© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

Different Approaches of Multiple Linear Regression (MLR) Model in Predicting Ozone (O_3) Concentration in Industrial Area

Nur Nazmi Liyana Mohd Napi[1], Samsuri Abdullah[1*], Amalina Abu Mansor[2], Nurul Adyani Ghazali[1], Ali Najah Ahmed[3], Nazri Che Dom[4], Marzuki Ismail[2]

[1]Air Quality and Environment Research Group, Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, MALAYSIA

[2]Faculty of Science and Marine Environment, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, MALAYSIA

[3]Institute of Energy Infrastructure (IEI), Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor, MALAYSIA

[4]Centre of Environmental Health and Safety, Faculty of Health Sciences, Universiti Teknologi MARA, 42300 Puncak Alam, Selangor, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2023.15.01.010
Received 3 July 2021; Accepted 26 October 2022; Available online 28 February 2023

Abstract: Meteorological conditions and other gaseous pollutants generally impacted the development of ozone (O_3) in the atmosphere. The purpose of this study was to create the best O_3 model for forecasting O_3 concentrations in the industrial area and to determine the variables that affect O_3 concentrations. Five-year data of meteorological and gaseous pollutants were used to analyze and develop the prediction model. Based on three distinct techniques, three separate multiple linear regression (MLR) prediction models of O_3 concentration were developed. MLR_3 had the highest correlation coefficient of 0.792 during development as compared to models MLR_1 and MLR_2. MLR_2 was deemed the best O_3 prediction model, however, since it had the lowest error values of root mean square error (3.976) and mean absolute error (3.548) when compared to other models. The establishment of an O_3 prediction model can offer local governments with early information that could help them reduce and manage air pollution emissions.

Keywords: Ozone, meteorological, gaseous pollutant, multiple linear regression, industrial

1. Introduction

The expansion in the industrial sector, such as in the production of petrochemical, polymer, steels, and manufacturing in East Coast Peninsular Malaysia, has become one factor that declines the air quality in Terengganu [1], [2]. These industrial activities emit an uncontrolled amount of air pollutants into the atmosphere. The rapid growth in industrial area can increase the traffic that brings consequence to air pollutant emission into the atmosphere [3]. The emission of air pollutants is 20% to 25% of air pollutants from the industrial sector, such as through power plants, combustion activities, refineries, and chemical reactions. Meanwhile, 70% to 75% of air pollutants came from mobile sources due to high volume traffic during peak hours due to the incomplete combustion process [4]. The World Health Organisation [5] has included the ground-level ozone (O_3) in six criteria pollutants, which have high-risk potential to human health. In 2020, the Department of Environment (DOE) Malaysia [6] had set a permissible limit of ground-level O_3 concentration in the

**Corresponding author: samsuri@umt.edu.my*

New Malaysia Ambient Air Quality Standard (NMAAQS) in which cannot exceed 0.18 ppm for an hour and 0.1 ppm for 8 hours.

Tropospheric ozone, commonly known as ground-level O_3, is a highly phytotoxic pollutant. Photochemical and oxidation reactions in the presence of sunlight and its precursors, such as nitrogen oxide (NO_x) and volatile organic compounds (VOCs), produce O_3, which is released into the atmosphere by both human and natural processes [7]. The interaction between NO_x and VOCs under sunlight resulted in O_3 secondary pollutant. The burning of fossil fuels has become the primary source of elevated O_3 levels in the atmosphere [8]. O_3 is formed when oxygen (O_2) is split up by ultraviolet radiation, creating an oxygen atom. The unstable and highly reactive oxygen atom will bind with O_2 molecules and form O_3 molecules [9]. Therefore, meteorological factors had become the primary influence contributing to high O_3 concentration in the atmosphere. Favorable meteorological circumstances, such as high ambient temperature and low relative humidity, accelerate the photochemical process of O_3 precursor, resulting in a high concentration of O_3 in the atmosphere [9], [10]. Wind helps O_3 to travel a hundred miles and thus affecting areas downwind. Stagnant wind conditions cause a high concentration of air pollution [11]. The El Nino event and the southwest monsoon (SWM) in Malaysia had a significant effect on the production of O_3 concentrations in the atmosphere [10], [12].

Uncontrolled O_3 emission in the atmosphere raises the risk-potential to human health, ecology, and environment. The massive amount of O_3 concentration in the atmosphere can have long-term and short-term effects, especially for certain human groups such as sensitive people like children and the elderly [13], [14]. Therefore, a high O_3 concentration can cause serious illnesses such as problems in the cardiovascular system, respiratory system, cancer, and even mortality when people are exposed to it for the long term [13], [15], [16]. Furthermore, a significant high O_3 concentration in the atmosphere can interrupt the ecosystem by inhibiting plant growth, thus resulting in the forest's abnormal development [17]. It also disrupts symbiotic relationship, regular plant-parasite interaction and increases species extinction, which can cause ecosystem functions impairment [18]. The O_3 concentration increase in the atmosphere will also increase the atmosphere temperature, resulting in global warming and climate change phenomena [19].

The multiple linear regression (MLR) model applied globally to forecast air pollution, as it can be computed and implemented efficiently [20], [21], [22]. The goal of the study was to develop the best MLR prediction model of O_3 concentration in the industrial area using three different MLR methodologies. We developed three methods; (1) Method 1 (MLR_1) in which parameters that have a strong correlation with O_3 concentration were used as input in MLR; (2) Method 2 (MLR_2) in which principal component analysis (PCA) output was considered as input in MLR; (3) Method 3 (MLR_3) in which all meteorological factors and gaseous pollutants were used as inputs to develop MLR forecasting model. The best prediction model of O_3 can help provide early information to local authorities for planning some mitigation strategies to decrease air pollution levels and improve air quality.

2. Materials and Methods

2.1 Study Area and Data Acquisition

Kemaman is on the East Coast of Peninsular Malaysia, facing the South China Sea with a total area of 2,535.60 km^2 and the estimated total population is about 201,100 in 2014. The Kemaman Municipal Council administers it, and the urban centre is located at Chukai [23]. It is also known as the second-largest city in Terengganu. The industrial areas in Terengganu consist of heavy industrial activities, such as the production of petrochemical, steel production, polymer, and manufacturing. The air quality monitoring station (AQMS) was installed by DOE Malaysia at Bukit Kuang, Teluk Kalong Primary School, Kemaman (N04º 16.260’: E103º 25.826’). It is near the city centre and industrial area with heavy traffic, especially during the peak hours (Fig1). Malaysian Department of Environment provided air pollutants and meteorological data from January 1, 2010 to December 31, 2014. The parameters including ozone (O_3, ppm), nitrogen oxide (NO, ppm), nitrogen dioxide (NO_2, ppm), carbon monoxide (CO, ppm), sulphur dioxide (SO_2, ppm), wind speed (WS, km/hr), ambient temperature (T,^{0}C), and relative humidity (RH, %). The data is tabulated and organised in Microsoft Excel Spreadsheet® 2016 before being analysed with Statistical Packages for the Social Sciences (SPSS®) Version 25. DOE Malaysia has entrusted Alam Sekitar Malaysia Sdn Bhd (ASMA) with the installation, operation, and maintenance of air pollution monitoring instruments and data [24].

DOE used Teledyne API Model 400/400E instrument through the ultraviolet (UV) absorption (Beer-Lambert) method with a 0.4 ppb detection limit and using the 0.5% of the precision level to measure the O_3 concentration hourly [25]. Model 200A measured the continuous monitoring of NO_2 and NO concentrations in the ambient air NO/NO_2/NO_x analyser by having chemiluminescence detection principles, as it provides sensible, stable, and easy usages [25]. The Teledyne API Model 100A/100E was used to measure SO_2 concentration by the lowest detection at 0.04ppb using UV fluorescence. The Teledyne API Model 300/300E with 0.5% precision and 0.04 ppm of the most insufficient detection using non-dispersive and infrared absorption (Beer Lambert) used to monitor and measure CO concentrations [25]. The meteorological parameters, ambient temperature, relative humidity, and wind speed were measured by Met One 062, Met One 083D, and Met One 010C sensor, respectively [25].

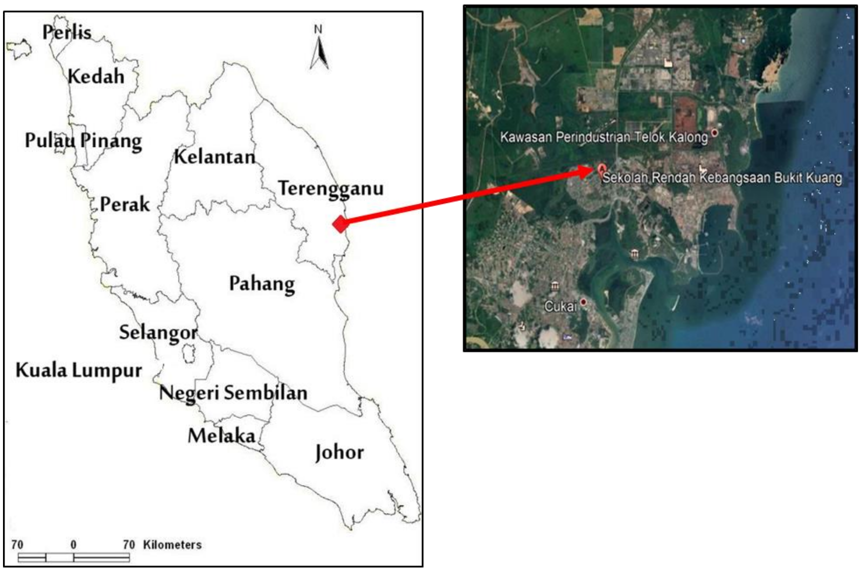

Fig. 1 - Location of Air Quality Monitoring Station (AQMS) at the Kemaman industrial area

Daily calibration with zero air and standard gas concentrations is used for quality control and data assurance. Monitored data are checked before being transfer to DOE [26]. Due to calibration and technical problems, missing data were deleted to produce unbiased prediction and conservative results [27]. We did statistical descriptives analysis and percentile plot to investigate the trends of O_3 concentration for five years—Normalizing data set by min-max techniques, ranging from 0 to 1. The normalization process was able to reduce biases in the analysis, as the parameters in the data set consisted of different types of International System of Units (SI) [24], [28]. Equation 1 shows the min-max normalization technique.

$$z_i = \frac{x_1 - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where, $x = (x_1 \ldots, x_n)$ and z_i is a normalized data.

2.2 Model Development and Validation

PCA is a statistical method to determine and classify variables based on their correlation coefficient in principal components (PCs). 70 per cent of the data set in this study uses as input in PCA, in which the PCA divided and grouped into PCs, which will operate as input in regression analysis [29]. Varimax rotation applies to specify the PCs based on greater than 1 values of the eigenvalues. The Kaiser-Meyer-Olkin (KMO) test is essential in this analysis because it measures sampling adequacy with $p > 0.50$. In contrast, Bartlett's test of sphericity uses to test factor analysis appropriation between correlation and variables with $p < 0.001$ [30], [31]. The advantage of this analysis is that it can reduce multicollinearity problem and ensure that a maximal variance of linear combination is chosen [29]. Equation 2 shows the equation for PCA [30].

$$PC_{ij} = l_{1i}X_{1j} + l_{2i}X_{2j} + \ldots + l_{ni}X_{nj} \tag{2}$$

Where PC is component score, l is component loading, X is the measured value of the variable, i is the component number, j is the sample number, and n is the total number of variables.

MLR is an established model that may connect two or more independent variables to one dependent variable. The stepwise MLR models in this analysis were developed using a 95 percent confidence interval. Dataset divided in respect of 7:3 ratio for model development and data validation [32]. The residuals assumed normally distribute by having zero mean, uncorrelated and constant variances [23]. The MLR equation shown in Equation 3.

$$y = b_0 + \sum_{i=1}^{n} b_i X_i + \varepsilon \tag{3}$$

b_i is the regression coefficient (X_i is independent variables), and ε is a stochastic error associated with the regression.

VIF measures the multicollinearity problem between the predictors (independent variables) in the regression model. VIF values are below ten show there is no multicollinearity problem between the independent variables [29]. The VIF equation presents in Equation 4.

$$VIF_i = \frac{1}{1-R_i^2} \tag{4}$$

Where,
VIF_i is the variance inflation factor with ith predictors
R_i^2 is the determination in the regression of the ith predictor on all other predictors

D-W test used to detect the autocorrelation in residuals from regression analysis. It could predict the O_3 in the following hours or next days based on the O_3 concentration in the current day. The test range values are between 0 to 4, showing no first-order autocorrelation as the residuals are uncorrelated for an evaluated value of 2 [29]. The equation of DW was present in Equation 5.

$$DW = \frac{\sum_{i=1}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2} \tag{5}$$

Where,
n = observations number
$e_i = y - y_i$ (y = observed values and y_i is the predicted values).

R^2 is used to establish if the data provide sufficient evidence to represent the entire model that contains information about the O_3 concentration prediction model or vice versa. It also had been used as an indicator to select the best-fitted prediction models [29]. The R^2 equation illustrated in Equation 6.

$$R^2 = (\frac{\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})}{n.S_{pred}.S_{obs}})^2 \tag{6}$$

Where, n = total number of measurements at a particular site, P_i = predicted values, O_i = observed values, $\bar{P}$ = mean of predicted values, $\bar{O}$= mean of observed values, S_{pred} = standard deviation of predicted values, and S_{obs} = standard deviation of the observed values.

The best-fit model is determined by the model's performance indicator of error and accuracy measurements. The error measures consisted of root mean square error (RMSE) and mean absolute error (MAE), while the determination of coefficient, R^2 used as an accuracy measure. The model that is having higher accuracy measure (the values are close to 1) and lower error values (relative to 0) considered as the best prediction model of O_3 concentration in the industrial area [24], [33]—the performance indicators displayed in Equation 7 to Equation 9.

a) Root Mean Square Error

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}[P_i - O_i]^2\right)^{1/2} \tag{7}$$

b) Mean Absolute Error

$$MAE = \frac{\sum_{i=1}^{n}|O_i - P_i|}{n} \tag{8}$$

c) Correlation Coefficient

$$R^2 = \left(\frac{\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})}{n.S_{pred}.S_{obs}}\right)^2 \tag{9}$$

3. Results and Discussion

The yearly trend in O_3 concentrations during a five-year period of data from an industrial location on Peninsular Malaysia's east coast has changed throughout the years [34]. Fig2 illustrates the fluctuated trends of O_3 concentration according to the percentiles 0%, 25%, 50%, 75%, and 100%. The highest maximum O_3 concentration recorded in 2013 with 0.098 ppm, and the lowest minimum O_3 concentration of 0.000 ppm recorded in 2010, 2011, and 2014. The highest mean value of O_3 concentration was 0.023 ppm (0.001 – 0.083 ppm), which was recorded in 2012, while the lowest mean value of O_3 concentration was 0.018 ppm (0.000–0.081 ppm) in 2010—the summary of five-year O_3 descriptive data displayed in Table 1. Therefore, the O_3 concentration in the industrial areas at Terengganu was still within the NMAAQS limit, which was 0.18 ppm for one-hour exposure, while 0.100 ppm for 8-hour exposure [6]. However, other industrial areas in Malaysia, such as Shah Alam, recorded a maximum O_3 concentration of about 0.174 ppm during the daytime, 0.089 ppm during night-time, and 0.113 ppm during critical conversion time [33]. The highest maximum O_3 concentration with values 0.124 ppm, 0.105 ppm, and 0.091 ppm recorded at Klang, Perai, and Pasir Gudang industrial areas, respectively, based on the 2009 data [25]. The reaction of O_3 precursor with sunlight in the daytime promotes photochemical occurrence, which is an essential factor in increasing O_3 concentration in the ambient air [10]. The O_3 precursor consists of nitrogen oxide, and VOCs emitted from industrial activities and motor vehicles [10], [35]. The presence of oxidant radicals (hydroperoxyl radicals, HO_2), organic peroxy radicals (RO_2), hydrocarbon and alkoxy radicals (RO) increase O_3 production by converting NO to NO_2 [19].

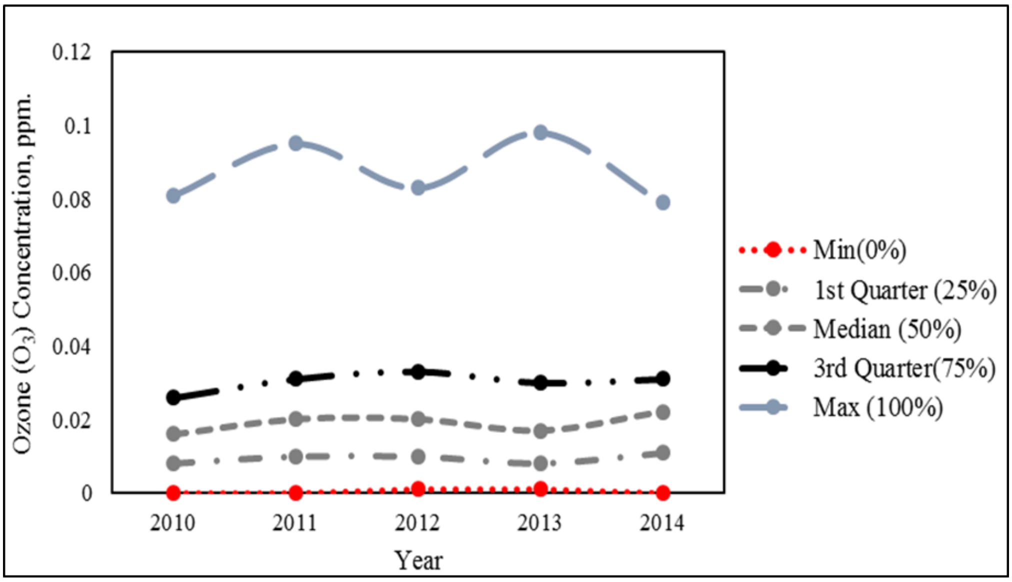

Fig. 2 - Annual trend of ozone (O_3) concentration from the year 2010 to 2014

Table 1 - Summary of descriptive analysis of O_3 concentration from the year 2010 to 2014

Descriptive Statistics	**2010 (N=5847)**	**2011 (N=4481)**	**2012 (N=5133)**	**2013 (N=5492)**	**2014 (N=3232)**
Mean (ppm)	0.018	0.022	0.023	0.021	0.022
Median (ppm)	0.016	0.020	0.020	0.017	0.022
Maximum (ppm)	0.081	0.095	0.083	0.098	0.079
Minimum (ppm)	0.000	0.000	0.001	0.001	0.000
Std Dev (ppm)	0.013	0.015	0.015	0.015	0.137

Spearman correlation connects the relationship between O_3, meteorological factors and O_3 precursors. Table 2 tabulated the correlation analysis. With coefficient values of r = 0.788, $p < 0.01$, and r = 0.702, $p < 0.01$, respectively, WS and T displayed a substantial positive relationship with O_3 concentration. The concentration of O_3 in the atmosphere displayed a significant negative correlation with RH. Other gaseous pollutants including NO (r = 0.788, p 0.01), SO_2 (r = 0.702, $p < 0.01$), NO_2 (r = 0.215, $p < 0.01$), and CO (r = 0.123, $p < 0.01$) displayed a weak positive relationship with the rise in O_3 concentration in Terengganu's industrial area. The mixed findings on correlational relationship of the bivariate parameters are slightly due to the different site characteristics, terrain, emission factors, and other related factors that influence the dispersion and fate of air pollutants in the atmosphere. The meteorological factors such as WS, T, and RH played huge roles in influencing the O_3 concentration in certain areas. The increase of WS helped in increasing the dispersion of O_3 and its precursor in the atmosphere by reducing the stability of the boundary layer and then transporting it from the surface layer to the upper layer [35], [36], [37]. The higher ambient temperature, T, and lower RH, which provided warm and dry conditions, promoted and speeded up the photochemical reaction and oxidation rates between O_3 itself and its precursor to produce a high O_3 concentration in the ambient air [35], [38]. Human activities emit the other gaseous pollutant parameters, such as open burning, emissions from industries, and motor vehicle emissions. SO_2 and CO usually produced from industrial emission and signified as the industrial emissions indicators that contribute to the

O_3 formation [35]. NO_x commonly generated through anthropogenic activities in which converted to NO and NO_2 through a chemical reaction which plays the main role in O_3 formation in the atmosphere [36], [38].

Table 2 - Summary of spearman bivariate correlation between O_3 concentration with meteorological factor and other gaseous pollutants

	O_3	WS	T	RH	NO	SO_2	NO_2	CO
O_3	1	0.788**	0.702**	-0.523**	0.141**	0.344**	0.215**	0.123**
WS		1	0.649**	-0.542**	0.266**	0.273**	0.084**	-0.006
T			1	-0.559**	0.242**	0.385**	0.245**	0.011
RH				1	-0.179**	-0.254**	-0.099**	0.173**
NO					1	0.241**	0.329**	0.278**
SO_2						1	0.324**	0.148**
NO_2							1	0.403**
CO								1

*Note: ** Correlation is significant at the 0.01 level (2-tailed)*

KMO and Bartlett's test of sphericity is important in PCA to determine the adequacy and appropriate factor analysis of the data in this study. Table 3 displays the result of KMO and Bartlett's test of this study in which the KMO value is 0.729 greater than 0.05, while Bartlett's test value is 0.000 lower than 0.001. Therefore, this study is proven to have adequate data. It fulfilled the appropriate factor analysis as the requirement for PCA in which the KMO value was greater than 0.05, and Bartlett's test value was lower than 0.001 [29].

Table 3 - KMO and Bartlett's Test

Kaiser-Mayer-Olkin Measure of Sampling Adequacy		0.729
Bartlett's Test of Sphericity	Approx. Chi-Square	51357.564
	df	28
	Sig.	0.000

Table 4 lists the eigenvalues for each linear component (factor), as well as the values before, after, and after rotation. Before initiating the extraction process, eight parameters were chosen. Following the extraction, two components were chosen as PCs: those with higher eigenvalues than one and those with less eigenvalues than one [29]. The eigenvalues used to measure the amount of each component variance (percentage). These two factors accounted for 62% of the percentage reliability. The selected eigenvalues again displayed in the extraction sums of squared loadings and rotation sums of squared loadings. The rotation optimised the structure of the factor, which equalizing the two factors. The percentage of the variance before extraction showed that Factor 1 (38.59%) was higher than Factor 2 (23.11%), and it still had the same value after the extraction. However, some changes in the percentage of variance after the rotation in Factor 1 and Factor 2, with 38.13% and 23.84%, respectively. Table 5 shows the results using the varimax rotation with Kaizer normalisation. The matrix explained the parameters in each PC. With values less than 0.5, the output was suppressed by having either a positive or negative sign. Principal Component 1 (PC-1) can be concluded as a meteorological factor because it consisted of wind speed, temperature, and relative humidity. The major pollutant in industrial areas came from industrial and motor vehicles emissions that emitted gaseous pollutants, such as NO, NO_2, and CO. Therefore, gaseous pollutants are also known as O_3 precursors indicated in Principal Component 2 (PC-2) [33], [35].

Table 4 - Total variance explained

	Initial Eigenvalues			**Extraction sums of squared loadings**			**Rotation Sums of Squared Loadings**		
Component	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	3.109	38.858	38.858	3.109	38.858	38.858	3.050	38.130	38.130
2	1.849	23.107	61.965	1.849	23.107	61.965	1.907	23.835	61.965
3	0.842	10.522	72.487						
4	0.742	9.280	81.767						
5	0.546	6.826	88.593						
6	0.401	5.017	93.610						
7	0.321	4.008	97.617						
8	0.191	2.383	100.000						

Table 5 - Varimax rotated component matrix

	Component	
	1	**2**
WS	0.863	
T	0.866	
RH	-0.781	
O_3	0.863	
NO		0.691
NO_2		0.811
CO		0.801

MLR models were established based on three different inputs. The first method (MLR_1) used the output from strong correlation parameters with the O_3 concentration, based on the Spearman correlation analysis. The second method (MLR_2) used the input from generated PCs through PCA, whereby this method is also known as principal component regression (PCR). It is a combination model of PCA and MLR [20], [30]. Meanwhile, the third method (MLR_3) used all the meteorological and other gaseous pollutant parameters as their input in the MLR model's development. Table 6 summarized all the three different equations of the MLR models.

Table 6 - Summary of three MLR Models for forecasting O_3 concentration based on three different Inputs

Method	Model	Remarks
1	$O_{3,\ t+1} = 0.030 + 0.434\ (WS) + 0.252\ (T)$	-
2	$O_{3,\ t+1} = 0.218 + 0.120\ (PC\text{-}1) + 0.020\ (PC\text{-}2)$	$PC\text{-}1 = 0.863\ (WS) + 0.866\ (T) - 0.781\ (RH) + 0.863\ (O_3)$ $PC\text{-}2 = 0.691\ (NO) + 0.811\ (NO_2) + 0.801\ (CO)$
3	$O_{3,\ t+1} = 0.792\ (O_3) + 0.086\ (T) + 0.234\ (NO) + 0.071\ (CO) + 0.021\ (WS) + 0.015\ (RH) - 0.106\ (NO_2) - 0.015$	-

As a result, WS and T were the significant prediction variables factor to the O_3 concentration increase in the industrial areas for MLR_1. The prediction of $O_{3,\ t+1}$ concentration increased to 0.434 units when the WS was raised by one unit and 0.252 units increasing by one unit of T. For MLR_2, the $O_{3,\ t+1}$ increased to 0.120 and 0.020 units when one unit of PC-1 and PC-2 was increased. PC-1 consisted of the meteorological parameters WS, T, RH, and O_3, while PC-2 consisted of NO, NO_2, and CO. The O_3, T, NO, CO, WS, RH, and NO_2 were significant predictor variables for MLR_3. The $O_{3,\ t+1}$ increased to 0.792 units, 0.086 units, 0.234 units, 0.071 units, 0.021 units, and 0.015 units by one unit of O_3, T, NO, CO, WS, and RH, respectively and decreased to 0.106 units for one unit of NO_2.

We found that, during the model development, the MLR_3 had higher values of determination correlation, R^2 (0.792) as compared to MLR_1 (R^2, 0.571) and MLR_2 (R^2, 0.646). The R^2 value had influenced the normal distribution of residuals in which it was negatively skewed, as illustrated in Fig3. The VIF ranges for these three MLR_1, MLR_2, and MLR_3 were 1.766, 1.000, and 1.198–2.550, respectively. Each model showed that they do not have a multicollinearity problem between the independent variables because the VIFs values were below 10. However, MLR_2 had the lowest VIF value compared to MLR_1 and MLR_3, as it was a hybrid model PCA and MLR in which PCR minimised the multicollinearity problem among the independent variables [29], [30]. These three models were also not having any first-order autocorrelation problem as the D-W values were within 2, which were 0.627 (MLR_1), 0.749 (MLR_2), and 1.436 (MLR_3) [29]. The fitted values against the prediction of the $O_{3,\ t+1}$ model’s residual for the three models are plotted in Fig4 to show the uncorrelated residuals as the data around the horizontal band with the constant variance. \

The meteorological factors such as WS, T, and RH played a crucial component in forming, transportation, dispersion, and dilution of O_3 in the ambient air. The high concentration of O_3 was related to the presence of higher ambient temperature and lower relative humidity, which provided intense solar radiation and dry condition due to less amount of rainfall that caused the photochemical to happen frequently [36], [39]. A high wind speed can affect either O_3 concentration itself or other gaseous pollutants, its precursors by travelling a hundred miles from the original emission source and then forming and increasing the O_3 in other areas, especially downwind areas. A low wind speed tends to allow chemical reactions to happen, making air pollution more concentrated [11], [35]. NO_x commonly emitted through the combustion of fossil fuel and exhaust fumes. The reactive oxygen-containing molecules (RO_2) during photolysis and oxidation reaction in converting NO_2 to NO and oxygen atom help in the formation of O_3 with sunlight [36]. Additionally, CO gases are one of the air pollutants emitted through the emission of motor vehicles and industry. The present hydroperoxyl radical (HO_2) during the oxidation reaction of CO also helps in contributing to the formation of O_3 in the atmosphere [40], [41].

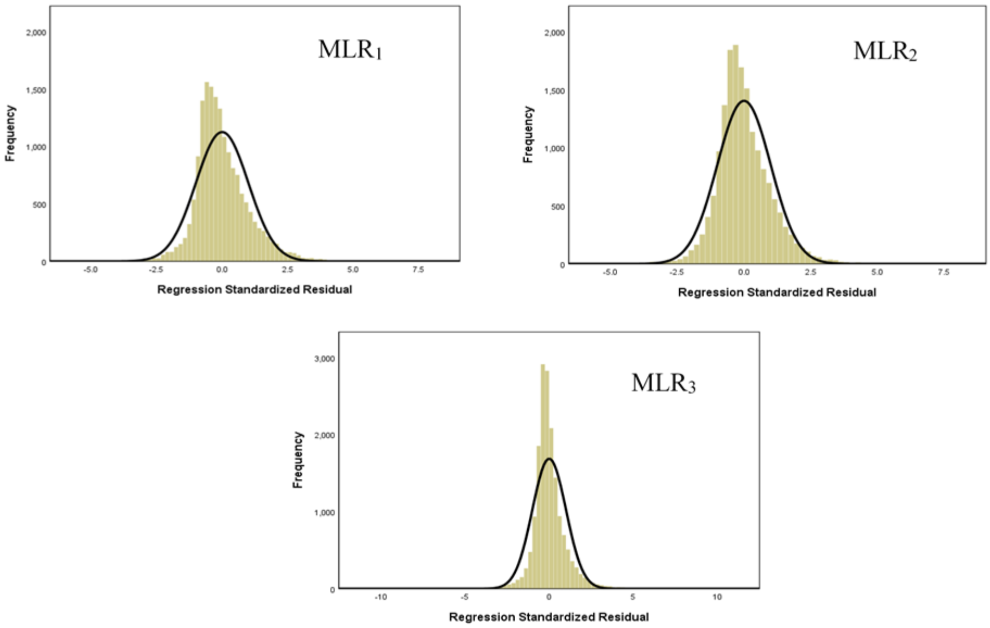

Fig. 3 - Standardized residual analysis of $O_{3,\,t+1}$ in all the three method

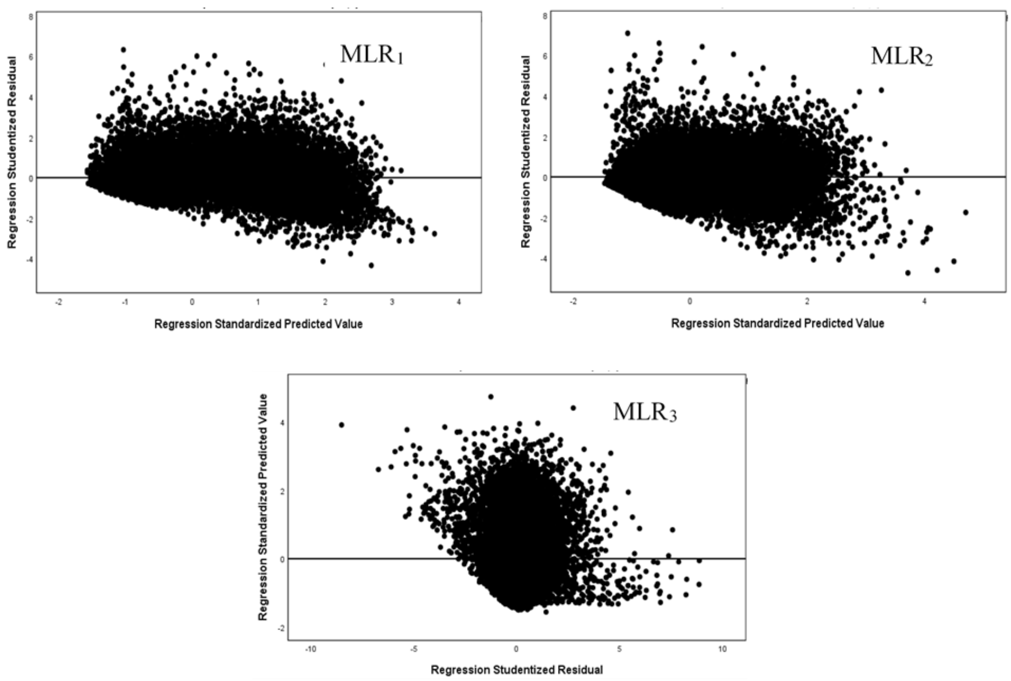

Fig. 4 - Testing assumption of variance and uncorrelated with mean equal to zero

Thirty per cent of the dataset used to plot the prediction of $O_{3,\,t+1}$ concentration against the observed O_3 concentration for the three different models to determine a best-fitted model for the industrial area in Terengganu, as shown in Fig5. MLR_1 had a correlation coefficient, R^2 values of 0.592, which was the highest compared to MLR_2 (0.439) and MLR_3 (0.262). Most of the points in each developed model were accumulated within a 95% confidence interval line, while the A and C lines were drawn as the upper and lower 95% confidence threshold for the MLR models.

In this study, the calculation of performance indicators is via error and accuracy measures. Error measure consisted of RMSE and MAE, while the accuracy measure consisted of R^2. Table 7 tabulated summary of the performance indicator. As a result, it showed that MLR_2 had two performance indicators, with the lowest value in the error measure of RMSE (3.977) and MAE (3.548) as compared to the values in MLR_1 (RMSE, 9.015; MAE, 8.834) and MLR_3 (RMSE, 6.806; MAE, 6.789), while the MLR_1 had the highest accuracy measure of R^2 (0.668) as compared to MLR_2 (0.431) and MLR_3 (0.359). Therefore, MLR_2 was selected as the best prediction model as it had an error value closest to zero and an accuracy measure close to one [33]. Based on a similar study by Pawlak and Jarosławski [42] conducted in Poland's rural and urban areas, the developed MLR model to predict O_3 concentration had managed to get the RMSE value of 16.3–15.9, MAE value of 13.0–15.9. and R^2 value of 13.0–15.9 for all models. The MLR models developed in two different urban areas in Hong Kong during four distinct seasons: summer, monsoon, post-monsoon, and winter. All the MLR models had RMSE, MAE, and R^2 with the range of 30.3–14.5, 29.6–11.4 and 0.64–0.54, respectively [43]. Awang et al. [33] established MLR and PCR models during the daytime, nighttime, and critical conversion time in the urban areas for forecasting the O_3 concentration. They found the best-fitted models selected based on two performance indicators: the model with the lowest RMSE value (20.28–7.01) and the highest R^2 value (0.74–0.23). The summary of a prediction model based on similar studies outcomes displayed in Table 8.

Table 7 - Summary of performance indicator

MLR Model	RMSE	MAE	R^2
1	9.015	8.834	0.668
2	3.977	3.548	0.431
3	6.806	6.789	0.359

Table 8 - The comparison of performance indicators in developing the O_3 concentration prediction models based on similar studies

Source	Country	Pollutant	Model	RMSE	MAE	R^2
Pawlak & Jarosławski, 2019 [42]	Poland	O_3	MLR	16.3-15.9	13.0-15.9	13.0-15.9
Zhang & Ding, 2017[43]	Hong Kong	O_3	MLR	30.3-14.5	29.6-11.4	0.64-0.54
Awang et al., 2015 [22]	Malaysia	O_3	MLR PCR	20.28-7.01	-	0.74-0.23

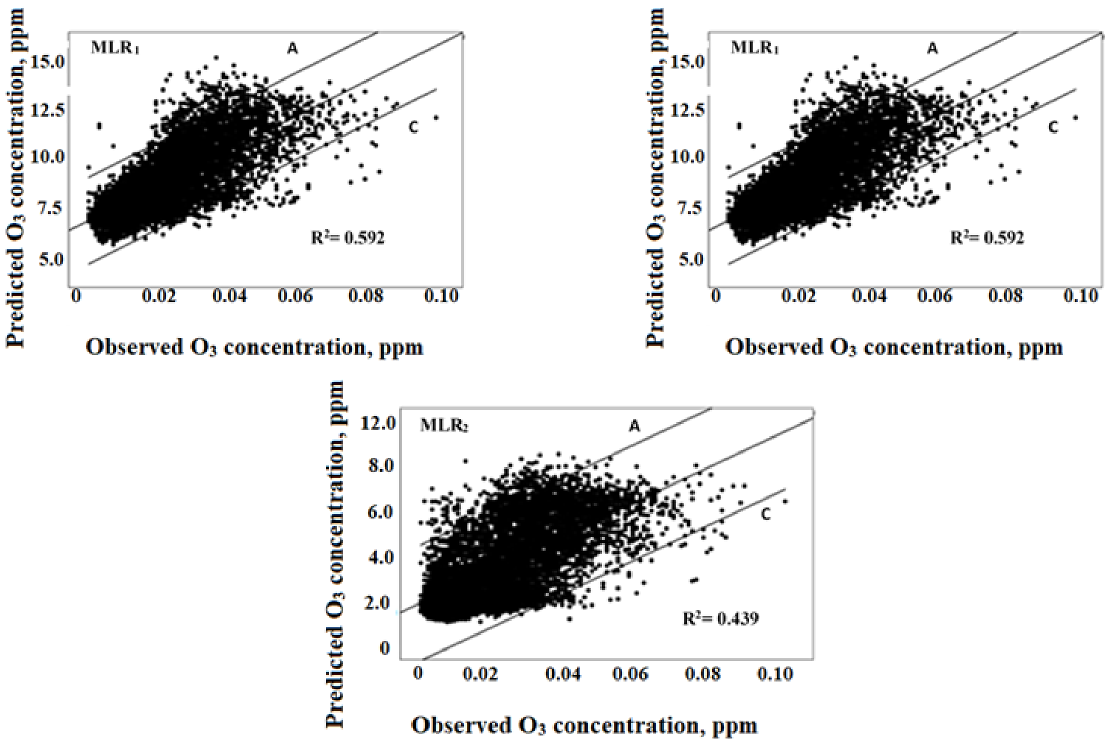

Fig 5 - Scatter plot of predicted O_3 concentration (ppm) against observed O_3 concentration (ppm)

4. Conclusion

The five-year data of O_3 concentration in the industrial area of Terengganu showed fluctuated trends from 2010 to 2014. WS and T exhibited a high positive association with the rise in O_3 concentration in the atmosphere, but RH had a high negative association with the rise in O_3 concentration. MLR_2 was chosen as the best-fitted prediction model for O_3 concentration based on its performance indicator since it had the lowest RMSE (3.977) and MAE (3.548) and a high R^2 (0.431).

Acknowledgement

This study is funded by the Fundamental Research Grant Scheme by the Malaysian Ministry of Higher Education (Ref: FRGS/1/2022/TK08/UMT/02/8) (VOT: 59716) and the Centre of Research and Innovation Management, Universiti Malaysia Terengganu. We would also like to thank the Air Quality Division, Malaysian Department of Environment to acquire air quality data.

References

[1] Ayuni, N. A., Juliana, J., & Ibrahim, M. H. (2014). Exposure to PM_{10} and NO_2 and Association with Respiratory Health among Primary School Children Living Near Petrochemical Industry Area at Kertih, Terengganu. Journal of Medical and Bioengineering, 3(4), 282–287. https://doi.org/10.12720/jomb.3.4.282-287

[2] Ismail, M., Abdullah, S., Si Yuen, F., & Ghazali, N. A. (2016). A ten-year investigation on ozone and it precursors at kemaman, Terengganu, Malaysia. EnvironmentAsia, 9(1), 1-8.

[3] Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. Air Quality, Atmosphere and Health, 3(1), 53–64. https://doi.org/10.1007/s11869-009-0051-1

[4] Rani, N. L. A., Azid, A., Khalit, S. I., Juahir, H., & Samsudin, M. S. (2018). Air pollution index trend analysis in Malaysia, 2010-15. Polish Journal of Environmental Studies, 27(2), 801–808. https://doi.org/10.15244/pjoes/75964

[5] World Health Organization, WHO (2005). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide, Summary Of Risk Assessment, Retrieved on April 17, 2020 from https://apps.who.int/iris/bitstream/handle/10665/69477/WHO_SDE_PHE_OEH_06.02_eng.pdf;jsessionid=11CD046AF1B03E5DCBBEF8DBC228537F?sequence=

[6] Department of Environment, Malaysia (2020), New Malaysia Ambient Air Quality Guidelines, Retrieved February 24, 2020 from https://www.doe.gov.my/portalv1/wp-content/uploads/2013/01/Air-Quality-Standard-BI.pdf

[7] 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.

[8] Perera, F. (2018). Pollution from fossil-fuel combustion is the leading environmental threat to global pediatric health and equity: Solutions exist. International journal of environmental research and public health, 15(1), 16.

[9] 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, 5(4), 238-260.https://doi.org/10.1007/s40726-019-00118-3

[10] Abdullah, A. M., Ismail, M., Yuen, F. S., Abdullah, S., & Elhadi, R. E. (2017). The relationship between daily maximum temperature and daily maximum ground level ozone concentration. Polish Journal of Environmental Studies, 26(2), 517–523. https://doi.org/10.15244/pjoes/65366

[11] United Stated Environmental Protection Agency, USEPA (2016), Ozone Concentration, Report on the Environment. Retrieved on April 16, 2020 from https://cfpub.epa.goa/roe/indicator_pdf.cfm?i=8

[12] Edwards, R. P., Engle, M., & Morris, G. (2020). Evaluation of El Niño-Southern oscillation influence on ozone exceedances along the United States Gulf Coast. Atmospheric Environment, 222(November 2019), 117127. https://doi.org/10.1016/j.atmosenv.2019.117127

[13] Simoni, M., Baldacci, S., Maio, S., Cerrai, S., Sarno, G., & Viegi, G. (2015). Adverse effects of outdoor pollution in the elderly. Journal of Thoracic Disease, 7(1), 34–45. https://doi.org/10.3978/j.issn.2072-1439.2014.12.10

[14] Sun, R., & Gu, D. (2008). Air pollution, economic development of communities, and health status among the elderly in urban China. American Journal of Epidemiology, 168(11), 1311–1318. https://doi.org/10.1093/aje/kwn260

[15] Xie, Y., Dai, H., Zhang, Y., Wu, Y., Hanaoka, T., & Masui, T. (2019). Comparison of health and economic impacts of $PM_{2.5}$ and ozone pollution in China. Environment International, 130, 104881. https://doi.org/10.1016/j.envint.2019.05.075

[16] Nuvolone, D., Petri, D., & Voller, F. (2018). The effects of ozone on human health. Environmental Science and Pollution Research International, 25(9), 8074–8088. https://doi.org/10.1007/s11356-017-9239-3

[17] Ismail, M., Suroto, A., & Abdullah, S. (2015). Response of malaysian local rice cultivars induced by elevated ozone stress. EnvironmentAsia, 8(1), 86-93. Retrieved from www.scopus.com

[18] Mills, G., Wagg, S., & Harmens, H. (2013). Ozone pollution: impacts on ecosystem services and biodiversity. NERC/Centre for Ecology & Hydrology.

[19] Bais, A. F., Lucas, R. M., Bornman, J. F., Williamson, C. E., Sulzberger, B., Austin, A. T., & Aucamp, P. J. (2018). Environmental effects of ozone depletion, UV radiation and interactions with climate change: UNEP Environmental Effects Assessment Panel, update 2017. Photochemical & Photobiological Sciences, 17(2), 127-179.

[20] 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

[21] 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

[22] Jumin, E., Zaini, N., Ahmed, A. N., Abdullah, S., Ismail, M., Sherif, M., Sefelnasr, A., and 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. doi:10.1080/19942060.2020.1758792

[23] Department of Statistics, Malaysia (2010), Population Distribution by Local Authority Areas and Mukims, Retrieved April 4, 2020 from https://web.archive.org/web/20150205090002/http://www.statistics.gov.my/portal/download_Population/files/population/03ringkasan_kawasan_PBT_Jadual1.pdf

[24] Abdullah, S., Napi, N.N.L.M., Ahmed, A.N., Mansor, W.N.W., Mansor, A.A., Ismail, M., Abdullah, A. and Ramly. Z.T.A. (2020). Development of Multiple Linear Regression for Particulate Matter (PM_{10}) Forecasting during Episodic Transboundary Haze Event in Malaysia. Atmosphere 11(3): 289. https://doi.org/10.3390/atmos11030289

[25] Awang, N. R., Elbayoumi, M., Ramli, N. A., & Yahaya, A. S. (2016). Diurnal variations of ground-level ozone in three port cities in Malaysia. Air Quality, Atmosphere and Health, 9(1), 25–39. https://doi.org/10.1007/s11869-015-0334-7

[26] 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

[27] 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

[28] Fan, J., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., & Xiang, Y. (2018). Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature. Renewable and Sustainable Energy Reviews, 94, 732–747. https://doi.org/10.1016/j.rser.2018.06.029

[29] 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 of Environmental Science, 2(2), 403–410. https://doi.org/10.6088/ijes.00202020003

[30] Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A., Rosaida, N., & Hamid, H. A. (2013). Future daily PM_{10} concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmospheric Environment, 77, 621–630. https://doi.org/10.1016/j.atmosenv.2013.05.017

[31] Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. Atmospheric Environment, 60, 172–181.https://doi.org/10.1016/j.atmosenv.2012.06.021

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

[33] Awang, N. R., Ramli, N. A., Yahaya, A. S., & Elbayoumi, M. (2015). 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

[34] Mohd Napi, N. N. L., Abdullah, S., Ahmed, A. N., Mansor, A. A., & Ismail, M. (2020). Annual and diurnal trend of surface ozone (O_3) in industrial area. Paper presented at the IOP Conference Series: Earth and Environmental Science, 498(1) doi:10.1088/1755-1315/498/1/012062 Retrieved from www.scopus.com

[35] Wang, J. and Ogawa, S. (2015). Effects of Meteorological Conditions on $PM_{2.5}$ Concentrations in Nagasaki, Japan. International Journal of Environmental Research and Public Health, 12(8): 9089–9101. https://doi.org/10.3390/ijerph120809089

[36] Roberts-Semple, D., Song, F., & Gao, Y. (2012). Seasonal characteristics of ambient nitrogen oxides and ground-level ozone in metropolitan northeastern New Jersey. Atmospheric Pollution Research, 3(2), 247–257. https://doi.org/10.5094/APR.2012.027
[37] Lal, S., Naja, M., & Subbaraya, B. H. (2000). Seasonal variations in surface ozone and its precursors over an urban site in India. Atmospheric Environment, 34(17), 2713–2724. https://doi.org/10.1016/S1352-2310(99)00510-5
[38] Quansah, E., Amekudzi, L. K., & Preko, K. (2012). The Influence of Temperature and Relative Humidity on Indoor Ozone Concentrations during the Harmattan. E. Journal of Emerging Trends in Engineering and Applied Sciences, 3(5), 863–867.
[39] Toh, Y. Y., Lim, S. F., & von Glasow, R. (2013). The influence of meteorological factors and biomass burning on surface ozone concentrations at Tanah Rata, Malaysia. Atmospheric Environment, 70, 435–446. https://doi.org/10.1016/j.atmosenv.2013.01.018
[40] Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Williams, M. L. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. Atmospheric Chemistry and Physics, 15(15), 8889–8973. https://doi.org/10.5194/acp-15-8889-2015
[41] 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
[42] Pawlak, I., & Jarosławski, J. (2019). Forecasting of surface ozone concentration by using artificial neural networks in rural and urban areas in central Poland. Atmosphere, 10(2). https://doi.org/10.3390/atmos10020052
[43] Zhang, J., & Ding, W. (2017). Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong. International Journal of Environmental Research and Public Health, 14(2), 1–19. https://doi.org/10.3390/ijerph14020114