelation of the sample independent variables or (IV). It is basically a test applied to the all possible model by removing multicollinearity sources occurred in each models. Multicollinearity exist if the correlation coefficient is greater than 0.95. Zainodin-Noraini multicollinearity remedial procedures had been adopted and details are explained in [3]. Pearson correlation analysis proves that there is a presence of multicollinearity between IV's in M81 and one variable (X12) have been removed from the models (M81.1.0).

Following the multicollinearity test, the coefficient test is carried out. The coefficient test on the other hand is used to test the coefficient of the corresponding variables and variables which are insignificant are eliminated as explained by [14]. To validate the elimination of the unuseful variable, Wald test [15] should be carried out to the possible models upon the end of all the removal procedure of insignificant variables. In this step, two irrelevant variables have been abolished from the model M81.1.0. At the end of this phase, only three variables have been left in the model (model M81.1.2) Table-4 shows the entered variable before the elimination procedure and the remaining variable after the removal of insignificant variables.

Table-4. Model 81.1.0 with entered variable before elimination procedure of insignificant variables and model M81.1.2 with remaining variable after elimination procedure of insignificant variables.

Model	Coefficients	Standard Error	t-Stat	P-value
Model M81.1.0				
Intercept	33.18623143	4.664163	7.115153	1.99E-08
x1	-0.161049406	0.094779	-1.69921	0.097674
x2	0.608553641	0.044664	13.62509	5.41E-16
x5	-0.206095251	0.07718	-2.67031	0.011193
x15	0.004418971	0.001496	2.954643	0.005417
x25	-0.001394539	0.000706	-1.97618	0.055626
Model M81.1.2				
Intercept	26.43243744	1.17759	22.44621	6.59E-24
x2	0.522959178	0.014869	35.1717	3.73E-31
x5	-0.100674173	0.033401	-3.01413	0.004513
x15	0.001848425	0.000408	4.529536	5.46E-05

Eight selection criteria (8SC)

Recognition of the best model should be based on 8SC as shown in [1]. The aim is to define a model with the smallest value of a criterion statistic. The calculation of the criterion statistics will be based on the Sum of Square Error (SSE), number of estimated parameters and the sample size.in Table-5.

From 192 possible models generated during first phase, only 57 models have been selected with the same SSE value and the number of model parameter. They are grouped together where any models from this group can be the selected models. The best model was then chosen from the selected models by using the 8SC based on the majority of least values on all 8 criteria as shown in Table-6. Thus, the best model generated is M81.1.2 as it has the lowest value from all 8SC.

Table-5. 8 Selection criteria for best model identification.

AIC [5]: $\left(\frac{SSE}{n}\right)(e)^{2(k+1)/n}$	RICE [17]: $\left(\frac{SSE}{n}\right)\left[1-\frac{2(k+1)}{n}\right]^{-1}$
FPE [4]: $\left(\frac{SSE}{n}\right)\frac{n+k+1}{n-(k+1)}$	SCHWARZ [20]: $\left(\frac{SSE}{n}\right)(n)^{2(k+1)/n}$
GCV [8]: $\left(\frac{SSE}{n}\right)\left[1-\frac{k+1}{n}\right]^{-2}$	SGMASQ [15]: $\left(\frac{SSE}{n}\right)\left[1-\frac{k+1}{n}\right]^{-1}$
HQ [10]: $\left(\frac{SSE}{n}\right)(\ln n)^{2(k+1)/n}$	SHIBATA [21]: $\left(\frac{SSE}{n}\right)\frac{n+2(k+1)}{n}$

Table-6. Values of 8SC selected model.

Model	SSE	AIC	RICE	FPE	SCHWARZ	GCV	SGMASQ	HQ	SHIBATA
M64.1.0	217.11	4.32	4.34	4.32	4.66	4.33	4.10	4.50	4.29
M69.0.2	299.97	5.75	5.77	5.75	6.06	5.76	5.55	5.92	5.74
M81.1.2	194.73	4.01	4.06	4.01	4.44	4.03	3.74	4.24	3.97
M65.1.0	3443.55	68.45	68.87	68.45	73.91	68.65	64.97	71.39	68.08
M66.0.1	3993.68	79.38	79.87	79.39	85.72	79.62	75.35	82.79	78.96
M67.0.1	5760.09	114.49	115.20	114.50	123.63	114.83	108.68	119.41	113.88
M83.1.2	2034.82	41.92	42.39	41.93	46.44	42.14	39.13	44.33	41.53
M106.4.3	2014.83	41.50	41.98	41.51	45.98	41.73	38.75	43.90	41.12
M68.1.1	6510.09	124.86	125.19	124.86	131.42	125.02	120.56	128.41	124.56
M73.0.1	4186.35	83.21	83.73	83.22	89.85	83.46	78.99	86.79	82.77
M74.0.1	5958.33	118.43	119.17	118.44	127.89	118.78	112.42	123.52	117.80
M75.1.0	5472.80	108.78	109.46	108.79	117.47	109.11	103.26	113.45	108.20
M76.0.1	7305.49	145.21	146.11	145.22	156.80	145.64	137.84	151.45	144.43
M77.1.1	10392.90	199.33	199.86	199.33	209.80	199.59	192.46	205.00	198.84
M78.1.1	13167.77	252.55	253.23	252.56	265.82	252.88	243.85	259.73	251.93
M85.3.1	3389.46	67.37	67.79	67.38	72.75	67.57	63.95	70.27	67.01
M86.0.2	3060.82	65.34	66.54	65.37	74.27	65.90	60.02	70.09	64.42
M93.2.2	295.02	5.86	5.90	5.86	6.33	5.88	5.57	6.12	5.83
M95.0.3	3621.20	74.59	75.44	74.61	82.64	75.00	69.64	78.90	73.90
M96.2.2	4093.16	81.36	81.86	81.37	87.85	81.60	77.23	84.85	80.92
M97.2.2	5472.80	108.78	109.46	108.79	117.47	109.11	103.26	113.45	108.20
M98.2.2	7305.49	145.21	146.11	145.22	156.80	145.64	137.84	151.45	144.43
M107.4.3	3472.55	71.53	72.34	71.55	79.25	71.92	66.78	75.66	70.87
M108.3.2	2538.98	56.17	57.70	56.22	65.50	56.87	50.78	61.10	55.05
M113.3.4	3587.84	73.91	74.75	73.93	81.88	74.30	69.00	78.17	73.22

Best model verification

Phase 4 and the last phase of the MR model building process is The Goodness-of-Fit test. It consists of the randomness test and normality test. It is commonly applied to the final best model. The randomness test is applied to determine that the residuals are randomly distributed. Since the value of Z = 0.637 < asymp. Sig (2-tailed) = 0.813, therefore, H_0 is accepted. There is enough evidence that the residual is randomly distributed. The normality test on the Kolmogorov-Smirnov statistics on the other hand is to ensure that the normality assumptions are adhered. Since the Kolmogorov-Smirnov statistics (0.637) and gives the significant p-value = 0.813>0.05, therefore, H_0 is also accepted. There is enough evidence at 0.05 significant levels that the standardized residual is normal. This statement is supported by the scatter plot and histogram in Table-6.From here, the best regression model would therefore be represented by:

$$WQI = 26.43 + 0.5229X_2 - 0.1006X_5 + 0.0018X_{15} \quad (4)$$

Table-7. One-sample Kolmogorov-Smirnov test.

Standardized Residual	
N	43
Kolmogorov-Smirnov Z	0.637
Asymp. Sig. (2-tailed)	0.813

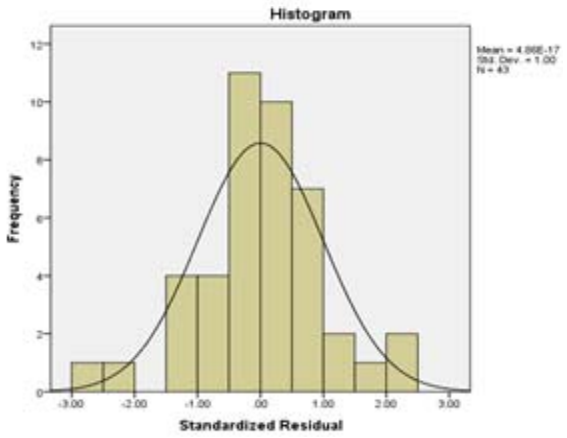

Figure-1. Histogram with normal curve.

Model accuracy measurement

The method used to measure the accuracy of a developed model is by using MAPE. It produces a measure of relative overall fit and is commonly used in quantitative forecasting methods in statistics. In [11] suggested that the absolute values of all the percentages errors are summed up and the average percentage is computed. MAPE is used to verify the best model and express its accuracy in percentage using the formula:

$$MAPE = \frac{1}{a}\sum_{t=1}^{q}\left|\frac{A_t - F_t}{A_t}\right| \times 100 \quad (5)$$

The actual value is denoted as A_t and the forecast or estimated value term as F_t. The difference between A_t and F_t, is divided by A_t. The absolute value of this calculation is summed for every fitted or forecast point in time and divided again by the total number of fitted points, a. In this case, the number of a = 3, which is reserved solely for this purpose.

Table-8. MAPE of random observations.

Y = At	x2	x5	x15	Ft	(At-Ft)/At	I(At-Ft)/At)I	x100
54.65	44.51	90.01	6650.84	52.94125	0.031267211	0.031267211	3.126721
43.96	45.69	67.47	2937.64	48.96397	-0.11383009	0.113830094	11.38301
84.00	100.87	79.00	7900.00	85.83148	-0.02180332	0.021803325	2.180332
							5.563354

The lower the MAPE value the better the model can be used in forecasting or predicting the missing values MAPE of 10% and lower is considered as highly accurate forecast, whereas a MAPE in the range between 11-25% is quite common which still gives a good forecast. 25%-50% is considered arguably a satisfactory forecast and anything above 50% is considered inaccurate and is not fit for forecasting. By substituting the remaining observation that has not been included in the model building analysis, the value of MAPE obtained is 5.56%. It is justified that this model could be use for forecasting, predicting WQI or estimating the missing value of parameters.

CONCLUSIONS

The WQI is one of the ways of measuring the health of a body of water. Even though Manjung River is a brackish type of river and it is unfit for domestic use and consumption due to its salinity level, its health is still vital for the survival of the ecosystem, living resources and its tranquility. In this study, it is clearly stated the contributions of each parameter in determining WQI. Suspended Solid (SS) are clearly dominant in determining the quality of the water as it levels determines the level of oxygen in the water, especially dissolved oxygen (DO) whereby there is a direct interaction with SS. The lower the level of SS means the higher the body of water can retain oxygen. As the samples are taken near industries and commercial farms and factories, the level of effluent, pollutants and other SS contributors are practically high, which marks its significance in the model. BOD is also significant as it determines the level of oxygen demand of living things in Manjung River. Other parameters such as pH, COD and NA are not significant mainly because of Manjung River is brackish and not fresh. Thus, any change in these three parameters does not have any significant impact compared to freshwater rivers, whereas it would change the characteristic drastically.

ACKNOWLEDGEMENTS

The data and samples obtained for this research is provided by Ms Nur Atika binti Mohd Rohani from Universiti Teknologi Petronas from her Final Year Project titled Study on Water Quality at Manjung River and Its Tributaries. The authors thank Nur Atika binti Mohd Rohani and her research team for providing the data set for this research. The authors would also like to thank the anonymous reviewers for their useful comments, suggestions and recommendations.

REFERENCES

[1] N. Abdullah, Z. H. Jubok and A. Ahmed. 2011. Improved stem volume estimation using p-value approach in polynomial regression models. Research Journal of Forestry. 5 (2): 50-65.

[2] A. M. Aenab and S. K. Singh. 2013. Evaluating water quality of Ganga river within Uttar Pradesh state by

water quality index analysis using C++ program. Civil and Environmental Research. 3(1): 57-65.

[3] A. H. Yahaya, N. Abdullah and H.J. Zainodin.2012. Multipleregression models up to first-order interaction on hydrochemistry properties. Asian Journal of Mathematics and Statistics. 5(4): 121-131.

[4] H. Akaike. 1970. Statistical predictor identification. Annals of the Institute of Statistical Mathematics. 22(1): 203-217.

[5] H. Akaike. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19(6): 716-723.

[6] Department of Environment Malaysia. 2006. Malaysia Environmental Quality Reports 2006. Ministry of Natural Resources and Environment Malaysia.

[7] N. Donia. 2011. Water quality management of lake Temsah, Egypt using geographical information system (GIS). International Journal of Environmental Science and Engineering.2: 1-8.

[8] G. H. Golub, M. Heath and G. Wahba. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics. 21(2): 215-223.

[9] Hair J. F., Black W. C. and Babin B. J. 2010. Multivariate Data Analysis: A Global Perspective. 7th Ed. Pearson Prentice Hall, New Jersey, USA.

[10] E. J. Hannan and B. G. Quinn. 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society: Series B (Methodological). 41(2): 190-195.

[11] Kristie W. 2011. Salinity Management Handbook. 2nd Ed. Department of Environment and Resource Management, Queensland, Australia.

[12] LevyP. S. and Lemeshow S. 2011. Sampling of Populations: Methods and Applications. 4th Ed. Wiley, New Jersey, USA.

[13] C. Y. Lin, M.H. Abdullah, S.M. Praveena, A.H.B. Yahaya and B. Musta. 2012. Delineation of temporal variability and governing factors influencing the spatial variability of shallow groundwater chemistry in a tropical sedimentary island. Journal of Hydrology.432: 26-42.

[14] Lind D. A., Marchal W. G. and Mason R. D. 2005. Statistical Techniques in Business and Economics. 16th Ed. McGraw-Hill Education, New York, USA.

[15] Pedhazur E. J. and Schmelkin L. P. 2013. Measurement, Design, and Analysis: An Integrated Approach. Psychology Press, New York, USA.

[16] Ramanathan R. 2005. Introductory Econometrics with Applications. South Western Educational Publishing.

[17] J. Rice. 1984. Bandwidth choice for nonparametric kernel regression. The Annals of Statistics. 12(4): 1215-1230.

[18] Ibrahim R. 2001. River water quality status in Malaysia. In: National Conference on Sustainable River Basin Management in Malaysia.

[19] H. Rubio-Arias, M. Contreras-Caraveo, R.M. Quintana, R.A. Saucedo-Teran and A. Pinales-Munguia. 2012. An overall water quality index (WQI) for a man-made aquatic reservoir in Mexico. International Journal of Environmental Research and Public Health. 9(5): 1687-1698.

[20] G. Schwarz. 1978. Estimating the dimension of a model. The Annals of Statistics. 6(2): 461-464.

[21] R. Shibata. 1981. An optimal selection of regression variables. Biometrika. 68(1): 45-54.

[22] K. Voudouris, A. Panagopoulos and J. Koumantakis. 2000. Multivariate statistical analysis in the assessment of hydrochemistry of the Northern Korinthia prefecture alluvial aquifer system (Peloponnese, Greece). Natural Resources Research. 9(2): 135-146.