Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 10, Number 6 (2014), pp. 869-874

http://www.ripublication.com

Estimation of Ozone Concentration in Rayong Province of Thailand Using MLR Method

J. Mekparyup[1], K. Saithanu[2] and S. Detudom[3]

[1,2,3]Department of Mathematics, Faculty of Science, Burapha University
169 Muang, Chonburi, Thailand
[1]jatupat@buu.ac.th, [2]ksaithan@buu.ac.th, [3]may_po02@hotmail.com

Abstract

The objective of this study was to estimate daily ozone concentration in Rayong province of Thailand using multiple linear regression (MLR) method. A dependent variable was ozone (O_3) and independent variables were Carbon Monoxide (CO), Nitrogen Monoxide (NO), Oxides of Nitrogen (NO_X), Sulfur Dioxide (SO_2), PM_{10}, wind speed, temperature, relative humidity (RH) and rainfall (RF). The results of the present study showed that the multiple linear regression equation for estimation daily ozone concentration in Rayong was $\hat{O}_3' = 5.951 + 0.664CO' - 0.619NO' + 0.666\,NO_X' - 0.610\,SO_2' - 2.939RH' + 0.058RF'$ with standard error of estimation 0.1544 and adjusted coefficient of determination 0.426.

Mathematics Subject Classification: 62J05

Keywords: MLR method, best subset method

INTRODUCTION

Thailand is a developing country in economic with expansion of industry to rural areas leading to increasing number of industrial factories which is faced the release of smog and concentration gas from these factories into the atmosphere causes air pollution problem.

O_3 is a major component of smog which is one of air pollutants spreading to air pollution and high level of O_3 is commonly found in large community or industrial areas. There are many factors that contribute to O_3, for example, rapidly increasing of the number of vehicles, burning waste in the open air and burning in the forest. Breathing with O_3 concentration more than standard level may make an impact on health problems such as irritation of respiratory system and cardiopulmonary

problems [1]. In addition, O_3 can affect lung function [2] and given the amount of O_3 for a long time may cause Chronic Bronchitis and Emphysema. The Notification of National Environmental Board No. 28, B.E 2550 (2007) is set O_3 concentration not more than 0.10 ppm in 1 hour and 0.07 ppm in 8 hours [3][4].

Rayong is a province in east of Thailand facing air pollutants problem of O_3 due to expanding industry rapidly, growing of the population and burning fuel and chemical in industrial process. Although the daily O_3 concentration is not still more than the air quality standard [5], O_3 concentration is trending to increase steadily. According to these reasons, the objective of present study aims to estimate daily O_3 concentration for planning to reduce air pollutant concentration of O_3 in Rayong province of Thailand using multiple linear regression (MLR) method.

MATERIALS AND METHODS

Air pollutant concentrations, Ozone (O_3), Carbon Monoxide (CO), Nitrogen Monoxide (NO), Oxides of Nitrogen (NO_X), Sulfur Dioxide (SO_2), PM_{10}, and meteorological factors, Wind Speed (WS), Temperature (T), Relative Humidity (RH), were collected from Air Monitoring Station at Rayong provincial agricultural extension office, Air Quality and Noise Management Bureau, Pollution Control Department, Thailand since May 4, 2014 to June 30, 2014.

1. CORRELATION

Correlation coefficient (R) is used to monitor relationship among air pollutant concentrations and meteorological factors.

2. THE MLR MODEL

In the present study, MLR method is used to analysis multivariate variables which consists of one dependent variable, daily ozone concentration (O_3), and 9 independent variables in each of day, carbon monoxide (CO), nitrogen monoxide (NO), oxides of nitrogen (NO_X), sulfur dioxide (SO_2), particulate matter 10 micrometers or less in diameter (PM_{10}), wind speed (WS), temperature (T), relative humidity (RH) and rainfall (RF). These variables are used to generate the MLR model following Equation 1.

$$\mathrm{O_3} = \beta_0 + \beta_1\mathrm{CO} + \beta_2\mathrm{NO} + \beta_3\mathrm{NO_X} + \beta_4\mathrm{SO_2} + \beta_5\mathrm{PM_{10}} + \beta_6\mathrm{WS} + \beta_7\mathrm{T} + \beta_8\mathrm{RH} + \beta_9\mathrm{RF} + \varepsilon \quad (1)$$

where $\beta_i =$ the regression coefficient $(i = 0,1,2,...,9)$ and ε = error of the regression model.

3. THE MLR EQUATION

For choosing the MLR equation, Mallows' C_p [11], standard error of estimation (S)

and adjusted coefficient of determination (R^2_{adj}) is considered using the best subset method. After received the MLR equation, then F test statistic of analysis of variance (ANOVA) is used to verify the equation.

4. ASSUMPTIONS OF THE MLR MODEL

There are four assumptions of the MLR model. (I) Normal distribution of the error is tested by Anderson-Darling statistic following Equation 2 [12].

$$AD = -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[\ln F(Y_i) + \ln(1 - F(Y_{n+1-i}))\right] \tag{2}$$

where F is the cumulative distribution function of the normal distribution and Y_i are the ordered data. (II) Independence of the errors is tested by Durbin-Watson statistic following Equation 3 [13].

$$DW = \sum_{i=1}^{n} (\varepsilon_i - \varepsilon_{i-1})^2 \Big/ \sum_{i=1}^{n} \varepsilon_i^{\,2} \tag{3}$$

(III) Homoscedasticity of the errors is tested by Breusch-Pagan statistic following Equation 4 [14].

$$BP = \frac{SSR^*}{2} \div \left(\frac{SSE}{n}\right)^2 \tag{4}$$

where SSR^* is the regression sum of squares using e^2 as a dependent variable and SSE is the error sum of squares using O_3 as a dependent variable. (IV) Multicollinearity among independent variables is tested by Variance Inflation Factor (VIF) following Equation 5.

$$VIF_j = \frac{1}{1 - R^2_{j|others}} \tag{5}$$

where $R^2_{j|others}$ is the coefficient of multiple determination with independent variable x_j on the $p-2$ other independent variables x in the multiple linear regression model (p is the number of independent variables).

If any of these four assumptions is violated, two ways will be used for solving these problems. First, the MLR equation will be rectified such as using the procedure of Box-Cox [15] or Johnson [16] to transform the dependent variable, etc. Secondly, others MLR equation will be chosen.

5. VALIDATION OF THE MLR EQUATION

Once the MLR equation which is in agreement of all assumptions is obtained, comparison between the observed data (OBS) and the estimated data (EST) are validated using time series plot, scatter plot and the percentage of error (PE) calculated following Equation 6.

$$PE = \frac{|OBS - EST|}{OBS} X100\% \quad (6)$$

RESULTS AND DISCUSSION

O_3 and RF showed the highest positive correlation coefficient with R=0.492 (P-value=0.000) and O_3 and RH was the second highly one with R=0.475 (P-value=0.000). However, O_3 and T presented the only one negative correlation coefficient with R=–4.43 (P-value=0.001).

CO, NO, NO_X, SO_2, RH and R was used to build the MLR equation using the best subset method which gave Mallows' C_p= 5.3. The selected variables influenced to O_3 were the same previous studies of [6][7][8][9][10]. The regression equation was shown in Equation 7 with S=2.9840 and R^2_{adj}=39.2 (F=7.12, P-value=0.000).

$$\hat{O}_3 = 54.671 + 25.088CO - 2.043NO + 0.883NO_X - 2.866SO_2 - 0.601RH + 3.647RF \quad (7)$$

Then, the assumptions of MLR analysis was monitored; (I) the error of the MLR distributed normality (AD=0.155, P-value=0.954), (II) the errors were not significant independence (DW=1.134, a critical value D_L=1.214) so the MLR equation was adjusted using the Box-Cox transformation. The adjusted MLR equation was then displayed as Equation 8 with S=0.1544 and R^2_{adj}=0.426 (F=7.44, P-value=0.000).

$$\hat{O}'_3 = 15.951 + 0.664CO' - 0.619NO' + 0.666NO'_X - 0.610SO'_2 - 2.939RH' + 0.058RF' \quad (8)$$

where $O'_3 = \ln O_3$, $CO' = \ln CO$, $NO' = \ln NO$, $NO'_X = \ln NO_X$, $SO'_2 = \ln SO_2$, $RH' = \ln RH$ and $RF' = \ln RF$. Reconsidering all assumptions of Equation 8; (I) the error of the adjusted MLR distributed normality (AD=0.308, P-value=0.548), (II) the errors were significant independence (DW=1.281, a critical value D_L=1.214), (III) the variance of errors was significant constant (BP=4.411, P-value=0.056), (IV) there was no multicollinearity problem among the independent variables with low VIF values ($VIF_{CO'} = 1.9$, $VIF_{SO'_2} = 1.4$, $VIF_{RH'} = 2.3$, $VIF_{RF'} = 1.7$) and high VIF values

$(\mathrm{VIF}_{NO'} = 10.0, \mathrm{VIF}_{NO'_X} = 10.6)$ [17] because these was an usual chemical correlation so it was not necessary to solve this problem.

Finally, the graphical validation of the MLR equation was considered. Figure 1a showed time series plot between the OBS and the EST values and Figure 1b showed scatter plot between the OBS and the EST values with correlation coefficient R=0.702 (P-value=0.000). Furthermore, the PE values illustrated that the highest and the lowest values were 9.96% and 0.25% consequently so the validation of the MLR equation was satisfied.

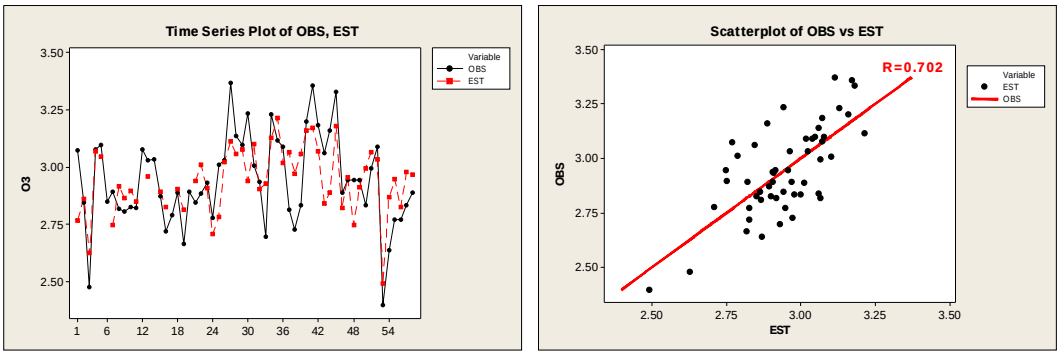

Figure 1: Comparison between OBS and EST in Rayong; (a) Time series plot, (b) Scatter plot

ACKNOWLEDGEMENT

The authors wish to thank to the Air Quality and Noise Management Bureau, Pollution Control Department, Thailand for kind support collecting data.

REFERENCES

[1] Protection, U. E, 2008, "Ozone Nation," Environmental health perspectives, 116(7).

[2] World Health Organization, 2003, "Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide," Report on a WHO Working Group.

[3] Notification of National Environmental Board No. 10, B.E.2538, 1995, "Air Quality and Noise Standards", the Royal Government Gazette No. 112 Part 52 dated May 25, B.E.2538.

[4] Notification of National Environmental Board No. 28, B.E.2550, 2007, "Air Quality and Noise Standards", the Royal Government Gazette No. 124 Part 58 dated May 14, B.E.2550.

[5] Word Health Organization, 2014, "WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide," Retrieved May 20, 2014 from Word Health Organization Web site:

http://whqlibdoc.who.int/hq/2006/WHO_SDE_PHE_OEH_06.02_eng.pdf

[6] Sousa, S.I.V., Martins, F.G., Pereira, M.C., & Alvim-Ferraz, M.C.M., 2006, "Prediction of ozone concentrations in Oporto city with statistical approaches," Chemosphere, 64(7), 1141-1149.

[7] Pai, T.Y., Sung, P.J., Lin, C.Y., Leu, H.G., Shieh, Y.R., Chang, S.C., ... & Jou, J.J., 2009, "Predicting hourly ozone concentration in Dali area of Taichung County based on multiple linear regression method," International journal of applied science and engineering, 7(2), 127-132.

[8] 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 & Software, 20(10), 1263-1271.

[9] Barrero, M.A., Grimalt, J.O., & Cantón, L., 2006, "Prediction of daily ozone concentration maxima in the urban atmosphere," Chemometrics and Intelligent Laboratory Systems, 80(1), 67-76.

[10] Lengyel, A., Héberger, K., Paksy, L., Bánhidi, O., & Rajkó, R., 2004, "Prediction of ozone concentration in ambient air using multivariate methods," Chemosphere, 57(8), 889-896.

[11] Hocking, R.R., & Leslie, R.N., 1967, "Selection of the best subset in regression analysis," Technometrics, 9(4), 531-540.

[12] Lewis, P.A.W., 1961, "Distribution of the Anderson-Darling Statistic," The Annals of Mathematical Statistics, 32(4), 1118-1124.

[13] Durbin, J., & Watson, G.S., 1951, "Testing for Serial Correlation in Least Squares Regression II," Biometrika, 38(2), 159-177.

[14] Breusch T.S., & Pagan, A.R., "A Simple Test for heteroscedasticity and Random Coefficient Variation," Econometrica, 47(5), 1287-1294.

[15] Sakia, R.M., 1992, "The Box-Cox transformation technique: a review," The statistician, 169-178.

[16] Johnson, N.L., 1949, "Systems of frequency curves generated by methods of translation," Biometrika, 149-176.

[17] O'brien, R.M., 2007, "A caution regarding rules of thumb for variance inflation factors," Quality & Quantity, 41(5), 673-690.