한국환경보건학회지, 제44권 제6호(2018)
J Environ Health Sci. 2018; 44(6): 539-547

pISSN: 1738-4087 eISSN: 2233-8616
https://doi.org/10.5668/JEHS.2018.44.6.539

원 저 | Original articles

Simple Forecasting of Surface Ozone through a Statistical Approach

Chang-Jin Ma* and Gong-Unn Kang†**

**Department of Environmental Science, Fukuoka Women's University*
***Department of Medical Administration, Wonkwang Health Science University*

ABSTRACT

Objectives: Ozone (O_3) advisories are issued by provincial/prefectural and city governments in Korea and Japan when oxidant concentrations exceed the criteria of the related country. Advisories issued only after exposure to high O_3 concentrations cannot be considered ideal measures. Forecasts of O_3 would be more beneficial to citizens' health and daily life than real-time advisories. The present study was undertaken to present a simplified forecasting model that can predict surface O_3 concentrations for the afternoon of the day of the forecast.

Methods: For the construction of a simple and practical model, a multivariate regression model was applied. The monitored data on gases and climate variables from Japan's air quality networks that were recorded over nearly one year starting from April 2016 were applied as the subject for our model.

Results: A well-known inverse correlation between NO_2 and O_3 was confirmed by the monitored data for Iksan, Korea and Fukuoka, Japan. Typical time fluctuations for O_3 and NO_x were also found. Our model suggests that insolation is the most influential factor in determining the concentration of O_3. CH_4 also plays a major role in our model. It was possible to visually check for the fit of a theoretical distribution to the observed data by examining the probability-probability (P-P) scatter plot. The goodness of fit of the model in this study was also successfully validated through a comparison ($r=0.8$, $p<0.05$) of the measured and predicted O_3 concentrations.

Conclusions: The advantage of our model is that it is capable of immediate forecasting of surface O_3 for the afternoon of the day from the routinely measured values of the precursor and meteorological parameters. Although a comparison to other approaches for O_3 forecasting was not carried out, the model suggested in this study would be very helpful for the citizens of Korea and Japan, especially during the O_3 season from May to June.

Keywords: Ozone, photochemical smog, nitrous oxide, regression model, health effect

I. Introduction

Ozone (O_3) is one of the pollutants of major concern because of its well-recognized effects upon human health and environment.[1,2] O_3 is artificially created by human activities, most abundantly when gasoline is burned incompletely in internal combustion engines like vehicles, airplanes, cargo ships, and anything else with an engine that exhausts carbon monoxide, nitrogen oxides, and hydrocarbons. Under high temperatures and solar radiation, these fugitive precursors become O_3 through photochemical reactions in the atmosphere. O_3 can be also created naturally from volatile organic substances emitted by lightning, soil, and plants. This ground-level O_3, in other words surface O_3, is one of the most significant atmospheric photochemical products.

In recent decades, due to the increasing of motor vehicle and industrial activities, surface O_3 concentrations have been substantially increased.[3] As mentioned

†**Corresponding author:** Department of Medical Administration, Wonkwang Health Science University, Iksan, Korea, Tel: +82-10-8629-7700, E-mail: gukang@wu.ac.kr
Received: 12 October 2018, Revised: 14 November 2018, Accepted: 22 November 2018

above, high level of O_3 on surface adversely affects human health (especially for people who spend long hours outside) ecological system, and cultural heritage buildings.[4,5] O_3 can cause headaches, loss of consciousness and other problems for people who spend long hours outside and can cause coughing and eye irritation in less serious cases.

While a high density of O_3 is usually most intense within the densely populated/industrialized areas, a high level of O_3 and its precursors can be transported thousands of kilometers away thus making a large contribution to global tropospheric O_3[6]. Pollutants and their precursors from China and their resultant problems are nothing new to the East Asia, especially, Korean Peninsular and Japan. In particular, the regions located in the west of Korea can be directly affected by China. As an example, O_3 concentrations exceeded the national criteria (8-hr average 0.06 ppm) in all monitoring sites in Jeollabuk-do, Korea.[7] This region faces increasing risks of problems related to O_3 as a result of enforced domestic industrialization for last decades and long-range transport from China.

In addition, as an interesting thing, two regions in the western part of Korea and Japan (i.e., Imsil, Korea and Banryu, Japan) had well-correlated monthly O_3 concentrations (R=0.79).[8]

In Japan, added to traditional Acid rain, PM2.5 and photochemical smog have been a concern for recent decades in Japan. Although pollution from China extends over nearly the entire Japanese, Western Japan, especially Kyushu island, is most vulnerable to pollutants from Chinese continent.[9] As the nation's first official health warning, the local governments of western Japan, especially Fukuoka Prefecture, have advised residents to stay indoors during the episodically high O_3 period. A large number of prefectures in Japan are issuing an increased number of photochemical smog warning.

A 1.4 time of the pollutants observed in the Kyushu region of Japan in May 2007, when levels of photochemical smog were particularly high, originated in China and 30 percent of those observed in the Tokyo area at that time were also from China (Environmental issues and pollution in Japan). Trends in concentration of photochemical oxidant has recently been increased in Fukuoka prefecture.[8][10] Its concentration sometimes exceeded the photo-chemical smog advisory level (120 ppb (v/v)).

Although Fukuoka Prefecture government is giving information of the forecasted PM2.5 for the afternoon of the day, however that of O_3 prediction is not executed in earnest yet. Forecasting of surface O_3 can contribute to improving not only our health risk but urban air quality because the frequent occurrence of high concentrations of aerosols is also most likely caused by photochemical reactions. High PM2.5 concentrations were observed in the west Japan when photochemical smog advisories were issued[9]. Therefore, in addition to PM2.5, O_3 forecasting for the frequent occurrence of high concentrations of O_3 from May to June is promptly required.

There are two elements of all the forecasting model assessment. One is the accuracy and the other is speed. To ensure the accuracy of the model, modelers can use a lot of the associated parameters. Added to this, a mechanistic model assuming that a complex system can be understood by examining the workings of its individual parts and the manner in which they are coupled should be chosen rather than statistical models.

Meanwhile, as a stochastic approach, the auto-regressive integrated moving average (ARIMA) modeling suggested by Box and Jenkins[11] that has been used in forecasting the daily maximum surface O_3 concentrations.[12]

Many papers related to surface O_3 prediction with statistical approach have been published.[13-16] However, a model performance that will meet the simplicity, accuracy, and speed, and will be a helpful reference for the citizens' activities of the afternoon on the day has not been realized.

The object of this study is to provide a good

model performance that will meet the simplicity, i.e., a simple and speedy model that can easily take advantage of the officer. Therefore, a multivariate regression model based on direct observation, measurement and extensive data records was applied. Also, in this study, to ensure the accuracy of model result, several major parameters that are accurately measurable with relatively few measuring errors were selected. The factors measured in real time were chosen to secure a rapid forecasting result.

II. Methods

1. Variable selection and statistical procedure for multi-step regression model

It is not easy to make the accurate O_3 forecasting because of the complexity of the mechanisms of O_3 formation in the troposphere, the complexity of surface meteorological conditions in urban areas, and the uncertainty in the measurements of all the parameters involved.[17] Therefore, in this study, to propose a simple and practical model for surface O_3 forecasting, two critical variables were selected from both precursor variables and meteorological variables. CH_4 and NO_2 were selected as the well-known precursor variables. In addition to precursor variables, the diurnal variations in O_3 concentration is also closely related to those in local meteorological conditions. The meteorological factors determining surface O_3 level are ambient temperature, solar radiation, relative humidity (RH), and wind speed.

One of earlier studies, the result of extensive spatiotemporal analyses of surface O_3 and related meteorological variables in South Korea had revealed that the correlation coefficients (r) between the daily maximum 8 h average O_3 and temperature varied from 0.11 to 0.47 and from 0.37 to 0.69 at coastal and inland, respectively, while solar intensity showed more strong correlation with O_3 at both regions (r=0.70 and 0.85 at coastal and inland, respectively.[18] Therefore, insolation radiation was selected instead of temperature in this study.

Although the evaluation of correlationships between O_3 concentration and RH was not performed in this study, Tu *et al.*[19] reported that relative humidity has a significant negative correlation with O_3. Therefore, in this study, the predicted insolation and RH data in that afternoon were selected out of many meteorological parameters.

2. Data and its measuring site

The National Institute for Environmental Studies of Japan provides hourly data of air pollutants concentration and meteorological data through the Atmospheric Environmental Regional Observation System (AEROS). Among air pollutants, O_3, NO_2, CH_4 have been measured in the ppbv unit by the ultraviolet absorption method (JIS, Japanese Industrial Standards, 7957), the chemiluminescence method (JIS 7953), and the flame Ionization Detector (FID) (JIS 7956), respectively. Their details on measurement, e.g., measuring instruments and their accuracy, reproducibility, and calibration, are referred to in each JIS number.

Kasumigaoka monitoring station (33°40'11.93"N, 130°26'9.57") of the AEROS, which is near in the west coast of Fukuoka City, Japan was selected as data acquisition site. This monitoring site is located in Higashi-ku of Fukuoka City, Japan. Higashi-ku is one of the seven wards of Fukuoka City in Japan. As of 1 March 2017, Fukuoka City has a population of 296,576, with 136,133 households and an area of 66.68 km^2. Details on Fukuoka City were described in elsewhere.[20]

To investigate the relationship between O_3 and meteorological variables, it is desirable to use data observed at the same station. As meteorological data, relative humidity and insolation data were also obtained from the Kasumigaoka monitoring station and they were measured bycapacitance method and thermopile method, respectively. This study was based on the average hourly values of surface O_3, its precursors, and meteorological data for about a year (from April 2016 to February 2017). The data

were limited to blue-sky.

3. Concept of multivariate analysis

Multivariate analysis is the most popular technique used in operational O_3 forecasting.[21] It attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Its form is as follows.

$$y_i = B_0 + B_1x_1 + ... + B_4x_4 + E \quad (1)$$

where

y_i=dependent variable–concentration of O_3

$B_{0\text{-}4}$=constant

x_1=independent variable–CH_4

x_2=independent variable–NO_2

x_3=independent variable–insolation

x_4=independent variable–RH

E=random error in prediction known as residuals.

III. Results and Discussion

1. Measured O_3 and NO_x

In order to determine the characteristics of observed data prior to model construction, we calculated the hourly averages for NO_2 and O_3 during the sunny day (clear sky day) of whole observed period (April of 2016). Their diurnal variations are shown in Fig. 1. In addition to the data of AERO, the hourly averaged NO_2 and O_3 data of the same period provided by the air quality monitoring network in Iksan, Korea are simultaneously displayed in Fig. 1. As mentioned above, the issue of high O_3 concentrations at Jeollabuk-do (especially Iksan city) located in the southwestern part of Korea is of great concern to the local community.

As shown in figure, the diurnal variation of O_3 shows a monomodal peak during afternoon (especially around 3 p.m.) and low concentrations during early morning. On the other hand, NO_2 shows bimodal curves at night and early morning and low concentrations during afternoon. Their timely fluctuation patterns are fairly matched. This timely fluctuation is mainly due to both photochemical reaction and meteorological conditions. From 08:00 to 15:00, an increase in global solar radiation and the height of the mixing layer results in a decrease in NO_x concentration and an increase in O_3.[22] Similar results have been reported by Han *et al.*[23] in the study on the ambient atmosphere of Tianjin, China.

The peaks of NO_2 at morning rush hours was probably caused by its daily injection from vehicles. In general, the morning peak of NO_2 appears 1-2 hours after the NO peak because NO is slowly converted to NO_2 by the thermal reaction.

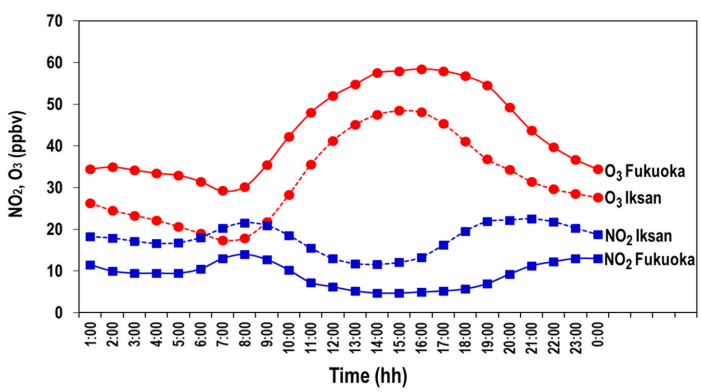

Fig. 1. Hourly variation of the average NO_2 and O_3 observed at Iksan, Korea and Fukuoka, Japan during the sunny days (clear sky day) during the whole observation period (from April 2016 to February 2017).

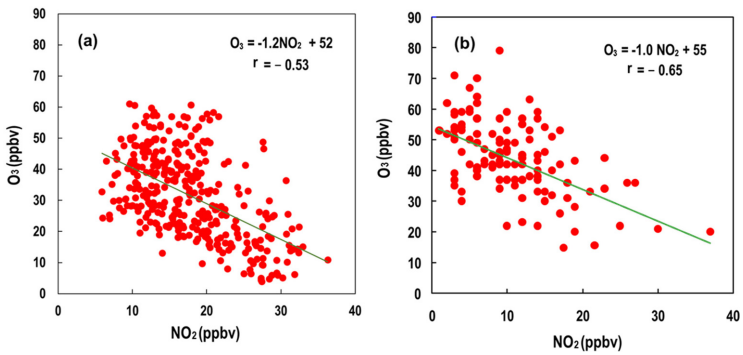

Fig. 2. Scatter plots of the NO_2 and ozone concentrations observed at Iksan, Korea (a) and Fukuoka, Japan (b).

Increasing of NO_2 at night was due to low mixing layer heights.

A scatter plot of the observed O_3 and NO_2 concentrations observed at both sites is shown in Fig. 2. O_3 has negative correlation with NO_2 in both regions. This result is similar to that observed at an urban site in China.[19)]

Since established by Leighton[24)] it is widely known that, at sunrise, photochemical reactions are initiated, and the reactions represented by equation (2) followed by (3) can be expected.

$$NO_2 + h\upsilon \rightarrow NO + O \quad (2)$$

$$O + O_2 + M \rightarrow O_3 + M \quad (3)$$

2. Ozone Modelling

2.1. Multicollinearity diagnostics

Fig. 3 illustrates the statistical procedure for forecasting the afternoon of the day O_3 levels at Fukuoka site. Since CH_4 and relative humidity are not being measured in the air quality monitoring network in Iksan, the data of Fukuoka was only used in model construction.

As the first step in multiple regression analysis, a multicollinearity must be checked. Although the explanatory variables must be independent from each other, in some applications of regression, the explanatory variables are related to each other. This is called the multicollinearity problem.[25)] For confirming the no multicollinearity, a multicollinearity index known as variance inflation factor (*VIF*) was calculated. The *VIFs* of four explanatory variables is calculated by the following equation[26)]:

$$VIF_i = \frac{1}{1 - R_i^2}, i = 1, 2, \ldots, n. \quad (4)$$

where n is the number of predictor variables and R^2_i is the square of the multiple correlation coefficient of the i th variable with the remaining (n-1) variables.

As shown in Table 1, the calculated *VIF*s among combination of four explanatory variables (from x_1 to x_4) varied from 1.88 to 3.07. Hossain *et al.*[26)] suggested that if $0 < VIF < 5$, there is no evidence of multicollinearity problem. In order word, multicollinearity test demonstrates that there is no correlation among four explanatory variables.

2.2. The coefficients for four kinds variables

Table 1 shows the coefficients for four kinds variables by multi step regression model. The estimated multiple linear regression model was as follow.

$$O_3\ (\text{ppb}) = -173.53 + 10.27CH_4 - 1.03NO_2 + 11.31\text{Insolation} - 0.08\text{RH} \quad (5)$$

According to above constructed model, insolation is the most influential factor in determining the concentration of O_3. Our model also indicates that in addition to insolation, CH_4 plays a major role in determining the high O_3 concentration. Meanwhile,

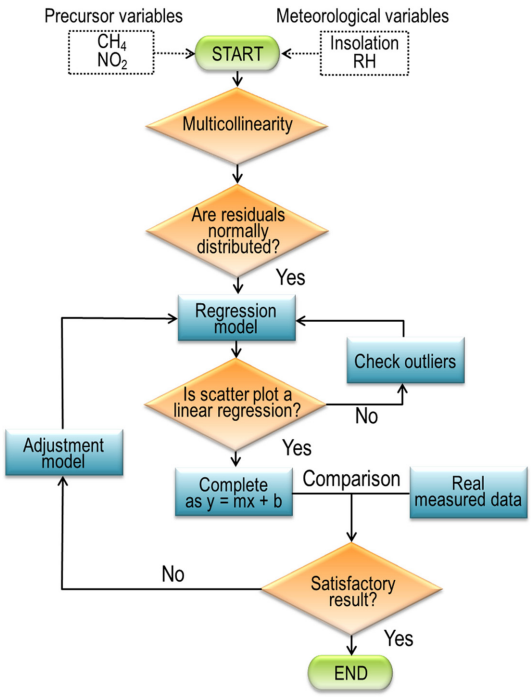

Fig. 3. Flowchart for multi step regression model with step-wise induction of variables.

relative humidity shows a bit of a role to lower O_3 concentration. This may seem reasonable because, as the previous study[19] reported, relative humidity has a negative correlation with O_3. Wang *et al.*[8] also indicated that the O_3 concentrations of Mondy and Japan sites had negative correlation with humidity.

2.3. Tests for residual normality and a normal probability

As a key part of next steps after model-building, a residual plot check is needed. Residuals are obtained by subtracting the observed data from the predicted data. Examining residuals tell us whether our assumptions are reasonable, and our choice of model is appropriate. The overall pattern of the residuals should be like the bell-shaped pattern. In general, a histogram can be used to assess the assumption that the residuals are normally distributed.

Figure 4 shows the histogram of regression standardized residuals with a normal density function on the histogram. Standardized residual terms frequency graphics for the dependent variable must show a normal dispersion. To the extent that the histogram matches the normal distribution, the residuals are normally distributed. Since the histogram graphic in Fig. 4 shows that such requirement has been met.

As a more sensitive graph for model validity test, a scatter plot between the observed cumulative probability and theoretical cumulative probability is also used. The cumulative probability of each residual can be calculated using the following formula:

$$P_{Cumulative}(i-th\ residual) = \frac{i}{N+1} \quad (6)$$

where P is the cumulative probability of a point, i

Table 1. The coefficients for four kinds variables by multi step regression model

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	VIF
	B	Std.Error	Beta			
Constant	-173.53	41.17		-4.21	0.00	
CH_4	10.27	2.16	0.36	4.75	0.00	2.62
NO_2	-1.03	0.15	-0.44	-6.76	0.00	1.88
Insolation	11.31	1.57	0.47	7.20	0.00	1.95
Relatve humidity	-0.08	0.12	-0.05	-0.64	0.52	3.07

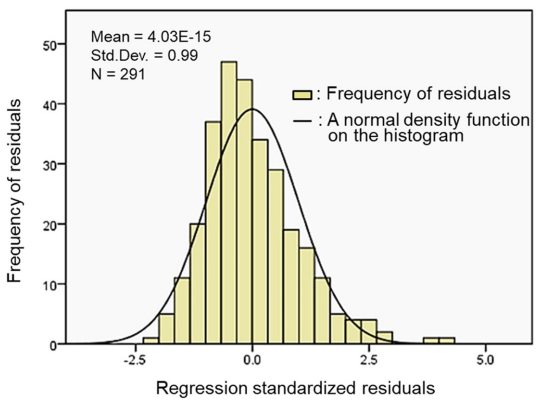

Fig. 4. Approximately normal distribution of standardized residuals produced by a model for a calibration process.

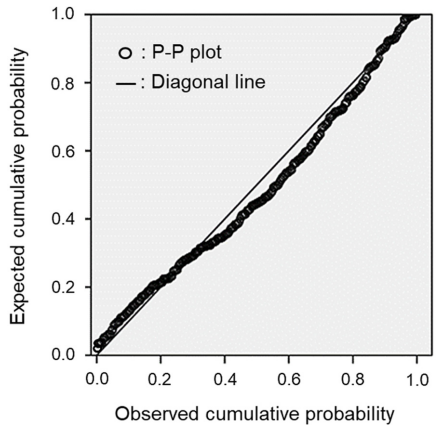

Fig. 5. Normal P-P (the observed probability and expected probability) scatter plot of regression standardized residuals.

is the order of the value in the list, and *N* is the number of entries in the list.

Figure 5 shows the P-P scatter plot of the standardized residuals, which are relatively evenly distributed on both sides of zero. As shown in figure, the expected cumulative distribution approximates the observed distribution well, then all points described as circles stand close to the diagonal line. It can be said that the model in this study is valid, and homoscedasticity is established.

3. Model Validation

In addition to investigating a residual normality and a normal probability, it is equally important to validate the model forecasts. Figure 6 shows the scatter plots with a solid regression line of the measured and predicted O_3 concentrations. This scatter plots turned out a quite higher correlation ($r = 0.87$ with significance level of 0.05) of between the measured and predicted O_3 concentrations. Therefore, the model designed in this study usually give the best forecasts of 7 h ahead. The ability of the model in making accurate O_3 forecasting was demonstrated by applying it to real data set of the observed ambient O_3. Figure 7 shows comparative results of time series of 7 h ahead O_3 concentration forecasted from in this study and those of observed.

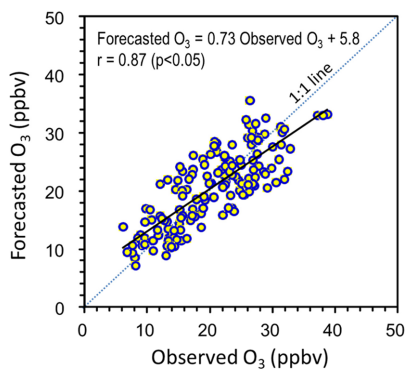

Fig. 6. Scatter plot of the observed and forecasted O3 concentrations.

A sustained daily fluctuation pattern is shown during the whole target period. As mentioned earlier, the daily maximum O_3 concentration is greatly affected by the O_3 concentration of the morning on the day. Although regression models assume that the observed O_3 data are statistically independent each other, however, it is partially dependent on the previous day's concentrations. As a high correlation between observed and forecasted O_3 has been proved through above scatter plot (Fig. 6), it turns out to match each other well. Although the detailed interpretation of annual O_3 cycles is not done because it is not the point of this study, O_3 level

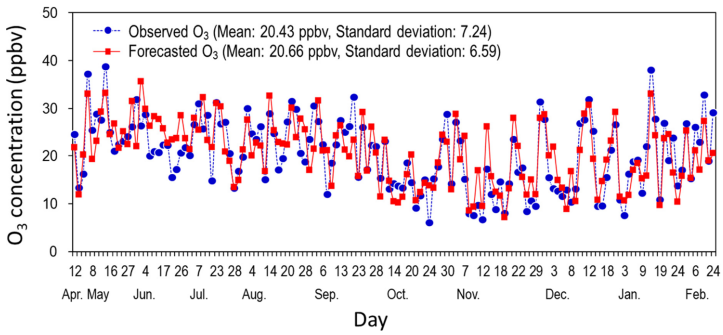

Fig. 7. Comparative results of time series of the 7 hours ahead O_3 concentration forecasted from in this study and those of observed.

increased during spring, reached the peak in late spring, and then decreased in summer and finally dropped to the bottom in autumn.

IV. Conclusions

Korea and Japan located in the downwind of China, there are a lot of people who are hindered in everyday life due to episodically high O_3 concentrations. This study aims at giving the forecasted O_3 data by the statistical model to the citizen who worry about being exposed to hazardous surface O_3 in their daily life. The model constructed in this study points out that insolation and CH_4 bring a significant impact on O_3 concentration. The advantage of our model that is capable of immediate forecasting the surface O_3 of the afternoon of the day from the routinely measured values of the precursor and meteorological parameters. Although a comparison to other approaches for O_3 forecasting was not carried out, the model suggested in this study is very helpful for the citizen of Korea and Japan for especially O_3 season from May to June. More studies are necessary to verify our model fitness at different sites in whole East Asian region. Further studies using a long-term (at least more than 10 years) observational data are also planning to make the accuracy of statistical model and to assess the validity of the model result.

Acknowledgment

This paper was supported by Wonkwang Health Science University in 2018.

References

1. Lee JH, Oh IB., Kin MH, Bang JH, Park SJ, Yoon SH, Kim YH., Change in the prevalence of allergic diseases and its association with air pollution in major cities of Korea - Population under 19 years old in different land-use areas -. Korean Journal of Environmental Health. 2017; 43(6): 478-490 (in Korean).
2. Bae HJ, Ha JS, Lee AK, Park JI. Age dependencies in air pollution-associated asthma hospitalization. Korean Journal of Environmental Health. 2008; 34(2): 124-130 (in Korean).
3. Sicard P, De Marco A, Troussier F, Renou C, Vas N Paoletti E. Decrease in ozone mean concentrations at Mediterranean remote sites and increase in the cities. Atmospheric Environment. 2013; 79: 705-715.
4. Contran N, Paoletti E. Visible foliar injury and physiological responses to ozone in Italian provenances of Fraxinus excelsior and Fornus. The Scientific World Journal. 2007; 7: 90-97.
5. Paoletti E. Impacts of ozone on Mediterranean forests: A review. Environmental Pollution. 2006; 144: 463-474.
6. Mauzerall DL., Jacob DJ, Fan SM, Bradshaw JD, Gregory GL, Sachse GW, Blake DR. Origin of tropospheric ozone at remote high northern latitudes in summer. Journal of Geophysical Research. 1996; 101: 4175-4188.

7. Jeonbuk Daily. 2017. http://www.jjan.kr/news/articleView.html?idxno=1139914
8. Wang E, Sakurai T, Ueda H. Assessment of ozone variability in East Asia during recent years. Acid Deposition Monitoring Network in East Asia (EANET) Science Bulletin. 2008; 1: pp. 3-20.
9. Hara Y, Uno I, Shimizu A, Sugimoto N, Matsui I, Yumimoto K, Kurokawa J, Ohara T, Liu Z. Seasonal characteristics of spherical aerosol distribution in eastern Asia: Integrated analysis using ground/space-based lidars and a chemical transport model. SOLA. 2011; 7: 121-124. doi: 10.2151/sola.2011-031
10. Iwamoto S, Oishi O, Tagami S, Chikara H, Yamamoto S. Classify causes for high concentration of photochemical oxidant in Fukuoka prefecture. Japan Society for Atmospheric Environment. 2008; 43: 173-179.
11. Box GEP, Jenkins GM. Time series analysis forecasting and control; Holden day: San Francisco, CA.1976.
12. Kumar K, Yadav AK, Singh MP, Hassan H, Jain VK. Forecasting daily maximum surface ozone concentrations in Brunei Darussalam-An ARIMA modeling approach. Journal of the Air & Waste Management Association. 2004; 54(7): 809-814, DOI: 10.1080/10473289.2004.
13. Ortiz-garcía EG, Salcedo-Sanz S, Pérez-Bellido ÁM, Portilla-Figueras JA, Prieto L. Prediction of hourly O_3 concentrations using support vector regression algorithms, Atmospheric Environment. 2010; 44: 4481-4488.
14. Sousa SIV, Martins FG, Alvim-Ferraz MCM, Pereira MC. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environmental Modelling Software. 2007; 22: 97-103.
15. Wang W, Lu W, Wang X, Leung AYT. Prediction of maximum daily ozone level sing combined neural network and statistical characteristics. Environment International. 2003; 29: 555-562.
16. Robeson SM, Steyn DG. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. Atmospheric Environment. 1990; 24B: 303-312.
17. Argiriou AA. Use of neural networks for tropospheric ozone time series approximation and forecasting. Atmospheric Chemistry and Physics Discussions. 2007; 7: 5739-5767.
18. Seo J, Youn D, Kim JY, Lee H. Extensive spatiotemporal analyses of surface ozone and related meteorological variables in South Korea for the period 1999-2010. Atmospheric Chemistry and Physics. 2014; 14: 6395-6415
19. Tu J, Xia ZG, Wang H, Li W. Temporal variations in surface ozone and its precursors and meteorological effects at an urban site in China. Atmospheric Research. 2007; 85(3-4): 310-337.
20. Ma CJ, Kim KH. Artificial and biological particles in the springtime atmosphere. Asian Journal of Atmospheric Environment. 2013; 7: 209-216.
21. Milionis AE, Davies TD. Regression and stochastic models for air pollution-I. Review, comments and suggestions. Atmospheric Environment. 1994; 28(17): 2801-2810.
22. Ulke AG, Mazzeo NA. Climatological aspects of the daytime mixing height in Buenos Aires City, Argentina. Atmospheric Environment. 1998; 32: 1615-1622.
23. Han S, Bian H, Feng Y, Liu A, Li X, Zeng F, Zhang X. Analysis of the Relationship between O_3, NO and NO_2 in Tianjin, China. Aerosol and Air Quality Research. 2011; 11: 128-139.
24. Leighton PA. Photochemistry of air pollution. Academic Press, New York. 1961; pp.300.
25. Chatterjee S, Hadi AS. Regression analysis by example. Wiley & Sons, New York. 2006.
26. Hossain MG, Sabiruzzaman M, Islam S, Ohtsuki F, Lestrel PE. Effect of craniofacial measures on the cephalic index of Japanese adult female students. Anthropological Science. 2010; 118(2): 117-121.