Single Order of Multiple Regression Model of Water Quality Index (WQI) in Manjung River and its Tributaries

A. H. Yahaya, N. M. Salih, M.K. Puteri Zarina, W.M Dahalan

Abstract: ***This research highlights a multi-variety technique to examine the relationship between dependent and independent variable in predicting the water quality index in Manjung Rivers and its affluents. The model building process been used to analyse and generate the data. There are 63 possible models for single order multiple regressions. The number of possible model started to reduce as we started to eliminate insignificant variable. This model then needs to run under eight selection criteria to identify the best model. The best model will be certified by using Mean absolute percentage error (MAPE) in order to measure the validity of the model.***

Index Terms: Multiple Regression Model, Water Quality Index, Single Order, Relationship

I. INTRODUCTION

Water is the vital element for the living organism. It is importance for us to maintain the quality of water from being contaminated by the foreign substance or naturally contaminated. The water quality index objective is to provide the water quality information in the understandable and useable form that able to be understood by the public [8][10]. Multiple regressions are one of the statistical methods that can be used to measure the water quality index. Regression analysis is a statistical process for estimating the relationship between variables. It helps in understanding the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. There is several type of regression model and one of it is multiple regressions. Multiple regressions are an expension of simple linear regression and we can use this model to predict the water quality index. The multiple regression been used in this study cases to isolate the influences of the changes in water quality from all influences factors that might effecting the purpose of the value changes [16]. The main idea of this research is to build a new model of multiple regressions to predict the water quality index.

This will provide the reader with the fundamental understanding of the research. The primary objective of this inquire about are to recognize the mains parameters that have a critical contribution within the water quality list in Manjung Rivers and its tributaries and to decide the leading single order multiple regression model to foresee the water quality list in Manjung Rivers and its tributaries.

II. MATERIALS AND METHODS

A. Study Area

The scope of the study focused along the Manjung River and its main tributaries. This tributary is linked straightforwardly to the most stream. The examples of tributaries in Manjung River are Sg. Air tawar, Sg. Setiawan and Sg. Lumut.

Fig. 1 Location of study

B. Data Collection

The information that has been investigate in this research was taken as a secondary data from a water testing exploration of Manjung River. It was taken along the Manjung stream basin at 6 testing stations with five times of recurrence for both tides (study period is within July 2012 and November 2012). Each parameter was inspected based on the Water Quality Standard and Regulation in Malaysia.

Among that data are the 6 variables that are taken as the autonomous factors considered as Sub index of parameter (SI), Dissolve Oxygen (DO), Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD, Ammonical Nitrogen (AN), Total Suspended Solid (TSS) and Salinity (pH).

Revised Manuscript Received on October 05, 2019.

A. H. Yahaya, Universiti Kuala Lumpur – Malaysian Institute of Marine Engineering Technology (UNIKL-MIMET) ,Lumut, Perak, Malaysia

N. M. Salih, Universiti Kuala Lumpur – Malaysian Institute of Marine Engineering Technology (UNIKL-MIMET) ,Lumut, Perak, Malaysia.

M. K. Puteri Zarina, Universiti Kuala Lumpur – Malaysian Institute of Marine Engineering Technology (UNIKL-MIMET) ,Lumut, Perak, Malaysia

W.M Dahalan, Universiti Kuala Lumpur – Malaysian Institute of Marine Engineering Technology (UNIKL-MIMET) ,Lumut, Perak, Malaysia.

Retrieval Number: L40131081219/2019©BEIESP
DOI: 10.35940/ijitee.L4013.1081219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

III. STATISTICAL ANALYSIS

A. Statistical Method

Multiple regression is a very useful extension of simple linear regression that used several variables rather than just one to forecast a value on a quantitatively measured criterion variable [6]. Multivariate methods such as multiple regression can paint a more complete picture of how the world works. The goal of multiple regression is to produce a model in the form of a linear equation that identifies the best weighted combination of IV in the study to optimally predict the DV. Interaction impacts constitute the combined impacts of factors on the basis or subordinate degree. When an interaction impact is display, the effect of one variable depends on the level of the other variable. Portion of the control of MR is the capacity to appraise and test interaction impacts when the indicator variables are either categorical or continuous. When interaction impacts are show, it implies that elucidation of the individual variables may be deficient or deceiving.The specific MR model can be stated as follows:

$$Y_{pred} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

In this equation, Y_{pred} is the predicted score on the criterion variable, the X_{si} are the predictor variables with interaction in the equation, and the β_s the weights or coefficients associated with the predictors. Because this is a raw score equation, it also contains a constant, shown as β_0 in the equation.

Table. 1 Description of variable shows in the model

VARIABLES	DESCIPTION
Y	Water Quality Index
X1	SI Dissolved Oxygen (DO)
X2	SI Biological Oxygen Demand (BOD)
X3	SI Chemical Oxygen Demand (COD)
X4	SI Ammonia Nitrates (AN)
X5	SI Suspended Solid (SS)
X6	SI Salinity (pH)

B. Model Result

Phase 1: All possible models

Phase 1 is to identify all possible model in the data from single order multiple regression up to 5th interaction. For single order multiple regression, there are 63 feasible models. Within the advancement of the MR models for this datasets, Water Quality Index (WQI) would be the Dependent Variable (DV) marked by Y, while, DO (X1), BOD (X2), COD (X3), AN (X4), SS (X5) and pH (X6) would be the Independent Variables (IV). N denoted as all possible models, can be computed by using the formula:

In this formula, N is the number of possible models generated and q is the number of variables and j = 1, 2... q [11]. For this research, q = 6 and the possible model are:

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

$$N = \left(C_1^6\right) + \left(C_2^6\right) + \left(C_3^6\right) + \left(C_4^6\right) + \left(C_5^6\right) + \left(C_6^6\right) = 63$$

Phase 2: Selected Model

Phase 2.1 Multicollinearity test

Multicollinearity (MC) is the result of strong correlations of IV. Correlation coefficient with higer value will surge the standard error of the beta coefficients and create evolution of the special role of each independent triggered in difficult or inconceivable result. Multicollinearity occur if Correlation Coefficient is greater than 0.95. Zainodin-Noraini multicollinearity alterative strategies has been implement in this study [1] [2].

Table. 2 Multicollinearity Test

	y	*x1*	*x2*
y	1		
x1	0.81154915	1	
x2	0.992103651	0.776110225	1

The MC is as it were recognized in case the Correlation Coefficient between independent variables is more than 0.95 and in case the MC is occur, the variables have to be expel since it don't shows any relationship between the dependent and independent variables. In case the MC isn't identified, the variable is still substantial to be utilized. This MC test in obligatory to each model in order to dispose of any irrelevant relationship. In this simulation purpose, M7 has been chosen to demonstrate the elimination process. There are, no MC problem has been discovered in M7 as seen in Table 2.

Phase 2.2 Coefficient test

Following of that, the coefficient test should be implemented as an elimination method of irrelevant variable by utilizing the backward elimination [17]. This is to disposed all variables with p value higher than 0.05.

Table. 3 Coefficient Test

	P-value
Intercept	0.0000
X1	0.0004
X2	0.0000

As seen in Tables 3, none of the variable have p value higher than 0.05, so all the variable can be utilized because it give a significant relationship. If the p values are more than 0.05, the variables must be eliminated.

Phase 3: Best Model

From 63 potential models produced during the phase of this analysis, as it were 18 models have been chosen with the same SSE value and number of model parameter.

These models at that point been assembled and any models from this group can be the chosen model.

The best model was at that point chosen from the selected models by utilizing the 8SC based on the majority of smallest values as appeared in Table 4.

Table. 4 8 selection criteria for best model identification

AIC:	$\left(\frac{SSE}{n}\right)(e)^{2(k+1)/n}$	RICE:	$\left(\frac{SSE}{n}\right)\left[1-\frac{2(k+1)}{n}\right]^{-1}$
FPE:	$\left(\frac{SSE}{n}\right)\frac{n+k+1}{n-(k+1)}$	SCHWARZ:	$\left(\frac{SSE}{n}\right)(n)^{2(k+1)/n}$
GCV:	$\left(\frac{SSE}{n}\right)\left[1-\frac{k+1}{n}\right]^{-2}$	SGMASQ:	$\left(\frac{SSE}{n}\right)\left[1-\frac{k+1}{n}\right]^{-1}$
HQ:	$\left(\frac{SSE}{n}\right)(\ln n)^{2(k+1)/n}$	SHIBATA:	$\left(\frac{SSE}{n}\right)\frac{n+2(k+1)}{n}$

Where, n would be the number of observations, (k+1) is the number of model's parameters and SSE the sum of square of error. The Akaike Information Criterion (AIC) [3] and Finite Prediction Error (FPE) [4] are developed by Akaike. The Generalised Cross Validation (GCV) is developed by Golub et al. [5] while the HQ criterion is suggested by Hannan and Quinn [7]. The RICE criterion is discussed by Rice [13] and the SCHWARZ criterion is discussed by Schwarz [14]. The SGMASQ is developed by Ramanathan [12] and the Shibata criterion is suggested by Shibata [15].

Table. 5 Best Model Selection

Selected Model	SSE	AIC	RICE	FPE	SCH WARZ	GCV	SGM ASQ	HQ	SHI BATA
M1.0.0	6510.09	124.85	125.19	124.86	131.42	125.02	120.55	128.41	124.55
M2.0.0	299.96	5.75	5.76	5.75	6.05	5.76	5.55	5.91	5.739
M3.0.0	6631.18	127.18	127.52	127.18	133.86	127.34	122.79	130.79	126.87
M4.0.0	10392.90	199.32	199.86	199.33	209.88	199.58	192.46	204.99	198.84
M5.0.0	13167.77	252.54	253.22	252.55	265.81	252.87	243.84	259.73	251.93
M6.0.0	17205.17	329.98	330.86	329.99	347.32	330.41	318.61	339.36	329.18
M7.0.0	**217.110**	**4.31**	**4.34**	**4.31**	**4.66**	**4.32**	**4.09**	**4.50**	**4.29**
M8.0.0	3443.55	68.44	68.87	68.45	73.91	68.65	64.97	71.38	68.08
M9.0.0	3993.68	79.38	79.87	79.38	85.71	79.61	75.35	82.79	78.95
M10.0.0	5760.08	114.49	115.20	114.50	123.63	114.83	108.68	119.41	113.87
M16.0.0	4186.34	83.21	83.72	83.21	89.85	83.45	78.98	86.78	82.76
M17.0.0	5958.33	118.43	119.16	118.44	127.88	118.784	112.42	123.51	117.79
M18.0.0	5593.43	111.17	111.86	111.19	120.05	111.51	105.53	115.95	110.58
M19.0.0	7305.49	145.20	146.10	145.22	156.80	145.64	137.83	151.44	144.43
M26.0.0	2034.81	41.91	42.39	41.92	46.43	42.14	39.13	44.33	41.52
M38.0.0	3621.20	74.59	75.44	74.61	82.64	74.99	69.63	78.89	73.90
M51.0.1	3392.82	69.88	70.68	69.90	77.42	70.26	65.24	73.92	69.24
M62.0.3	291.21	5.78	5.82	5.78	6.25	5.80	5.49	6.03	5.75

Table 5 shows the best model selection process. Based on the result in Table 5, M7.00 has been chosen as best model.

Phase 4: Best model verification

The ultimate stage of model building is implementation of the Goodness-of-Fit on the final best model. The goodness-of-fit consist of the randomness test and normality test. Randomness test is to verify that the residuals are randomly dispersed and normality test on the Kolmogorov-Smirnov statistics is to guarantee that the normality assumptions are not violated. Since the value of Z = 0.645 < Significant value = 0.799,

hence, H0 is not rejected and this test affirm the conclusion that there's sufficient prove that the error is randomly dispersed. Since the Kolmogorov-Smirnov statistics (0.645) gives the significant
p-value = 0.799 > 0.05, hence, H0 is accepted. There's sufficient prove at 0.05 significant levels that the standardized residual is normal. This explanation is supported by the histogram in Figure 2.From here, the best regression model would therefore be represented by:

WQI= 20.7422+ (0.1091*DO) + (0.5252*BOD)

Where X_1 is SI*DO, and X2 SI*BOD. This interaction factor between SI*DO and SI*BOD may be causes whereby increased biochemical oxygen demand can affect the dissolved oxygen.

Model accuracy measurement

The Mean Absolute Percentage Error (MAPE) is used in quantitative forecasting methods because it produces a measure of relative overall fit. The absolute values of all the percentages errors are summed up and the average is computed [9]. In this study MAPE is used to verify the best model obtain. It is usually express accuracy as percentages and is defined by the formula:

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

The definition of MAPE formula involved, A_t as the actual value and F_t, as the forecast (estimated) value. The contrast between At and Ft, is separated by the genuine value At once more. The outright value of this computation is summarize for each fitted or predicted point in time and divided once more by the overall number of fitted point's a. For this study, the number of a = 3, number of data withdrawn at the early stages for this purpose. The interpretation of MAPE results is the average to judge the accuracy of the forecast as less than 10% is a highly accurate forecast, 11% to 20% is a good forecast, 21% to 50% is a reasonable forecast, 51% or more is an inaccurate forecast By substituting the reserved value that has not been included in the model building analysis, the value of MAPE obtained is 3.11901%. This value indicates that this model could be highly accurate for estimation of missing value or forecasting.

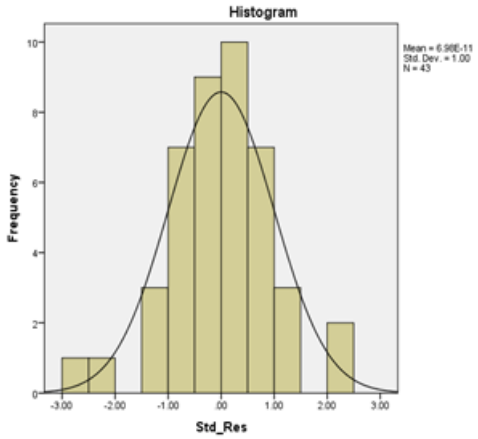

Fig. 2 Histogram with normal curve

IV. CONCLUSION

At the end of this research we able to distinguish the mains parameters that have a significant contribution in the water quality index in Manjung Rivers and its tributaries. We also able to identify the best single order multiple regression models to forecast the water quality index in Manjung Rivers and its tributaries. This new technique is more simple and accurate. The number of parameter that been used are lesser than the previous technique.

REFERENCES

1. Abdullah, N., H.J. Zainodin and A. Ahmed, 2011. Improved stem volume estimation using p-value approach in polynomial regression models. Res. J. Forest., 5: 50-65.
2. Abdullah, N., H.J. Zainodin and J.B.N. Jonney, 2008. Multiple regression models of the volumetric stem biomass. WSEAS Trans. Math., 7: 492-502.
3. Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Stat. Math., 22: 203-217.
4. Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control, 19: 716-723.
5. Golub, G.H., M. Heath and G. Wahba, 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21: 215-223.
6. Hair, J.F., W.C. Black and B.J. Babin, 2010. Multivariate Data Analysis: A Global Perspective. 7th Edn., Pearson Education Inc., New Jersey, USA., ISBN: 9780135153093, Pages: 800.
7. Hannan, E.J. and B.G. Quinn, 1979. The determination of the order of an autoregression. J. R. Stat. Soc. Ser. B, 41: 190-195.
8. Kristie, W., 2007. Salinity Management Handbook. West Region Publ., South Queensland, Australia
9. Levy, P. and S. Lemeshow, 1991. Sampling of Populations: Methods and Applications. John Wiley and Sons Inc., New York, USA.
10. Lin, C.Y., M.H. Abdullah, S.M. Praveena, H.Y. Aminatul and B. Musta, 2012. Delineation of temporal variability and governing factors influencing the spatial variability of shallow groundwater chemistry in a tropical sedimentary Island. J. Hydrol., 432-433: 26-42.
11. Pedhazur, E.J. and L.P. Schmelkin, 1991. Measurement, Design and Analysis: An Integrated Approach. Routledge, Hillsdale, NJ., USA., ISBN-13: 9780805810639, Pages: 819.
12. Ramanathan, R., 2002. Introductory Econometrics with Applications. 5th Edn., South-Western/Thomson Learning, Ohio, USA., ISBN-13: 9780030341861, Pages: 688.
13. Rice, J., 1984. Bandwidth choice for nonparametric kernel regression. Ann. Statistics, 12: 1215 1230.
14. Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat., 6: 461-464.
15. Shibata, R., 1981. An optimal selection of regression variables. Biometrika, 68: 45-54.
16. Voudouris, K., A. Panagopoulos and J. Koumantakis, 2000. Multivariate statistical analysis in the assessment of hydrochemistry of the Northern Korinthia prefecture alluvial aquifer system (Peloponnese, Greece). Nat. Resour. Res., 9: 135-143.
17. Zainodin, H.J., A. Noraini and S.J. Yap, 2011. An alternative multicollinearity approach in solving multiple regression problem. Trends Applied Sci. Res., 6: 1241-1255.

AUTHORS PROFILE

First Author Aminatul Hawa Yahaya served in Universiti Kuala Lumpur, Malaysian Institute of Marine Engineering Technology (UniKL MIMET) for over 13 years. Immediately after graduation, she joined Universiti Teknologi MARA (UiTM) as a lecturer for 3 years before joining UniKL MIMET in 2006. She is currently a Lecturer in Maritime Management Section, UniKL MIMET.

She also appointed as Final Year Project Coordinator under Research and Innovation Section. She previously holds posts in the institute where she was a Head of Section Applied Science and Technology. She graduated with Bachelor of Science (Honour) majoring in Statistics in 2002 from Universiti Putra Malaysia (UPM) and with Master of Applied Statistics from Universiti Malaya (UM). Her research interest are in applied statistics focusing on regression model development, model building procedure and logistic regression. The main programming tool for his research was SPSS. She got expertise in SPSS and practicing it since 2010. She also had a great contribution in publishing articles, journals, symposium papers in the field of Applied Statistics.

Second Author Noorazlina is a teaching staff for Marine Electrical and Electronic Technology Section in Universiti Kuala Lumpur Malaysian Institute of Marine Engineering Technology (UniKL MIMET) in Lumut Perak. She has working experience in educational field for at least 17 years since she had her first experience in Twintech University College back in year 1999. She obtained her first Bachelor degree in Electrical (Power) in De Montfort University, Leicester, United Kingdom in 1997. Her Master Degree is in Telecommunication and Information Engineering from Universiti Teknologi MARA, Shah Alam, Selangor in 2008. She has an administrative experience for about 10 years as the Dean and Coordinator program for engineering faculty at the University College. She is currently teach EE courses such as Digital Electronic System, Electronic Communications and Introduction to Electrical and Electronics..

Third Author Wardiah Mohd Dahalan received her B.Eng. (Hons) in Electrical & Electronics Engineering in 1996 from University of Dundee, UK and Master in Decision Science from Universiti Utara Malaysia. She received PhD degree from the University of Malaya, Kuala Lumpur, Malaysia in 2014. She is currently Senior Lecturer at Department of Marine Electrical and Electronics Engineering, University of Kuala lumpur (UniKL) and Head of Research and Innovation. Her research interests are renewable energy, reconfiguration area and optimization techniques. She is a member of IEEE, Rina-ImaRest and Malaysia Board of Technologist (MBOT).

Forth Author Puteri Zarina Megat Khalid is currently attached to Universiti Kuala Lumpur Malaysian Institute of Marine Engineering Technology (UniKL MIMET), Lumut as a senior lecturer-cum-Deputy Dean of Student Development & Campus Lifestyle. Among her research interests are language teaching and learning, modality analysis, pragmatics, English for Specific Purposes (ESP), genre analysis and Systemic Functional Linguistics. She received her Ph. D. in English Language from University of Glasgow, Scotland. She is also an appointed member of UniKL Research Journal Editorial Board and Editorial Board Member for UniKL MIMET Research Bulletin “Marine Frontier@UniKL” and several other publications. A member of International Systemic Functional Linguistics Association (ISFLA), Malaysian Society for Engineering & Technology (MySET) and Institute of Marine, Science and Technology (IMAREST).

.