VOL. 16, NO. 2, JANUARY 2021 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

INPUT VARIABLE SELECTION FOR HOURLY OZONE (O_3) CONCENTRATION PREDICTION IN URBAN AREA

Nur Nazmi Liyana Mohd Napi[1], Samsuri Abdullah[1], Amalina Abu Mansor[2], Ali Najah Ahmed[3], Wan Nurdiyana Wan Mansor[1] and Marzuki Ismail[2,4]
[1]Air Quality and Environment Research Group, Faculty of Ocean Engineering Technology and Informatics, University Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia
[2]Faculty of Science and Marine Environment, University Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia
[3]Department of Civil Engineering, Institute of Energy Infrastructure (IEI), College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang, Selangor Darul Ehsan, Malaysia
[4]Institute of Tropical Biodiversity and Sustainable Development, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia
E-Mail: samsuri@umt.edu.my

ABSTRACT

The higher emission of the air pollutants from motor vehicles in urban areas increases the formation of ozone in the atmosphere and give a negative impacts on human health, especially. The multicollinearity becomes a problem as many parameters were considered in the development of air pollutant prediction models. The aimed of the study is to develop the best $O_{3,\ t+1}$ prediction model in Terengganu's urban areas to help local authorities by providing early information about the air quality level. The air pollutants and meteorological data from the year 2010 to 2013 were obtained from the Department of Environment, Malaysia. There are three different models were developed based on three different input selection method. MLR_3 had highest correlation coefficient R^2, 0.411 compared with MLR_1 (0.382) and MLR_2 (0.399). Hence, it was selected as the best-fitted model to predict the O_3 concentration in Terengganu's urban areas.

Keywords: forecasting, multiple linear regression, ozone, meteorological, ozone precursor, urbanization.

INTRODUCTION

Urban areas are very scenario with having a high density of population, which influenced the higher number of motor vehicles used in daily activities. The higher number of motor vehicles resulting in higher emission of air pollutants which becomes the main concern as it influenced the air quality in urban areas (Zhou *et al*., 2018; Cho & Choi, 2014; Bereitschaft & Debbage, 2013). O_3 was categories as six criteria pollutants that give an adverse impact to human health such as cardiovascular disease, respiratory problem, cancer and even mortality that involved infant, children and adulteries (Ghorani-Azam *et al*., 2016; Ismail *et al*., 2015).

The regression statistical prediction model has been widely developed in several countries to predict the air pollutant especially on O_3 concentration in the different background areas (Liu *et al*., 2019; Capilla, 2016). The meteorological parameters and its precursors become the main variables in predicting the O_3 concentration in the ambient air (Lu *et al*., 2019; Zainordin *et al*., 2017). However, those parameters (independent variables) might introduce the multicollinearity problem which can affect the accuracy of model prediction result (Alin, 2010). In this study, three different multiple linear regression (MLR) models were developed based on three different selections of input variables to predict the next hour of O_3 concentration in an urban area at Terengganu.MLR_1 - using all parameters that involve in this study as input, MLR_2 - based on the correlation analysis, and MLR_3 - selected the input based on the Principal Component Analysis (PCA). The study aimed is to develop the best $O_{3,\ t+1}$ prediction model to help the local authorities by giving the preliminary alert information about the air quality level at the certain areas, as to help them to prepare the mitigation measure to preserve the better air quality.

MATERIALS AND METHODS

Study Area

The study area was carried out at Terengganu which is located on the east coast of Peninsular Malaysia with the total area 13, 035 km^2 and had the total population about 1, 015, 776 in 2010 (Department of Statistics, 2010). The air quality monitoring station (AQMS) in the Chabang Tiga Primary School, Kuala Terengganu (N05˚18.455'; E103˚07.213') was selected as it is near the Kuala Terengganu city centre which experienced the high volume of traffic during peak hour. Figure-1 shows the selected location of AQMS in an urban area at Terengganu.

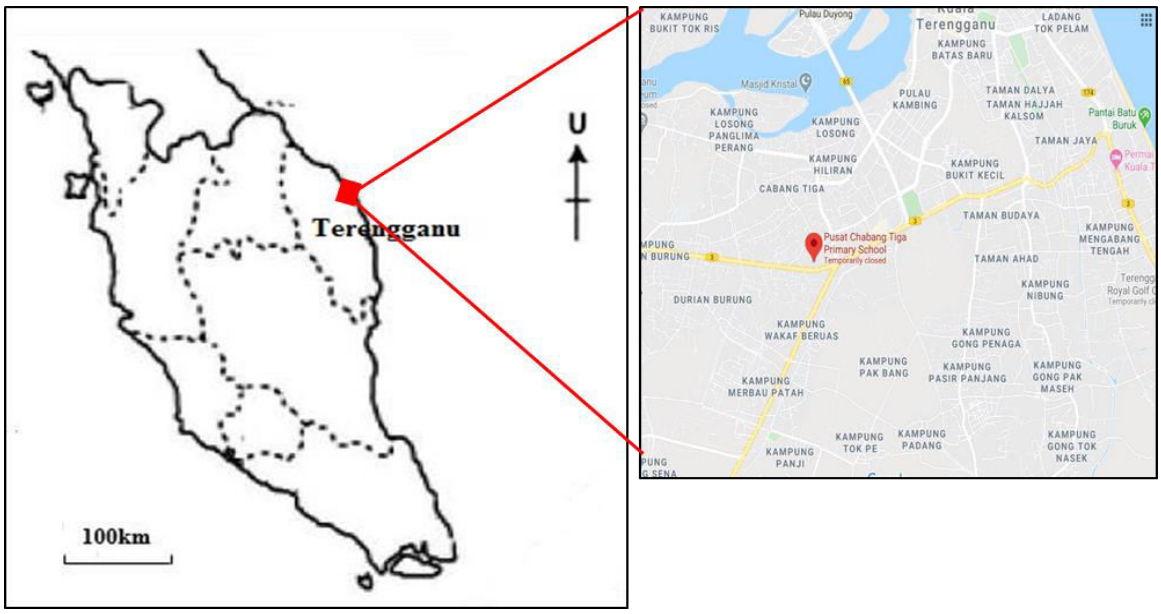

Figure-1. Location of the study area.

Data Acquisition

This study is based on 4-years data from the year 2010 until 2013 that acquired from the Air Quality Division, Department of Environment (DOE), Ministry of Environment and Water, Malaysia. It was arranged and tabulated in the Microsoft Excel Spreadsheet ® 2016 and analysed using Statistical Packages for Social Sciences (SPSS®) version 25. The parameters that involved in this study were ground-level ozone (O_3, ppm), ozone precursors such as nitrogen oxide (NO, ppm), nitrogen dioxide (NO_2, ppm) and carbon monoxide (CO, ppm), and meteorological factors, ambient temperature (T, ^{0}C), relative humidity (RH, %) and wind speed (WS, km/hr). Alam Sekitar Sdn. Bhd (ASMA) was appointed by the DOE to monitor, managed, and ensured the quality control and assurance of the monitoring data by all the instrument were having daily calibration using zero air and standard gas concentration (Abdullah *et al*., 2020; Banan *et al*., 2013). All the missing data that occurred due to equipment malfunction and calibration were omittedto avoid introducing a bias in the data analysis (Kang, 2013). The data was divided into 7:3 ratio for model development and validation (Abdullah *et al*., 2018; Roy and Ambure, 2016). All the parameters data were normalized by using min-max technique and range it from 0 to 1 value before it had been using in model development as to ensure no bias have occurred in the statistical data analysis result (Abdullah *et al*., 2018). The equation 1 is shown the min-max technique equation from the normalization process.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where,

$x = (x_1 \ldots, x_n)$
z_i = Normalized data.

Spearman Correlation

Spearman correlation is the non-parametric statistical method that used to investigate the relationship in term of strength and direction between two variables (Wang *et al*., 2019; Borkowf, 2000). The correlation coefficient, r had been scaled from -1 to 1 depending on how the strength of the relationship between two variables (Gingrich, 2004). In this study, the correlation of O_3 with the temperature, wind speed, relative humidity, NO, NO_2, and CO was determined and as result, the strong correlation parameter with O_3 had been selected to use as an input in MLR, MLR_2 method.

Principal Component Analysis (PCA) and Regression (PCR)

PCA is the method that had been widely used to reduce the multicollinearity problem between the independent parameters in the regression analysis. The parameters were grouped in the principal components (PCs) based on the eigenvalues that greater than 1 (Roy and Ambure, 2016). Afterwards, those selected PCs will be used as MLR input in MLR_3 method or also known as principal component regression (PCR). The varimax rotation had been applied in this study to ensure each original variable were maximally correlated in each PCs (Awang *et al*., 2015; Abdul- Wahab *et al*., 2005). The PCA equation is presented in equation 2.

$$PC_i = l_{1i}X_i + l_{2i}X_2 + \ldots + l_{ni}X_n \quad (2)$$

Multiple Linear Regression (MLR)

MLR was introduced in several years as a simple computed and implemented a technique to predict air pollution. It could predict the air pollution concentration in the next prediction time based on the current air pollutants data (Jumin *et al*., 2020). In this study, stepwise MLR with 95% confidence interval was applied to predict the next

hour of the O_3 concentration (dependent variables) based on the meteorological factors and its precursors (independent variables) (Abdullah *et al*., 2020;). The general MLR equation as illustrated in equation 3.

$$y = b_0 + \sum_{i=}^{n} b_i X_i + \varepsilon \quad (3)$$

Where the b_i as regression coefficient, which is independent variables and the ε is the stochastic error that correlates with the regression. The residuals in the MLR model were assumed that it's been having normally distributed residual with zero means, uncorrelated and constant variance (Abdullah *et al*., 2018).

RESULT AND DISCUSSIONS

Diurnal trends of O_3 concentration, ambient temperature and relative humidity

The diurnal trend of O_3 concentration, ambient temperature and relative humidity fluctuated in four years of data at Terengganu's urban areas. When the temperature is higher and the relative humidity was lower especially during the afternoon, the O_3 concentration was increasing as displayed in Figure-2. The O_3concentrationwas decreasing start from the evening until morning as the temperature is lower and the relative humidity were higher. It had been proven that the increase of O_3 concentration was having a strong connection with the meteorological factors (Quansah *et al*., 2012). Warm and humid provides ideal environmental condition to promote the more frequent photochemical reaction of O_3 precursor to occur and resulting in higher O_3 concentration in the atmosphere (Lu *et al*., 2019; Izzah *et al*., 2019; Chaiyakhan *et al*., 2017). The summary of the descriptive data was tabulated in Table-1. The average daily maximum O_3 concentration is 0.020ppm (0.00-0.090 ppm) which still within the Malaysia New Ambient Air Quality Standard. The average daily maximum temperature and relative humidity were 27.10^{O}C (20.00-39.30^{O}C) and 82.53% (36.00-100.00 %), respectively. However, the emission of other gaseous pollutants from motor vehicles also need to consider as it also a huge contributor to high O_3 concentration in an urban area at Terengganu (Zhou *et al*., 2018; Bereitschaft & Debbage, 2013).

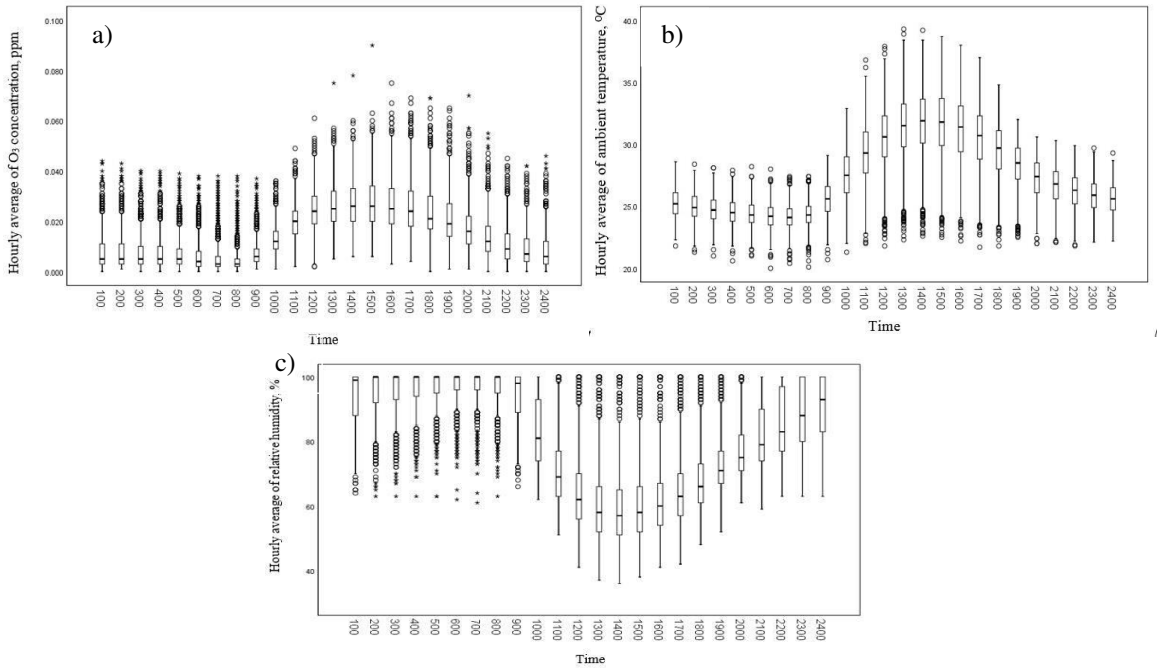

Figure-2. Hourly average trend of a) O_3 concentration (ppm) b) ambient temperature, ^{O}C c) relative humidity.

Table-1. Summary of descriptive analysis of O_3 concentration, ambient temperature and relative humidity.

Parameters	Average Time	Minimum	Maximum	Mean	Median	Std. Dev	NMAAQS*
O_3, ppm	1h	0.00	0.09	0.02	0.012	0.011	0.20
Ambient Temperature, ^{O}C	1h	20.00	39.30	27.10	26.300	3.160	-
Relative Humidity, %	1h	36.00	100.00	82.53	86.000	16.812	-

*New Malaysia Ambient Air Quality Standards

Model Development

Three MLR models were established to predict the next hour $O_{3,\ t+1}$ concentration by using three different input variables and each model were namely as MLR_1, MLR_2, and MLR_3. The MLR_1 was developed based on the all variables input that used in this study. The variables were consisting of WS, T, RH, NO, NO_2 and CO. Meanwhile, the output from Spearman correlation and PCA analysis were used as input in MLR_2 and MLR_3, respectively. Table-2 shows that WS (r = 0.686), T (r = 0.750) and CO (r = 0.650) were having strongly positive relationship with the increase of O_3 while RH (r = -0.767) and NO (r = - 0.579) were showed strongly negative relationship with the O_3 concentration. NO_2 was having a moderate and positive correlation with the O_3 concentration in the atmosphere. As a result, the strong either positive or negative correlation was selected as an input in the MLR_2.

Table-2. Spearman correlation analysis.

	O_3	WS	T	RH	NO	NO2	CO
O_3	1	0.686**	0.750**	-0.767**	-0.579**	-0.320**	0.650**

** Correlation is significant at the 0.01 level

In MLR_3, the data was adequate and it fulfils the relevant factor analysis as the requirement for PCA by having 0.751 ($p > 0.05$) and 0.000 ($p < 0.001$) for the Kaiser-Meyer-Olkin (KMO) and Bartlett's test, respectively (Nazif *et al.,* 2018; Awang *et al*., 2015b). The multicollinearity problem can be reduced through PCA by generating PCs (Roy and Ambure, 2016). Hence, two PCs were generated from six parameters that undergo the PCA. The WS had been excluded from the PCA as it has the communalities values less than 0.5 for better analysis interpretation (Taherdoost *et al*., 2014). The PCs were selected based on the eigenvalues that greater than 1 (Ul-Saufie *et al.*, 2011). Before rotation and after extraction, the percentage variance of PC-1 and PC-2 are having values which about79.23% and 57.94% for PC-1, respectively. The summary of total variance explained was depicted in Table-3 together with the excluded component. The PC-1 consists of parameters from meteorological factor (RH and T) and O_3 itself while PC-2 has come from the O_3 precursors which were NO, NO_2 and CO (Izzah *et al*., 2019; Ul-saufie *et al.,* 2011). The varimax rotated component matrix were explained the parameters that had been grouped in each PC with the suppressed output values less than 0.5 as shown in Table-4.

Table-3. Total variance explained.

	Initial Eigen values			Extraction sums of squared loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.48	57.94	57.94	3.48	57.94	57.94	2.65	44.12	44.12
2	1.28	21.28	79.22	1.28	21.28	21.28	2.11	35.10	79.12
3	0.65	10.79	90.01						
4	0.27	4.54	94.55						
5	0.23	3.84	98.39						
6	0.10	1.61	100.00						
7	0.32	4.01	97.62						
8	0.19	2.38	100.00						

Table-4. Varimax rotated component matrix.

	Component	
	1	2
T	0.939	
RH	-0.936	
O3	0.846	
NO		0.7280
NO2		0.838
CO		0.855

In this study, the models were established by assuming that they do not have any first-order autocorrelation and multicollinearity problem based on Durbin - Watson (D-W) values is within 2 and Variance Inflation Factor (VIF)within 10 (Ul-Saufie *et al.,* 2011). The D-W in this study were MLR_1(1.440), MLR_2(0.510) and MLR_3(0.734). Meanwhile, for VIF range value the MLR_1, MLR_2 and MLR_3 were obtained 1.728-6.037, 1.295-5.502, and 1.000, respectively. MLR_1 was obtained the highest coefficient determination (R2), 0.799 compared to model MLR_2 (0.552) and MLR_3 (0.667).

All the models were summarized in Table-5. The influence variables in MLR_1 were consist of T, RH, NO, CO_2, NO_2 and O_3. The $O_{3,\ t+1}$ concentration was increasing about 0.829, 0.140, 0.265 and 0.039 units when the one unit of O_3, T, NO, RH were increased. However, it was decreased about 0.020 and 0.094 units as the one unit of CO and NO_2 decreased. MLR_2, the $O_{3,\ t+1}$ concentration was increasing 0.318 and decreasing about 0.259 and 0.147 units when the one unit of T was increasing, and one unit of CO and RH were decreasing. However, MLR_3 was contained only two PCs which grouped based on the meteorological factors and O_3 precursor. The $O_{3,\ t+1}$ concentration was increased and decrease about 0.102 and 0.033 when the increase and decrease of one unit in PC1 and PC2, respectively. Overall, these three models were showed that the temperature, relative humidity and the CO were played an important role in increasing the O_3 in the atmosphere (Liu *et al.*, 2019). Therefore, CO generated through the incomplete combustion process in the motor vehicle. It's one of the O_3 precursors which produce O_3 concentration through the photochemical process (Teixeira *et al.*, 2009).

Table-5. The summary of the three different MLR models.

Method	Model	Remarks
MLR_1	$O_{3,\ t+1}$ = 0.829 O_3 + 0.140 T + 0.165 NO + 0.039 RH - 0.020 CO - 0.094 NO2 - 0.040	
MLR_2	$O_{3,\ t+1}$ = 0.195 + 0.318 T - 0.159 CO - 0.147 RH	
MLR_3	$O_{3,\ t+1}$ = 0.173 + 0.102 PC1 - 0.033 PC2	PC1 = 0.939 T – 0.936 RH + 0.846 O_3 PC2 = 0.728 NO + 0.838 NO_2 + 0.855 CO

The predicted against observed O_3 concentration graph was plotted for all models as illustrated in Figure-3. All the models were normally distributed. However, the R^2 has influenced the direction of the residuals which it has negatively skewed. Figure-4 were plotted between the fitted values against the prediction values of $O_{3,\ t+1}$ model's residual for all models. It showed the residuals were uncorrelated as it's accumulated around the horizontal band with the constant variance with mean equal to zero.

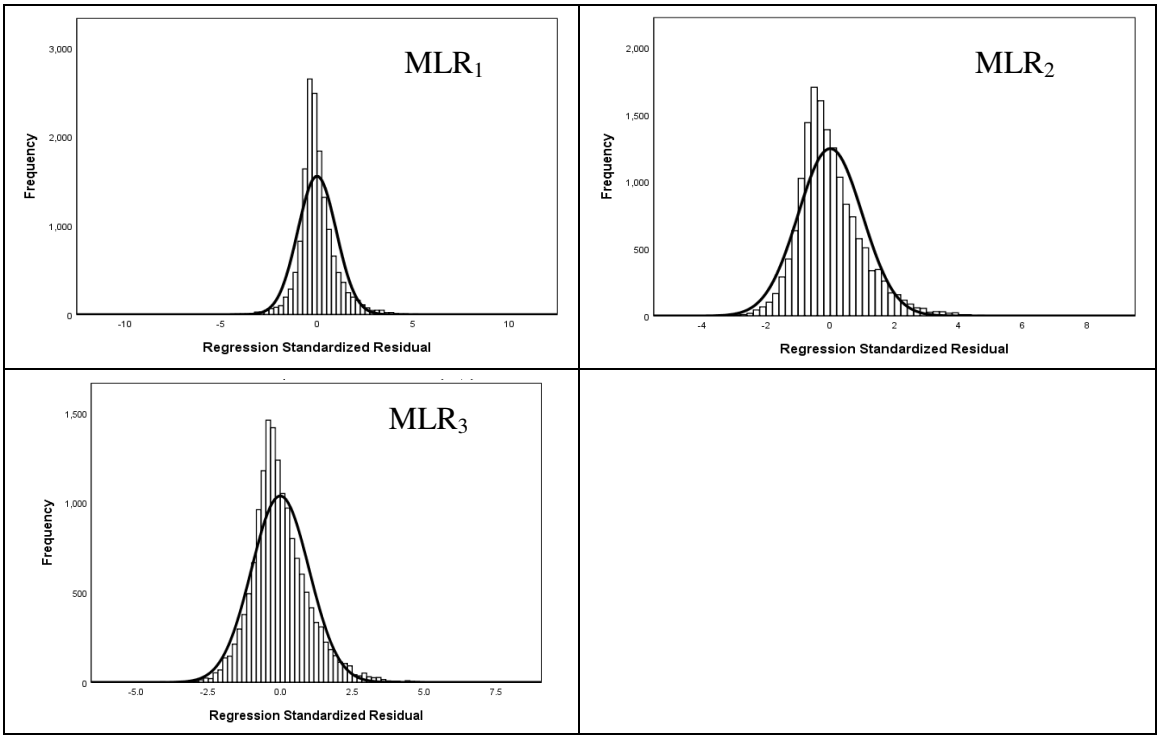

Figure-3. Standardized residual analysis of MLR_1, MLR_2, and MLR_3 model.

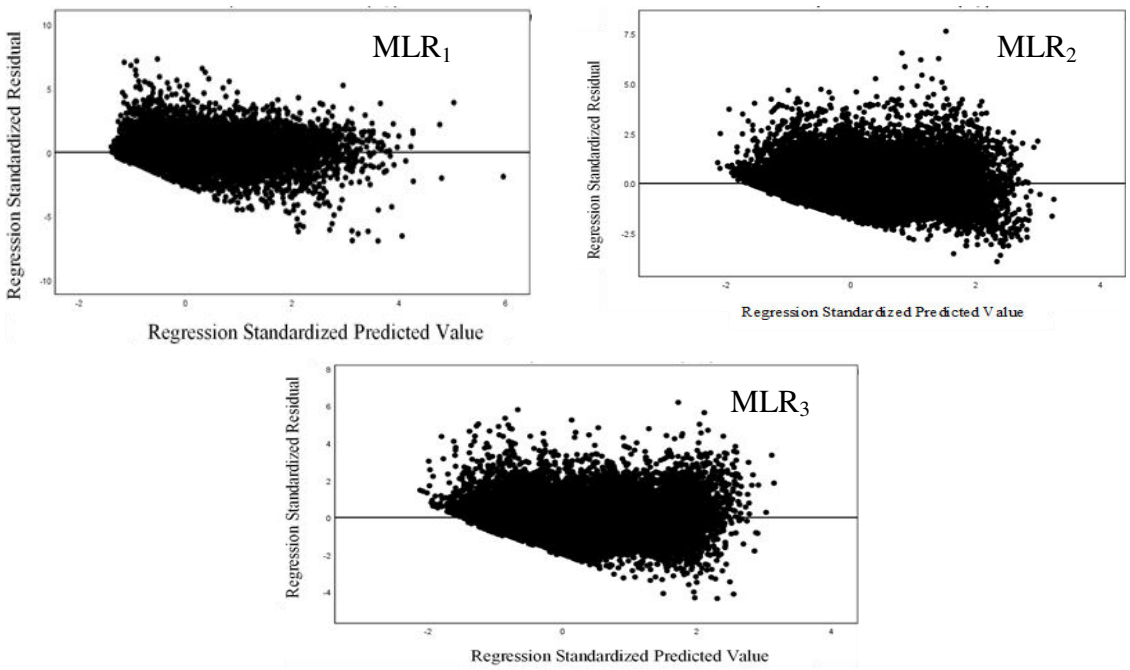

Figure-4. Assumptionof variance and uncorrelated residuals of MLR_1, MLR_2, and MLR_3 model.

Model Validation

Thirty percent of the data set was for model's validation. As a result, MLR_3 was having the highest correlation coefficient, R^2 (0.411) compare to MLR_1 (0.382) and MLR_2 (0.399). Hence, the MLR_3 was selected as the goodness-fit model to predict the $O_{3,\ t+1}$ concentration in an urban area at Terengganu. The graph predicted O_3 concentration, ppm versus observed O_3 concentration; ppm was plotted in Figure-5. Therefore, most of the points were accumulated within a 95% confidence interval. The upper and lower lines of 95% confidence interval limit were drawn as line A and line C.

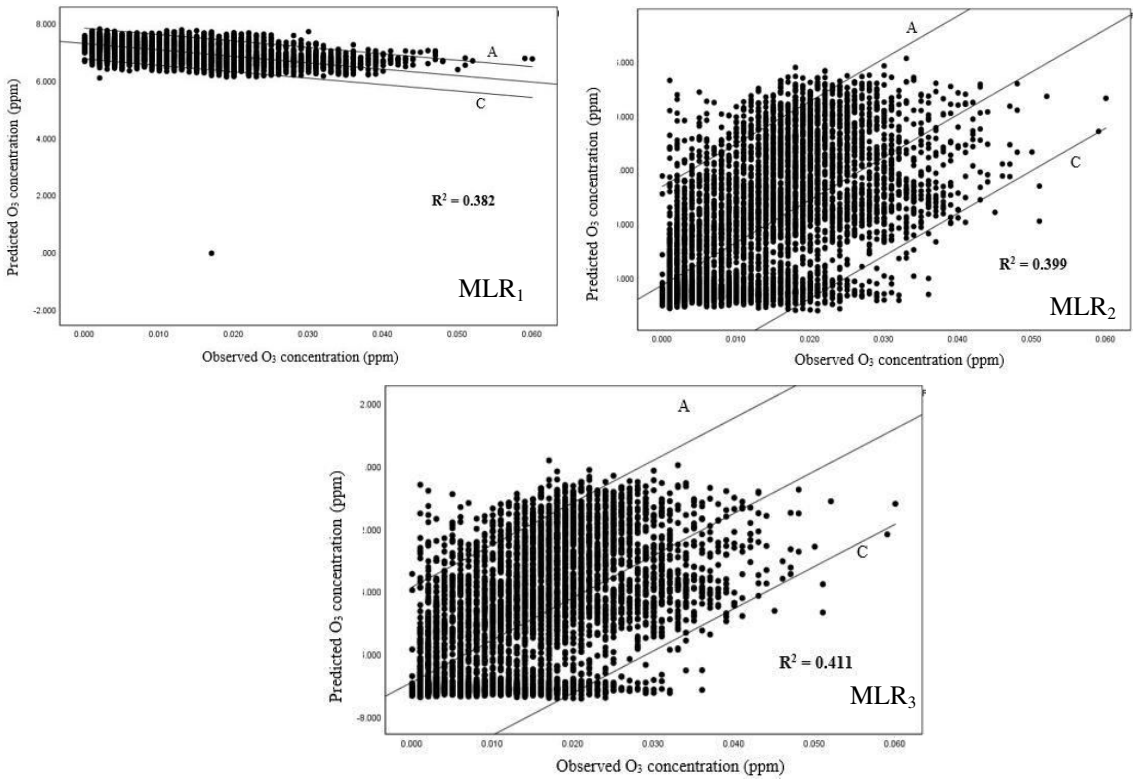

Figure-5. The predicted O_3 concentration (ppm) against observed O_3 concentration (ppm) in Terengganu's urban area.

CONCLUSIONS

As a conclusion, the diurnal trend of O_3, temperature and relative humidity fluctuated, and it's were correlated among each other. MLR_3 was selected as the best prediction models to predict the O_3 concentration in Terengganu's urban area by having the highest correlation coefficient, R^2 (0.411) compared to the other two models.

ACKNOWLEDGEMENT

This study is funded by the Fundamental Research Grant Scheme for Research Acculturation of Early Career Researchers by the Malaysian Ministry of Education (RACER/1/2019/TK10/UMT//1). We thanks to the Talent and Innovation Center, University Malaysia Terengganu for the funding. We also would like to thank the Air Quality Division, Malaysian Department of Environment for the acquisition of air quality data.

REFERENCES

Abdul-Wahab S. A., Bakheit C. S. & Al-Alawi S. M. 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environmental Modelling and Software, 20(10): 1263-1271. https://doi.org/10.1016/j.envsoft.2004.09.001

Abdullah S., Ismail M., Samat N. N. A. & Ahmed A. N. 2018. Modelling particulate matter (PM10) concentration in industrialized area: A comparative study of linear and nonlinear algorithms. ARPN Journal of Engineering and Applied Sciences. 13(20): 8227-8235.

Alin A. 2010. Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3): 370-374. https://doi.org/10.1002/wics.84

Awang N. R., Ramli N. A., Yahaya A. S. & Elbayoumi M. 2015a. Multivariate methods to predict ground level ozone during daytime, nighttime, and critical conversion time in urban areas. Atmospheric Pollution Research, 6(5): 726-734. https://doi.org/10.5094/APR.2015.081

Banan N., Latif M. T., Juneng L. & Ahamad F. 2013. Characteristics of surface ozone concentrations at stations with different backgrounds in the Malaysian Peninsula. Aerosol and Air Quality Research, 13(3): 1090-1106. https://doi.org/10.4209/aaqr.2012.09.0259

Bereitschaft B. & Debbage K. 2013. Urban form, air pollution, and CO2 emissions in large U.S. metropolitan areas. Professional Geographer, 65(4): 612-635. https://doi.org/10.1080/00330124.2013.799991

Borkowf C. B. 2000. New nonparametric method for variance estimation and confidence interval construction for Spearman's rank correlation. Computational Statistics and Data Analysis, 34(2): 219-241. https://doi.org/10.1016/S0167-9473(99)00077-8

Capilla C. 2016. Prediction of hourly ozone concentrations with multiple regression and multilayer perceptron models. International Journal of Sustainable Development and Planning, 11(4): 558-565. https://doi.org/10.2495/SDP-V11-N4-558-565

Chaiyakhan K., Chujai P., Kerdprasop N. & Kerdprasop K. 2017. Hourly Ground-level Ozone Concentration Prediction using Support Vector Regression. Lecture Notes in Engineering and Computer Science, 2227, 306-311.

Cho H. S. & Choi M. J. 2014. Effects of compact urban development on air pollution: Empirical evidence from Korea. Sustainability (Switzerland), 6(9): 5968-5982. https://doi.org/10.3390/su6095968

Ghorani-Azam A., Riahi-Zanjani B. & Balali-Mood M. 2016. Effects of air pollution on human health and practical measures for prevention in Iran. Journal of Research in Medical Sciences, 21(5), https://doi.org/10.4103/1735-1995.189646

Gingrich P. 2004. Chapter 11 Association between Variables. Introductory Statistics for the Social Sciences - Http://Uregina.ca/~gingrich/Text.Htm, 794-835. Retrieved from http://uregina.ca/~gingrich/text.htm

Ismail M., Suroto A. & Abdullah S. 2015. Response of Malaysian Local Rice Cultivars Induced by Elevated Ozone Stress. Environment Asia, 8(1). DOI: 10.14456/ea.2015.11

Izzah Mohd Hashim N., Alia Izzati Mohd Yusoff N., Mohamed Noor N. & Zia Ul-Saufie 2019. Assessment of Surface Ozone Concentration in Northern Peninsular Malaysia. IOP Conference Series: Materials Science and Engineering, 551(1). https://doi.org/10.1088/1757-899X/551/1/012100

Jumin E., Zaini N., Ahmed A. N., Abdullah S., Ismail M., Sherif M., … El-Shafie A. 2020. Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction. Engineering Applications of Computational Fluid Mechanics, 14(1): 713-725. https://doi.org/10.1080/19942060.2020.1758792

Kang H. 2013. The prevention and handling of the missing data. Korean Journal of Anesthesiology, 64(5): 402-406. https://doi.org/10.4097/kjae.2013.64.5.402

Liu X., Tan W. & Tang S. 2019. A Bagging-GBDT ensemble learning model for city air pollutant concentration prediction. IOP Conference Series: Earth and Environmental Science, 237(2): 0-7. https://doi.org/10.1088/1755-1315/237/2/022027

Lu X., Zhang L. & Shen L. 2019. Meteorology and Climate Influences on Tropospheric Ozone: a Review of Natural Sources, Chemistry and Transport Patterns.

Current Pollution Reports, (1). https://doi.org/10.1007/s40726-019-00118-3

Nazif A., Mohammed N. I., Malakahmad A. & Abualqumboz M. S. 2018. Regression and multivariate models for predicting particulate matter concentration level. Environmental Science and Pollution Research, 25(1): 283-289. https://doi.org/10.1007/s11356-017-0407-2

Quansah E., Amekudzi L. K. & Preko K. 2012. The Influence of Temperature and Relative Humidity on Indoor Ozone Concentrations during the Harmattan Corresponding Author: Quansah E. Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS), 3(5): 863-867.

Roy K. & Ambure P. 2016. The double cross-validation software tool for MLR QSAR model development. Chemometrics and Intelligent Laboratory Systems, 159(October): 108-126. https://doi.org/10.1016/j.chemolab.2016.10.009

Taherdoost H., Sahibuddin S. & Jalaliyoon N. 2014. Exploratory factor analysis: Concepts and theory. 2nd International Conference on Mathematical, Computational and Statistical Sciences. 375-382.

Teixeira E. C., de Santana E. R., Wiegand F. & Fachel J. 2009. Measurement of surface ozone and its precursors in an urban area in South Brazil. Atmospheric Environment, 43(13): 2213-2220. https://doi.org/10.1016/j.atmosenv.2008.12.051

Ul-Saufie A. Z., Yahya A. S. & Ramli N. A. 2011. Improving multiple linear regression model using principal component analysis for predicting PM_{10} concentration in Seberang Prai , Pulau Pinang. International Journal of Environmental Science, 2(2): 403-410. https://doi.org/10.6088/ijes.00202020003

Wang Z., Lv J., Tan Y., Guo M., Gu Y., Xu S. & Zhou Y. 2019. Temporospatial variations and Spearman correlation analysis of ozone concentrations to nitrogen dioxide, sulfur dioxide, particulate matters and carbon monoxide in ambient air, China. Atmospheric Pollution Research, 10(4): 1203-1210. https://doi.org/10.1016/j.apr.2019.02.003

Zainordin N. S., Ramli N. A. & Elbayoumi M. 2017. Distribution and temporal behaviour of O_3 and NO_2 near selected schools in Seberang Perai, Pulau Pinang and Parit Buntar, Perak, Malaysia. Sains Malaysiana, 46(2): 197-207. https://doi.org/10.17576/jsm-2017-4602-03

Zhou C., Li S. & Wang S. 2018. Examining the impacts of urban form on air pollution in developing countries: A case study of China's megacities. International Journal of Environmental Research and Public Health, 15(8). https://doi.org/10.3390/ijerph15081565