Multiple Linear Regression (MLR) and Principal Component Regression (PCR) for Ozone (O_3) Concentrations Prediction

IOP Conf. Series: Earth and Environmental Science **616** (2020) 012004 doi:10.1088/1755-1315/616/1/012004

Multiple Linear Regression (MLR) and Principal Component Regression (PCR) for Ozone (O_3) Concentrations Prediction

Nur Nazmi Liyana Mohd Napi[1], Mohammad Syazwan Noor Mohamed[1], Samsuri Abdullah[1], Amalina Abu Mansor[2], Ali Najah Ahmed[3], Marzuki Ismail[2,4]

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

[2]Faculty of Science and Marine Environment, University Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.

[3]Institute of Energy Infrastructure (IEI), Department of Civil Engineering, College of Engi-neering, Universiti Tenaga Nasional (UNITEN), Kajang, Selangor Darul Ehsan 43000, Malaysia

[4]Institute of Tropical Biodiversity and Sustainable Development, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

E-mail: samsuri@edu.umt.my

Abstract. Rapid economic growth has led to an increase in ozone (O_3) concentration which significantly affecting human health and environment. The prediction of O_3 is complicated due to the redundancy of influencing parameters which introduce the multicollinearity problem. The aim of this study is to assess the best prediction model for O_3 concentration which is Multiple Linear Regression (MLR) and Principle Component Regression (PCR). Data from 2012 to 2014 were used including O_3, nitrogen dioxide (NO_2), nitrogen oxide (O_2), temperature, relative humidity and wind speed on hourly basis. Principle Component Analysis (PCA) was used in order to reduce multicollinearity problem, prior to the implementation of MLR. The hybrid model of PCR was selected as best -fitted models as it had higher correlation coefficient, R^2 values compared with MLR model. In conclusion, the information from best-fitted prediction model can be used by local authorities to plan the precaution measure in combating and preserve the better air quality level.

1. Introduction

High concentrations of O_3 pollutant and exposure to it tends to be incredibly dangerous to human wellbeing [1,2], particularly as it can make extraordinary harm the respiratory framework and initiate asthma in youngsters and hypertension in people who practice in outside situations. High O_3 concentration introduces a consequence to the ecosystems and increase global warming through greenhouse gases emission [3]. New Malaysian Ambient Air Quality Standard (NMAAQS) by Department of Environment Malaysia has suggested that the average of O_3 concentration for 1-hour is 200 μg/m^3. Multiple Linear Regression (MLR) is widely used for the air pollutant prediction based on several predictors by having better understanding of O_3 variation influenced factors. However, the multicollinearity problem become main concern in the MLR as it can reduce the reliability of the data

[4,5]. Hence, the Principal Component Regression, PCR has been introduced to reduce the multicollinearity problem. PCR is the hybrid models combination of Principal Component Analysis, PCA with the MLR [6]. This study intended to establish best prediction model for ozone in urban area. The developed model can be used by respective authorities and industrial players for mitigating air pollution, reducing human exposure towards detrimental high ozone concentrations and technological advancement for measurement of air pollutants.

2. Material and Methods

2.1. Data Monitoring

The study area is located at Klang, Selangor as it classifies as one of the urban areas. The monitoring station is pointed at Secondary School of Perempuan Raja Zarina, Pelabuhan Klang with latitude 3^0 35.9784" N and longitude 101^0 24' 30.1464" E. The air quality data that were used in this study are gained from Air Quality Division, Department of Environment (DOE), Ministry of Environment and Water, Malaysia through long term observing by private company, Alam Sekitar Sdn Bhd (ASMA) [6]. Hourly air quality data covering the period 2012 to 2014. The air pollutant parameters that were used in this study are O_3, SO_2, NO_X, NO, NO_2, and CO. Also, the meteorological parameters which are wind speed, temperature and relative humidity. The missing data were removed from this study to avoid from the bias [7]. The data was arranged and analysed by using Microsoft Excel Spreadsheet ® 2013 and Statistical Packages for Social Sciences (SPSS®) version 25. The Spearman Bivariate Correlation analysis was applied to identify the relationship between O_3 and independent variables as data set was non-parametric and not normally distributed [8]. The min-max normalization technique was applied as the parameters having various type of international system of units (SI) and the data were scaled from 0 to 1 value. The normalization of data was obtained through mathematical equation [7] as shown in equation (1).

$$z_i = ((x_i) - \min(x)) / (\max(x) - \min(x)) \quad (1)$$

Where, $x = x_1, \ldots, x_n$ (variables values) and z_i = Normalized data.

2.2. Multiple Linear Regression (MLR)

MLR attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observe the data [7]. Stepwise method had been used to develop the MLR model in this study. The residuals of MLR model was assumes it have normal distribution with zero mean, uncorrelated and constant variance [9]. General equation of MLR is as shown in equation (2).

$$y = b_0 + \Sigma_i{=}1 b_i X_i + \varepsilon \quad (2)$$

Where, b_i = Regression coefficients, X_i = Independent variables and ε = Stochastic error

2.3. Principal Component Regression (PCR)

PCR is the combination method of PCA and MLR or known as hybrid model. Grouping and selection of input for MLR were performed via PCA. Groups of PCs executed as a result of orthogonal transformation in PCA, which in each PC, the parameters are correlated to each other [10]. The PCA equation is shown as in equation (3).

$$PC_i = l_{1i} X_i + l_{2i} X_2 + \ldots\ldots + l_{ni} X_n \quad (3)$$

Where, $PC_i = i^{th}$ principal component and X_{ij} = The loading of the observed variable X_i.

3. Results and Discussion

3.1. The variation of ozone influencing factors

Correlation analysis is a statistical method used to evaluate the strength of relationship between two variables. The Spearman Bivariate Correlation Analysis is deemed suitable for the dataset as it is non-parametric and not normally distributed [11]. The correlation analysis between O_3, other gaseous pollutants and meteorological parameters was carried out as tabulated in table 1. In Klang, the O_3 concentration was having strong and positive correlation with WS ($r = 0.643$, $p<0.05$)) and T ($r = 0.668$, $p<0.05$)) while strong and negative relationship with RH ($r = 0.606$, $p<0.05$)). The O_3 precursor which are NO ($r = 0.476$, $p<0.05$)), NO_2 ($r = -0.393$, $p<0.05$)) and CO ($r = -0.488$, $p<0.05$) were having negatively moderate correlation with the O_3 concentration. SO_2 ($r = -0.019$, $p<0.05$) was having weak and negative effect to the O_3 concentration. Overall, T and RH parameter were become the main influencing factors that increase the O_3 concentration at both areas as the higher ambient temperature provide warm and dry condition promoted more frequent photochemical process to occur [12]. The other gaseous pollutants especially O_3 precursor such as NO, NO_2 and CO that emitted through incomplete combustion process from motor vehicles and industrial activities emission become the main contributor to the higher O_3 concentration in the atmosphere. An oxidant radical (hydroperoxyl radicals, HO_2), organic peroxyl radicals (RO_2), hydrocarbon and alkoxy radicals (RO) elements that contain in O_3 precursors help them converted to the O_3 by photochemical reaction [3].

Table 1. Summary of Spearman Correlation Analysis (r) between O_3 concentration with meteorological factor and other gaseous pollutants.

Parameter	O_3	WS	T	RH	NO	SO_2	NO_2	CO
O_3	1	0.643^{**}	0.668^{**}	-0.606^{**}	-0.476^{**}	-0.019^{*}	-0.393^{**}	-0.488^{**}

Note: ** Correlation is significant at the 0.01 level (2-tailed); * Correlation is significant at the 0.05 level (2-tailed); O_3= Ozone; WS = Windspeed; T = Temperature; RH = Relative humidity; NO = Nitrogen oxide; SO_2 = Sulphur dioxide; NO_2 = Nitrogen dioxide; CO = Carbon monoxide.

3.2. Development of the models

Multiple Linear Regression (MLR) and Principal Component Regression (PCR) were developed based on seventy percent of the data from the dataset. The determination of coefficient, R^2 for PCR (0.405) was higher compare to MLR (0.325). All the established models were not having the multicollinearity problem, as the variance inflation factor (VIF) was within 10 with range values 1.087 to 8.836 [4]. Models did not have any first order autocorrelation as the Durbin-Watson (D-W) values were within 2 with 0.764 (MLR) and 0.728 (PCR). The summary of the all models was depicted in table 2.

Table 2. Summary of developed models.

Method	Model	R^2	Range of VIF	DW
MLR	$O_3 = 0.121 + 0.441$ (T) - 0.147 (NO) + 0.14 (WS) – 0.186 (SO_2) – 0.119 (CO) + 0.046 (RH)	0.325	1.098 – 8.836	0.764
PCR	$O_3 = 0.154 + 0.232$ PC1 – 0.098 PC2 – 0.181 PC3	0.405	1.087 – 1.376	0.728

T, WS, RH, NO, SO_2, and CO are significant predictor variables in this MLR model. The increase of one unit of T, WS, and RH will increase of 0.441, 0.140 and 0.046 units of O_3. The O_3 concentration decreased about 0.147, 0.186 and 0.119 units when the one unit of NO, SO_2 and CO decreasing. The Kaiser Meyer Olkin (KMO) and Barlett's test was applied to determine the adequacy of the data and the degree of correlation between the input variables before the PCA can proceed [7,10]. In this study, the data having the KMO values 0.780 and Barlett's Test of Sphericity of less than 0.001 which fulfil the requirements. Rotated component matrix utilizing Kaiser Normalization of the three components was portrayed in table 3. The output is suppressed with values less than 0.4. The higher the factor loading, the more the variable contributes to the variation of the PCs. PC-1 was considered as meteorological factors as it consisted of the T, RH and WS parameters while PC-2 was categorized as ozone precursors that influences the O_3 in the ambient air. Open burning source was considered as PC-3.

Table 3. Rotated matrix.

Parameter	PC1	PC2	PC3
Wind speed	0.718		
Temperature	0.936		
Relative Humidity	-0.894		
NO		0.793	
SO_2			0.982
NO_2		0.775	
CO		0.841	

The residuals MLR and PCR models were normally distributed with zero mean and constant variances and distributed along the horizontal band which indicates the homogeneity of the variances and uncorrelated as illustrated in figure 1. The developed MLR and PCR models were validated through the scatter plot as shown in figure 2, based on the thirty percent of the data from the dataset. Upper and lower lines show the upper bound and lower bound of 95% confidence interval of dataset. Lines A and C are the upper and lower 95% confidence limit for regression model, respectively [6]. Most of the data were accumulated within the upper and lower bound of 95% confidence interval. Hence, the PCR was considered as best-fitted model with the correlation coefficient, R^2 is 0.405 which is higher than R^2 of MLR model, 0.325.

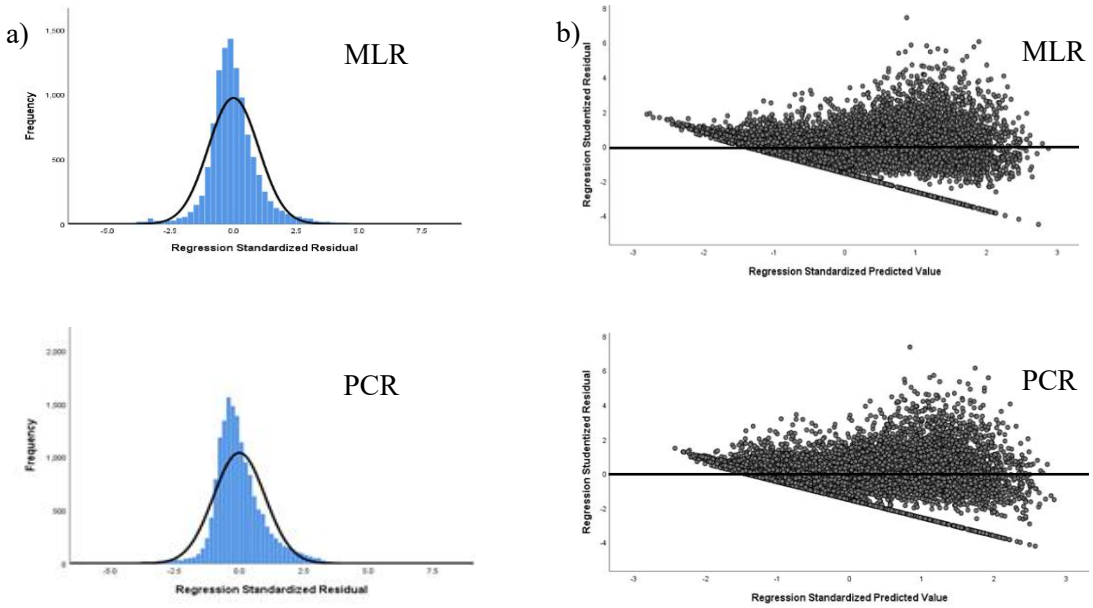

Figure 1. Residuals of MLR and PCR model with a) normal distribution b) Constant variance.

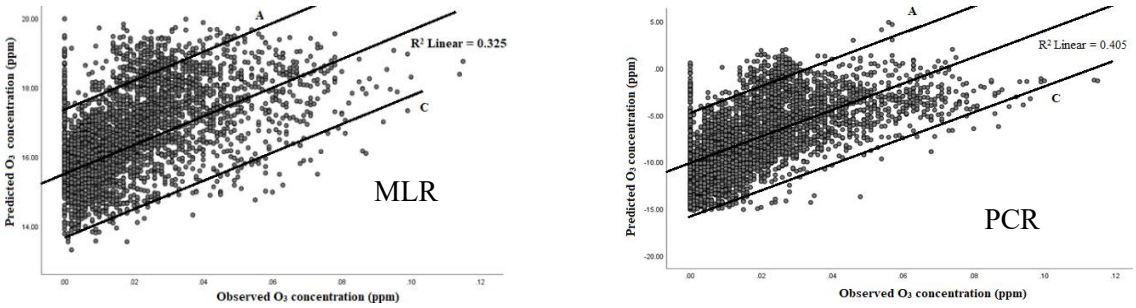

Figure 2. Validation of MLR and PCR model.

4.0 Conclusions

In conclusion, the meteorological factors and other gaseous pollutants, especially O_3 precursors were responsible for the increase of O_3 concentration in the urban area in Klang. PCR was selected as the best-fitted prediction model to predict the O_3 concentration, having higher of R^2 compared to MLR model.

References

[1] Nuvolone D, Petri D, & Voller F 2018 *Environmental Science and Pollution Research International* **9** 8074–8088
[2] Fuhrer J, Val Martin M, Mills G, Heald C L, Harmens H, Hayes F, & Ashmore M R 2016 *Ecology and evolution* **24** 8785-8799
[3] Lu X, Zhang L, & Shen L 2019 *Current Pollution Reports* **1**
[4] Daoud J I 2018 *Journal of Physics: Conference Series* **1** 949
[5] Tan K C, San Lim H, & Jafri M Z M 2016 *Atmospheric Pollution Research* **3** 533-546
[6] Abdullah S, Ismail M, Fong S Y, & Ahmed N 2016 *Environment Asia* **2**

[7] Awang N R, Ramli N A, Yahaya A S, & Elbayoumi M 2015 *Atmospheric Pollution Research* **5** 726–734
[8] Ul-Saufie A Z, Yahya A S, & Ramli N A 2011 *International Journal of Environmental Science* **2** 403–410
[9] Uyanık G K, & Güler N 2013 *Procedia – Social and Behavioral Sciences* **106** 234–240
[10] Nazif A, Mohammed N I, Malakahmad, A, & Abualqumboz M S 2018 *Environmental Science and Pollution Research* **1** 283–289
[11] Hauke J, & Kossowski T 2011 *Quaestiones Geographicae* **2** 87–93
[12] Edwards R P, Engle M, & Morris G 2020 *Atmospheric Environment* **222** 117-127