International Journal of Applied Science and Engineering
2009. 7, 2: 127-132

Predicting hourly ozone concentration in Dali area of Taichung County based on multiple linear regression method

Tzu-Yi Pai [a, *], Pao-Jui Sung [a, b], Chung-Yi Lin [b], Horng-Guang Leu [c], Yein-Rui Shieh [c], Shuenn-Chin Chang [c], Shyh-Wei Chen [c], Jin-Juh Jou [c]

[a] *Department of Environmental Engineering and Management, Chaoyang University of Technology, Wufeng, Taichung, 41349, Taiwan*
[b] *Dali City Administration, Taichung County Government, Dali, Taichung, 41261, Taiwan*
[c] *Environmental Protection Administration, Taipei, 10042, Taiwan*

Abstract: In this study, multiple linear regression (MLR) method was used to establish the relationship between the O_3 at time t + 1 and other indices including hourly air pollutant concentrations and meteorological conditions at time t. Then O_3 was predicted using the obtained best-fitting MLR. The results indicated that the relationship between the O_3 at time t + 1 and other indices including hourly air pollutant concentrations and meteorological conditions at time t agreed with MLR well, The values of mean absolute percentage error (MAPE), correlation coefficient (R), coefficient of determination (R^2), mean square error (MSE), and root mean square error (RMSE) were 29.09 %, 0.95, 0.90, 45.33, and 6.73, respectively when determining the best-fitting equation. In addition, MLR could predict hourly ozone concentrations successfully. The values of MAPE, R, R^2, MSE, and RMSE were 10.37 %, 0.93, 0.86, 0.33, and 0.57, respectively when predicting. It also indicated that the hourly air pollutant concentrations and meteorological conditions at time t could be applied on the prediction of ozone of time t +1.

Keywords: multiple linear regression; ozone; air quality; meteorological conditions; photochemical reaction

1. Introduction

In the past two decades, air pollution has improved in most cities in Western Europe, North American as well as Japan. Air pollution reductions have resulted mainly from greater efficiency and pollution-control technologies in factories, power plants, and other facilities [1] Although improvements are also achieved in transportation, the regulation efficiencies of mobile pollution sources are not as significant as those of stationary pollution sources because their mobile, emitted characteristics [2-6].

Among all air pollutants, the elevated O_3 concentrations at ground level are of particular concern, because of the harm to human health and vegetation. Gao and Niemeier [7] indicated that ozone pollution was caused by photochemical reactions of precursor volatile organic compounds (often called non-methane hydrocarbons, NMHC) and nitrogen oxides,

* Corresponding author; e-mail: bai@ms6.hinet.net
Accepted for Publication: January 25, 2010,

of which transportation emissions are the single major source. Several literatures also proved that the mobile sources had a significant influence on ozone formation [7-9]. In addition, the emissions of NMHC are one of the main contributors to ozone formation [10].

The relationship between ozone and its precursors is complicated due to the fact that meteorological and chemical reaction rates range from very fast to very slow. Such relationships between meteorological condition and ozone concentrations have been explored in several studies which have utilized statistical regression, graphical analysis, fuzzy theory, and cluster analysis.

Multiple linear regression method (MLR) is one of the most widely used methods for expressing the dependence between a dependent variable and several independent variables. If the relation between the O_3 and other hourly air pollutant concentrations could be found using MLR, a better control strategy could be sought.

Since the predicting performance of MLR is good, so the objectives of this study are using MLR to establish the relationship between the O_3, other hourly air pollutant concentrations, and meteorological conditions, then to predict O_3.

2. Materials and methods

2.1. Data set

The monitoring data from air quality monitoring station locating in Dali area of Taichung County was selected for study. The concentrations of O_3, other hourly air pollutant concentrations and from 29th of July to 16th of August 2008 were investigated. They were sampled and investigated every hour and their total number of data was 456. Among the data, 384 data points were used to obtain the coefficients of the models and 72 data points were used as the observed values when evaluating the performance of the model. The mean value and standard deviation of O_3 was 25.0 ppb and 21.2 ppb, respectively. The hourly air pollutant concentrations included methane (CH_4), carbon monoxide (CO), carbon dioxide (CO_2), NMHC, nitrogen monoxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NOx), fine particulates ($PM_{2.5}$ and PM_{10}), and sulfur dioxide (SO_2). The meteorological conditions included ambient temperature (Temp), rainfall, relative humidity (RH), wind speed (WS), and ultraviolet B (UVB).

2.2. Brief description on MLR

MLR determines the relationship between two or more independent variables and a dependent variable by fitting a linear equation to observed data. Every value of the independent variable is associated with a value of the dependent variable. In this study, the regression equation for independent variables including hourly air pollutant concentrations and meteorological conditions is defined as follows:

$$[O_3]_{t+1} = a_1\times[CH_4]_t + a_2\times[CO]_t + a_3\times[CO_2]_t + a_4\times[NMHC]_t + a_5\times[NO]_t + a_6\times[NO_2]_t + a_7\times[NOx]_t + a_8\times[O_3]_t + a_9\times[PM_{10}]_t + a_{10}\times[PM_{2.5}]_t + a_{11}\times[SO_2]_t + a_{12}\times[Temp]_t + a_{13}\times[rainfall]_t + a_{14}\times[RH]_t + a_{15}\times[WS]_t + a_{16}\times[UVB]_t + Constant \quad (1)$$

Equation (1) described how the average response of O_3 at time t +1 changed with the independent variables at time t, i.e. hourly air pollutant concentrations and meteorological conditions.

For the least-squares model, the best-fitting equation for the observed O_3 data at time t+1 was calculated by minimizing the sum of the squares of the vertical deviations from each data point to the regression equation (if a data point lay on the fitted line exactly, then its vertical deviation was 0). The least-squares estimates a_1 to a_{15} and constant were computed by MATLAB software.

3. Results and discussion

3.1. Determination of best-fitting equation

For determining the best-fitting equation, the observed O_3 data at time t+1 was calculated by minimizing the sum of the squares of the vertical deviations from each data point to the regression equation. The least-squares estimates a_1 to a_{15} and constant were computed as shown in Table 1.

Table 1. The coefficients in best-fitting equation

Item	CH_4	CO	CO_2	NMHC	NO	NO_2	NOx	O_3	PM_{10}
Coefficient	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
Value	3.28	12.61	0.01	-14.79	-0.07	-0.69	0.28	0.70	0.11
Item	$PM_{2.5}$	SO_2	Temp	rainfall	RH	WS	UVB	Constant	
Coefficient	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	Constant	
Value	0.01	0.65	-1.41	1.09	-0.22	-0.92	1.68	48.40	

According to Table 1, CH_4, CO, CO_2, NOx, O_3, PM_{10}, $PM_{2.5}$, SO_2, rainfall, and UVB at time t showed a positive effect on O_3 concentration at time t+1. Contrarily, NMHC, NO, NO_2, Temp, RH, and WS at time t showed a negative effect on O_3 concentration at time t+1.

The pollutants including CH_4, CO, CO_2, NMHC, NO, NO_2, and NOx were precursors of O_3. Theoretically, a rise in ozone concentration was associated with a drop in the levels of CH_4, CO, CO_2, NMHC, NO, NO_2, and NOx (Abdul-Wahab et al., 2005). But, there existed an industrial park and traffic loading was heavy in Dali area, various emission conditions resulted in different level of precursors. Therefore some precursors showed opposite trends with theoretical reaction.

Besides, O_3 concentration at time t + 1 was negatively but strongly affected by Temp at time t, it was positively and strongly affected by UVB at time t. Abdul-Wahab et al. [11] indicated that the reaction of O_3 showed a different behavior at day time and night time.

During daylight hours, temperature and UVB were high, but temperature was lower and UVB became zero during night time. Since the O_3 concentrations of 24 hours were calculated for determining the best-fitting equation in this study, the obtained coefficients for different indices showed a whole effect on O_3 concentration either at day time or night time.

3.2. Simulation of O_3

Since the best-fitting equation was determined, in order to evaluate the fitting and predicting performance of MLR, the mean absolute percentage error (MAPE), correlation coefficient (R), coefficient of determination (R^2), mean square error (MSE), and root mean square error (RMSE) were employed and described as,

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{obs_i - pre_i}{obs_i}\right| \times 100\% \tag{2}$$

$$R = \frac{\sum_{i=1}^{n}(obs_i - \overline{obs})(pre_i - \overline{pre})}{\sqrt{\sum_{i=1}^{n}(obs_i - \overline{obs})^2 \sum_{i=1}^{n}(pre_i - \overline{pre})^2}} \tag{3}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(pre_i - \overline{obs_i}\right)^2}{\sum_{i=1}^{n}\left(pre_i - \overline{obs_i}\right)^2 + \sum_{i=1}^{n}\left(pre_i - obs_i\right)^2} \tag{4}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(obs_i - pre_i\right)^2 \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(obs_i - pre_i\right)^2} \quad (6)$$

where obs_i is the observed value, pre_i is the prediction value, $\overline{obs}$ and $\overline{pre}$ are the average values of observed values and prediction values, respectively.

Figure 1 depicts the best-fitting results of ozone using MLR when determining the best-fitting equation. The values of MAPE, R, R^2, MSE, and RMSE were 29.09, 0.95, 0.90, 45.33, and 6.73, respectively when determining the best-fitting equation as shown in Table 2. It showed the relationship between the O_3 concentrations at time t +1 and other hourly indices at time t agreed with Equation (1) well.

Figure 2 shows the prediction results of ozone using MLR when predicting ozone at time t + 1. According to Table 2, the values of MAPE, R, R^2, MSE, and RMSE were 10.37, 0.93, 0.86, 0.33, and 0.57, respectively when predicting. It showed that the indices at time t could predict the O_3 concentrations at time t +1 well using MLR.

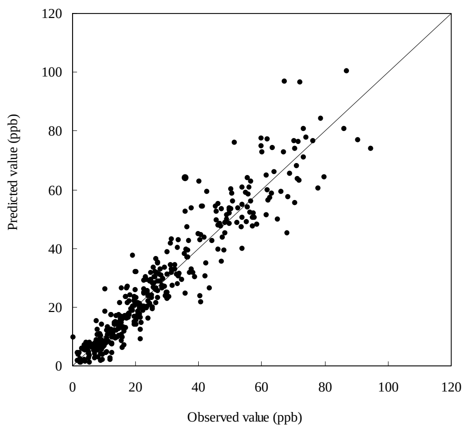

Figure 1. The best-fitting results when determining the best-fitting equation.

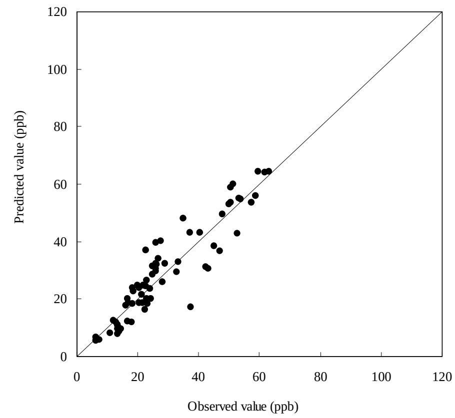

Figure 2. The prediction results using MLR.

Table 2. The performance of MLR

	Best-fitting	Predicting
MAPE	29.09	10.37
R	0.95	0.93
R^2	0.90	0.86
MSE	45.33	0.33
RMSE	6.73	0.57

Comparable observations were similarly made by Abdul-Wahab et al. (2005). Abdul-Wahab et al.[11] employed data on the concentrations of seven environmental pollutants (CH_4, NMHC, CO, CO_2, NO, NO_2 and SO_2) and meteorological variables (WS and direction, Temp, RH and solar radiation) to predict the concentration of ozone in the atmosphere using both multiple linear and principal component regression methods. They found that R^2 for the day and night periods, were of 0.82 and 0.76, respectively. In this study, the R^2 of 0.86 was obtained using MLR. Although the hourly O_3 concentration could be predicted well using regression model, the regression model should be tested with a new set of data to present it predicting power in the future.

4. Conclusions

In this study, MLR was used to establish the relationship between the O_3 at time t + 1 and other indices including hourly air pollutant concentrations and meteorological conditions at time t. Then O_3 was predicted using the obtained best-fitting MLR. The results can be drawn as follows:

According to the results, the relationship between the O_3 at time t + 1 and other indices including hourly air pollutant concentrations and meteorological conditions at time t agreed with MLR well, The values of MAPE, R, R^2, MSE, and RMSE were 29.09 %, 0.95, 0.90, 45.33, and 6.73, respectively when determining the best-fitting equation. In addition, MLR could predict hourly ozone concentrations successfully. The values of MAPE, R, R^2, MSE, and RMSE were 10.37, 0.93, 0.86, 0.33, and 0.57, respectively when predicting. It also indicated that the hourly air pollutant concentrations and meteorological conditions at time t could be applied on the prediction of ozone of time t +1.

After prediction, it is suggested that the MLR can be used as the objective function or constrains in optimization for best control or management in the future study.

References

[1] Cunningham, W. P., and Cunningham, M. A. 2006. Principles of Environmental Science. Inquiry & Applications. Mc Graw- Hill Company, New York.

[2] Faiz, A., Gautam, S., and Burki, E. 1995. Air pollution from motor vehicles: issues and options for Latin American countries. *The Science of the Total Environment,* 169, 1-3: 303-310.

[3] Fischer, P. H., Hoek, G., van Reeuwijk, H., Briggs, D. J., Lebret, E., van Wijnen, J.H., and Kingham, S. 2000. Traffic-related differences in outdoor and indoor concentrations of particles and volatile organic compounds in *Amsterdam. Atmospheric Environment,* 34, 22: 3713-3722.

[4] Kingham, S., Briggs, D., Elliott, P., Fischer, P., and Erik, L. 2000. Spatial variations in the concentrations of traffic-related pollutants in indoor and outdoor air in Huddersfield, England. *Atmospheric Environment,* 34, 6: 905-916.

[5] Lipfert, F. W., Wyzga, R. E., Baty, J. D., and Miller, J. P. 2006. Traffic density as a surrogate measure of environmental exposures in studies of air pollution health effects: long-term mortality in a cohort of US veterans. *Atmospheric Environment,* 40, 1: 154-169.

[6] Pai, T. Y., Hanaki, K., Ho H. H., and Hsieh, C. M. 2007. Using grey system theory to evaluate transportation on air quality trends in Japan, Transportation Research Part D: *Transport and Environment,* 12, 3: 158-166.

[7] Gao, H. O., 2007. Day of week effects on diurnal ozone/NOx cycles and transportation emissions in Southern California. Transportation Research Part D: *Transport and Environment,* 12, 4: 292-305.

[8] Gao, H. O., and Niemeier, D. A. 2007. The impact of rush hour traffic and mix on the ozone weekend effect in southern California. Transportation Research Part D: *Transport and Environment,* 12, 2: 83-98.

[9] Wang, G., Bai, S., and Ogden, J. M., 2009. Identifying contributions of on-road motor vehicles to urban air pollution using travel demand model data. Transportation Research Part D: *Transport and Environment*, 14, 3: 168–179.

[10] Delucchi, M. A., Greene, D .L., and Wang, M. Q. 1994. Motor-vehicle fuel economy: The forgotten hydrocarbon control strategy. Transportation Research Part A: *Policy and Practice,* 28, 3: 223-244.

[11] Abdul-Wahab, S. A., Bakheit, C. S., and 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.