Prediction of Future Ozone Concentration for Next Three Days Using Linear Regression and Nonlinear Regression Models

IOP Conf. Series: Materials Science and Engineering **551** (2019) 012006 doi:10.1088/1757-899X/551/1/012006

Prediction of Future Ozone Concentration for Next Three Days Using Linear Regression and Nonlinear Regression Models

Nazirul Mubin Zahari[1], Raja Ezzah Shamimi[2], Mohd Hafiz Zawawi[3], Ahmad Zia Ul-Saufie[4] and Daud Mohamad[5]

[1,2,3,5]Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional, Jln IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
[4]Faculty of Computer and Mathematical Science, Universiti Teknologi MARA, Pulau Pinang, Malaysia.

E-mail: mubinzahari@gmail.com

Abstract. The aim of this research is to predict the ozone concentration level for the next three days. Linear regression model and a nonlinear regression model are used to measure the air pollution data and the result was compared. The performance indicator used to evaluate the accuracy of the methods is Index of Agreement (IA), Prediction Accuracy (PA) and Coefficient of Determination (R2). While Normalized Absolute Error (NAE) and Root Mean Square Error (RMSE) are for error measured. The results show that the prediction for the next three days. The highest ozone concentration of the linear regression model is 0.085ppm at Petaling Jaya, Selangor. While the lowest concentration for the linear regression model is 0.015 ppm at Klang, Selangor. Besides, the highest ozone concentration for the nonlinear regression model is 0.1 ppm at Petaling Jaya, Selangor for the second-day prediction. Comparison between the linear regression model and a nonlinear regression model indicates that nonlinear regression model can as an alternative method to the linear regression model.

1. Introduction

Malaysia is a newly developed country that experienced rapid urbanization and underwent an economic boom in the last ten years. Many industries have been growing and created new demands in housing, transportation, and urbanization as the population increases and cause environmental catastrophes especially pollution [1].

In term of air quality, Malaysia is in ranks 117th the worst country among 180 countries worldwide [2]. Air pollution structure contains six mains characteristics that always monitored by Department of Environment, which are ozone (O3), lead (Pb), carbon monoxide (CO), sulfur dioxide (SO2), nitrogen dioxide (NO2) and particulate matter (PM10). Air pollution is rick to human health, animal, damage crops and nature environment [3]. Studies have been proved that the dangerous impact of air pollution on human health specifically on patients suffering from cardiovascular and respiratory diseases [4]. Air pollution is the world's main environment risk for health and mostly this disease has suffered the population of developing countries [5]. In the past study, various types of statistical analysis have been applied to fit the air pollutant concentrations [3]. Based on the past study, researchers have applied regression models for predicting ozone concentration in different areas include in Malaysia [6-8]. This study focus on prédiction of linear regression and nonlinear regression model for future ozone

concentration level for the next three days. The final results will be compared to determine the best model.

2. Methodology

2.1 Study area

There is five air monitoring station in Malaysia that was selected for this research. It is to predict the future ozone concentration for the next day (D+1), next two days (D+2) and next three days (D+3). All the location selected based on the urban area, industrial area and the area that will influence the reading of air quality. This air monitoring station is monitored by the Department of Environment (DOE) Malaysia. The area of this study is Petaling Jaya, Perai, Seberang Jaya, Klang, and Nilai.

2.2 Linear Regression

The linear regression model is a statistical method to predict the relationship between of two variable which is dependent and independent variable. Linear regression attempts to model the relationship by fitting the linear equation to experimental data. This is the basic equation for the linear regression model:

$$\text{Response} = \text{constant} + \text{parameter} * \text{predictor} + .. + \text{parameter} * \text{predictor} \quad (1)$$

$$Y = bo + b1X1 + b2X2 + ... + bkXk \quad (2)$$

2.3 Nonlinear Regression

Nonlinear is extra complicated compared to linear to create. It is because the function is produced by a sequence of approximation and trial-and-error. To determine the equation is nonlinear or linear is if the equation does not meet the criteria of the linear equation, it is the nonlinear equation.

$$Y = E\ (b\ o + b1X1 + b2X2 + ... + bkXk) \quad (3)$$

2.4 Performance Indicator

The performance indicator for accuracy measures is the root mean square error (RMSE), index of agreement (IA), prediction accuracy (PA), and coefficient of determination (R2). While indicator for error measures are a normalized absolute error (NAE), root means square error (RMSE). This performance indicator is used to determine the best method in predicting ozone concentration for the next three days.

Table 1. Performance Indicator Parameter.

Performance Indicator	Equation	Description
Normalized Absolute Error (NAE)	$NAE = \frac{\sum_{t=1}^{n}(P_t - O_t)}{\sum_{t=1}^{n} O_t}$	NAE values close to 0 Method is correct
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(O_t - P_t)^2}{n}}$	RMSE values close to 0 Method is correct
Index of Agreement (IA)	$IA = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}\left\|P_i - \overline{O}\right\| - \left\|O_i - \overline{O}\right\|^2}$	IA values close to 1 Method is correct
Prediction Accuracy (PA)	$PA = \frac{\sum_{t=1}^{n}(P_t - \overline{O})^2}{\sum_{t=1}^{n}(O_t - \overline{O})^2}$	PA values close to 1 Method is correct
Coefficient of Determination (R2)	$R^2 = 1 - \frac{\sum_{t=1}^{n}(P_t - \overline{O})^2}{\sum_{t=1}^{n}(O_t - \overline{O})^2}$	R2 values close to 1 Method is correct

3. Results and Discussion

3.1 Descriptive Analysis

The ozone concentration data for 2009 until 2015 at five air monitoring station which are Petaling Jaya, Perai, Seberang Jaya, Klang, and Nilai is summarised in table 2. The highest average ozone concentration is Petaling Jaya compared to others study area. This five study area shows the ozone concentration data is skewed to the right that shows it has an extreme event. Based on the kurtosis value, Petaling Jaya and Nilai have positive kurtosis value. While Perai Klang and Seberang Jaya have negative kurtosis value.

Table 2. Descriptive Statistic.

Descriptive Analysis	Study Area				
	Petaling Jaya	Perai	Seberang Jaya	Klang	Nilai
Air Monitoring Station	Sek. Keb. Bandar Utama, Petaling Jaya	Sek. Keb. Cederawasih, Taman Inderawasih, Perai	Sek. Keb. Seberang Jaya II, Seberang Jaya	Sek. Men. (P) Raja Zarina, Klang	Tmn. Semarak (Phase II), Nilai
Coordinate	(3.14,101.61)	(5.39,100.39)	(5.38,100.40)	(3.01,101.41)	(2.82,101.82)
N	2109	1677	2288	2102	2121
Mean	0.047	0.042	0.039	0.045	0.045
Variance	0.000	0.000	0.000	0.000	0.000
Std. deviation	0.021	0.0182	0.018	0.020	0.018
Minimum	0.0009	0.0010	0.0020	0.0000	0.0020
Maximum	0.125	0.105	0.105	0.124	0.119
Skewness	0.419	0.435	0.410	0.464	0.424
Kurtosis	0.119	-0.123	-0.336	-0.025	0.322

3.2 Model Development

Linear regression model and a nonlinear regression model were developed using IBM SPSS Statistic 23 as shown in table 3 and table 4. The model was developed for the next day, next two days and the next three days at five research area which are Petaling Jaya, Perai, Seberang Jaya, Klang, and Nilai. These models were used to predict the ozone concentration.

Table 3. Model Development for Linear Regression Model.

Day of Prediction	Model
The next day (D+1)	$O3_{,D+1} = b_o + b_1T + b_2H + b_3SO_2 + b_4NO_2 + b_5O_3 + b_6CO + b_7PM_{10}$
The next two days (D+2)	$O3_{,D+2} = b_o + b_1T + b_2H + b_3SO_2 + b_4NO_2 + b_5O_3 + b_6CO + b_7PM_{10}$
The next three days (D+3)	$O3_{,D+3} = b_o + b_1T + b_2H + b_3SO_2 + b_4NO_2 + b_5O_3 + b_6CO + b_7PM_{10}$

Table 4. Model Development for Nonlinear Regression.

Day of Prediction	Model
The next day (D+1)	$O3_{,D+1} = E\,(b_o + b_1T + b_2H + b_3SO_2 + b_4NO_2 + b_5O_3 + b_6CO + b_7PM_{10})$
The next two days (D+2)	$O3_{,D+2} = E\,(b_o + b_1T + b_2H + b_3SO_2 + b_4NO_2 + b_5O_3 + b_6CO + b_7PM_{10}$
The next three days (D+3)	$O3_{,D+3} = E\,(b_o + b_1T + b_2H + b_3SO_2 + b_4NO_2 + b_5O_3 + b_6CO + b_7PM_{10}$

3.3 Prediction Model

As shown in figure 1-5, the model developed by the IBM SPSS Statistic 23 were used to predict ozone concentration in the future. Linear regression model and a nonlinear regression model were applied to get the best fit value to predict the ozone concentration level for the next three days. Based on the graph below, show the prediction value for the next three days. This research is based on the air pollution data from five monitoring station in Malaysia. After that, the results of the linear regression model and a nonlinear regression model were comparing to get the best method.

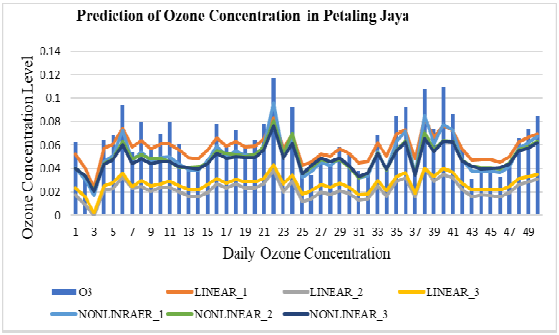

Figure 1. Prediction of Ozone Conc. in Petaling Jaya.

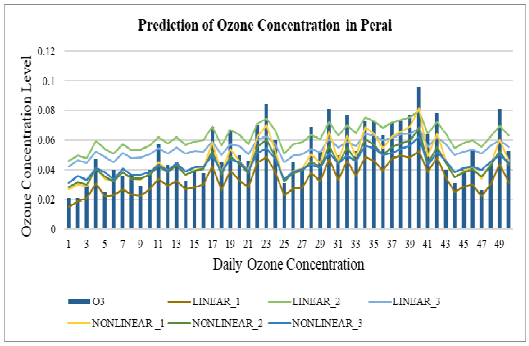

Figure 2. Prediction of Ozone Conc. In Perai.

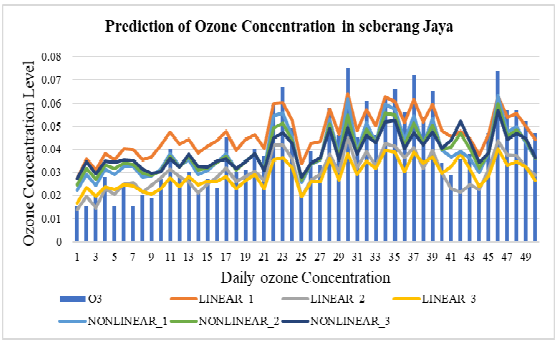

Figure 3. Prediction of Ozone Conc. in Seberang Jaya.

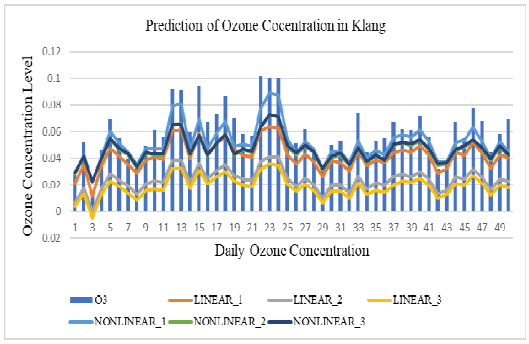

Figure 4. Prediction of Ozone Conc. in Klang.

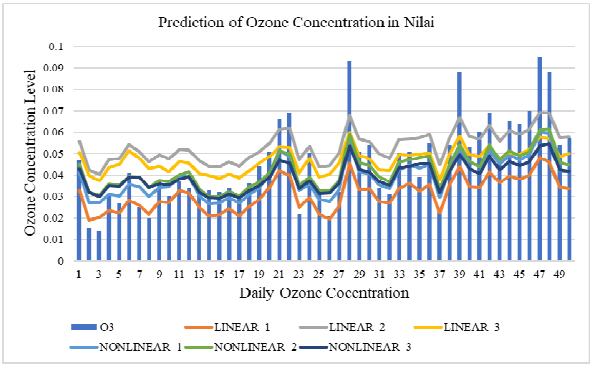

Figure 5. Prediction of Ozone Concentration in Nilai.

3.4 Identification Model

Table 5 shows a descriptive analysis on the performance indicator. The comparison between the two methods which is linear and non-linear. From the comparison of the performance indicators in Petaling Jaya, Perai, Seberang Jaya, Klang and Nilai, almost all the value of the NAE, RMSE, PA, R^2 and IA values between validation model and verification model are very close. The gap for each performance indicator is very small. The final value after compares to 0 and 1 between performances indicators also balance. So, we can conclude the nonlinear regression model can be the best alternative to a linear regression model to predict the ozone concentration in Malaysia. It is because the nonlinear regression model has slightly the same results with a linear regression model.

Table 5. Descriptive Statistic.

Sites	Prediction	Method	NAE	RMSE	PA	R^2	IA
Petaling Jaya, Selangor	D + 1	Linear	0.330202	0.018501	0.516835	0.266815	0.633424
		Nonlinear	0.311657	0.018625	0.478207	0.228422	0.652168
	D + 2	Linear	0.607149	0.033141	0.402198	0.161579	0.489003
		Nonlinear	0.332674	0.019401	0.382433	0.146088	0.571791
	D + 3	Linear	0.512534	0.029348	0.357862	0.127919	0.500480
		Nonlinear	0.340345	0.019767	0.343716	0.118006	0.534684
Perai, Pulau Pinang	D + 1	Linear	0.381158	0.018949	0.551498	0.303728	0.613635
		Nonlinear	0.285191	0.014568	0.536769	0.287721	0.692372
	D + 2	Linear	0.440168	0.020489	0.484774	0.234680	0.570974
		Nonlinear	0.304539	0.015398	0.468581	0.219263	0.611537
	D + 3	Linear	0.512534	0.029348	0.357862	0.127919	0.500480
		Nonlinear	0.340345	0.019767	0.343716	0.118006	0.534684
Seberang Jaya, Pulau Pinang	D + 1	Linear	0.307691	0.014158	0.633161	0.400485	0.753295
		Nonlinear	0.296841	0.014253	0.609894	0.371593	0.750966
	D + 2	Linear	0.390888	0.018531	0.545251	0.296996	0.619722
		Nonlinear	0.319681	0.015115	0.536364	0.287393	0.691487
	D + 3	Linear	0.426308	0.019958	0.498615	0.248364	0.580040
		Nonlinear	0.332746	0.015489	0.494118	0.243905	0.653095
Klang, Selangor	D + 1	Linear	0.342321	0.020378	0.606437	0.367364	0.663020
		Nonlinear	0.288836	0.017134	0.574277	0.329433	0.730889
	D + 2	Linear	0.651349	0.033240	0.523946	0.274219	0.514041
		Nonlinear	0.310164	0.018259	0.508485	0.258275	0.659017
	D + 3	Linear	0.734617	0.036782	0.485142	0.235106	0.482622
		Nonlinear	0.317200	0.018403	0.476569	0.226870	0.642043
Nilai, Negeri Sembilan	D + 1	Linear	0.356841	0.020533	0.446514	0.199150	0.555302
		Nonlinear	0.288206	0.016633	0.425909	0.181194	0.594215
	D + 2	Linear	0.363522	0.019390	0.386465	0.149187	0.533569
		Nonlinear	0.299020	0.016849	0.365729	0.133607	0.551138
	D + 3	Linear	0.308922	0.017003	0.336818	0.113319	0.484296
		Nonlinear	0.304960	0.017220	0.335922	0.112716	0.502955

4. Conclusions

In conclusion, nonlinear regression can be the alternative method to the linear regression. These two regressions are slightly the same in predicting ozone concentration level for the next three days in Malaysia. These prove that the linear regression model and the nonlinear regression model are approximately the same methods to predict the ozone concentration for the next three days in

Malaysia. Linear regression model and a nonlinear regression model can be used in prediction air quality data.

Acknowledgments
This study was supported by Universiti Tenaga Nasional internal grant (UNIIG) and special thanks to those who contributed to this project directly or indirectly.

References
[1] Afroz R *et al.* 2003 *Review of air pollution and health impacts in Malaysia* **92**
[2] EPI 2017 *Malaysia's Performance in Environmental Performance Index*
[3] Zawawi M H *et al*. *Journal of Scientific Research and Development* **3(4)** 106.
[4] Huebnerov Z *et al*. 2014 *Atmos. Pollut. Res* **5(3)** 471.
[5] Azam A G 2016 *J. Res. Med. Sci.* **21** 65.
[6] Ul-Saufie A Z *et al*. 2011 *International Journal of Applied Science and Technology* **1** 4
[7] Zahari N M 2018 *AIP Conference Proceedings* **2030** 020236.
[8] Isa M *et al*. 2018 *AIP Conference Proceedings* **765** 232.