British Journal of Applied Science & Technology
14(3): 1-8, 2016, Article no.BJAST.23910
ISSN: 2231-0843, NLM ID: 101664541

SCIENCEDOMAIN *international*
www.sciencedomain.org

Modeling of Air Pollutants and Ozone Concentration by Using Multivariate Analysis: Case Study of Dimitrovgrad, Bulgaria

S. G. Gocheva-Ilieva[1*], A. V. Ivanov[1] and I. P. Iliev[2]

[1]Department of Applied Mathematics and Modeling, Faculty of Mathematics and Informatics, Plovdiv University 'Paisii Hilendarski', 24 Tsar Asen St., 4000 Plovdiv, Bulgaria.
[2]Department of Physics, Technical University-Sofia, Branch Plovdiv, 25 Tsanko Diustabanov St., 4000 Plovdiv, Bulgaria.

Authors' contributions

This work was carried out in collaboration between all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJAST/2016/23910
Editor(s):
(1) Xu Jianhua, Department of Geography, East China Normal University, China.
Reviewers:
(1) Amaury De Souza, Federal University of Mato Grosso do Sul, Campo Grande, Brazil.
(2) Mietek Szyszkowic, Population Studies Division, Health Canada, Canada.
Complete Peer review History: http://sciencedomain.org/review-history/13095

Original Research Article

Received 28th December 2015
Accepted 18th January 2016
Published 28th January 2016

ABSTRACT

Air pollution is one of the key problems in urban areas and its investigation is vital both for people's health and for the environment as a whole. In particular, ground ozone is a secondary air pollutant with concentrations dependent mainly on changes in the levels of other pollutants and meteorological conditions within a given region. This paper presents a statistical study based on multivariate analysis of hourly data on 9 air pollutants and 6 meteorological variables in the town of Dimitrovgrad, Bulgaria over a period of 7 years and 3 months. Yeo-Johnson power transformation is applied to each air pollutant variable to improve normality of the time series. The dominant patterns in the considered data are examined with the help of Principal Component Analysis (PCA) and factor analysis. Furthermore, particular focus is given for determining the concentration levels of ozone in relation to the other air pollutants and/or 6 meteorological time series using principal component regression (PCR). The good fitting of the obtained models with coefficients of determination R2 over 78% is obtained. An example of using the model to forecast the

**Corresponding author: E-mail: snow@uni-plovdiv.bg;*

concentrations of ozone for 24 hours ahead is given. The obtained results could be used as an assessment in all analyses of the air quality of the town Dimitrovgrad, including the official reports of the Environmental Agency and also as an independent alternative to the official alerting systems.

Keywords: Air pollution modeling; ozone concentration; principal component analysis; principal component regression; Yeo-Johnson power transformation.

1. INTRODUCTION

The monitoring, investigation and control of ambient air quality is a topical issue, very important for preserving the human health and the environment. Directives and regulatory restrictions are established in all European countries and worldwide for permissible concentrations of air pollutants [1-3]. In Bulgaria, 12 types of pollutants are systematically monitored by more than 36 automated stations run by the Executive Environment Agency which manages and coordinates activities related to the control and environmental protection of the country. The availability of a huge amount of collected data allows their statistical examination and makes it possible to find significant patterns, as well as dependencies within the data enabling the prediction of future states.

In literature, numerous similar investigations have been carried out in recent years, applying various mathematical methods supported by different statistical software. Multivariate statistical analysis is a well-established approach in this field. Recent studies where Principal Component Analysis (PCA), factor analysis, regression analysis and other methods were applied include for example [4-8]. Other popular techniques are the stochastic Box-Jenkins ARIMA methods [9], neural networks [10], etc.

The objective of this paper is to investigate the relationships between a large number of air pollutants in Dimitrovgrad, a town in Bulgaria, over an extended period of 7 years and 3 months. The specific goals of the study are: (i) Establishing the existence and determining the type of correlations between the investigated air pollutants; (ii) Defining patterns and possible groups of pollutants which are closely related, and classifying the pollutants; (iii) Obtaining multivariate regression models which describe the changes in ozone pollution levels in relation to the other pollutants and/or meteorological data; (iv) Applying the model for a short-time forecast.

2. MATERIALS AND METHODS

2.1 Study Area and Data Description

We examine air quality in the town of Dimitrovgrad, a typical urban region in South-central Bulgaria, 220 km away from the capital city of Sofia. The town is located in the Thracian valley on the banks of the river Maritza at an altitude of 125 m above sea level. It has about 40 000 inhabitants.

The study was carried out based on hourly data about the concentration of air pollutants between 1st January 2007 and 7 March 2014. Measurements were taken by an automated monitoring station in the town run by the official Executive Environment Agency.

The following 9 air pollutants are considered: nitrogen oxides (NOx, ppb); nitrogen dioxide (NO2, $\mu g/m^3$), nitrogen oxide (NO, $\mu g/m^3$), ozone (O3, $\mu g/m^3$), carbon monoxide (CO, $\mu g/m^3$), sulphur dioxide (SO2, $\mu g/m^3$), hydrogen sulfide (H2S, $\mu g/m^3$), ammonia, or azane, a compound of nitrogen and hydrogen (NH3, $\mu g/m^3$), particulate matter with diameter of 10 micrometres or less (PM10, $\mu g/m^3$).

Basic descriptive statistics of the data are given in Table 1, where *n* means the initial number of observations (number of hours). From Table 1, we have to note that although the mean value of ozone is not very high, the examined data indicate some values exceeding systematically the permissible limits established by the health regulations as set out in [1-3]. This is the reason to model additionally this particular secondary air pollutant.

The following 6 meteorological variables are also used: wind speed (WS, m/s), wind direction (SIGMA, degree), air humidity (HUMIDITY, %), air temperature (TEMP, ℃), sun radiation (GSR, W/m^2) and atmospheric pressure (PRESSURE, mbar). To note, that in our data SIGMA is not a circular variable and was preliminary transformed to non-circular one.

Table 1. Descriptive statistics of the observed air pollutants*

Variable	*n*	Minimum	Maximum	Mean	Std. deviation	Skewness	Kurtosis
O3	60236	0.00	1094	49.75	35.353	3.674	77.472
NOx	61612	0.00	468	14.63	23.318	5.676	48.422
NO	60621	0.00	451	6.47	21.323	7.541	77.478
NO2	61013	0.00	212	18.42	16.739	2.248	7.635
CO	58110	0.00	10	0.53	0.719	3.665	20.947
SO2	61641	0.00	1662	31.40	60.123	6.388	73.867
H2S	60570	0.00	0.06	0.003	0.003	5.141	54.114
NH3	60570	0.00	0.16	0.003	0.004	6.724	144.895
PM10	61261	0.03	896	57.08	57.165	3.738	22.080

**Std. error of Skewness for all variables is 0.010; Std. error of Kurtosis for all variables is 0.020*

2.2 Data Transformation

In principle, applying multivariate analysis requires an assumption for normal or close to normal distribution of participating variables, as well as other assumptions [11,12]. As shown in Table 1, the Skewness and the Kurtosis of the considered data are significantly different from zero, therefore, the normality condition is strongly violated. In such cases, it is advisable to improve the distribution by a suitable preliminary data transformation [9,12,13]. Firstly, the variables of all air pollutants were standardized with zero mean and standard deviation 1, using the formula

$$z_i = \frac{x_i - \overline{x}}{s},\ i = 1, 2, ..., n\ .$$

Then, the power transformation of Yeo-Johnson [14] was applied by the formula:

$$\psi_{YJ}(\lambda, x) = \begin{cases} \left\{(x+1)^{\lambda} - 1\right\}/\lambda & x \geq 0,\ \lambda \neq 0 \\ \log(x+1) & x \geq 0,\ \lambda = 0 \\ -\left\{(-x+1)^{2-\lambda} - 1\right\}/(2-\lambda) & x < 0,\ \lambda \neq 2 \\ -\log(-x+1) & x < 0,\ \lambda = 2 \end{cases}, \quad \lambda \in [-2, 2] \tag{1}$$

where $\lambda \in [-2, 2]$ is a parameter. The optimal values of λ were chosen to give the smallest possible value of the Jarque-Bera test of normality [15], defined as

$$JB = n\left[\frac{Sk^2}{6} + \frac{Ku^2}{24}\right], \tag{2}$$

where *n* is the number of cases, *Sk* and *Ku* are the Skewness and the excess Kurtosis of the transformed sample. The obtained transformed variables are denoted in square brackets.

Table 2 shows the obtained optimal values for λ and Skewness and Kurtosis of the transformed variables. With the small coefficients of skewness and kurtosis it could be concluded that the distributions of transformed variables are closed to the normal distribution.

2.3 Multivariate Statistical Methods Used in the Modeling Procedure

The examination of the correlation matrices for the initial and the transformed data of the 9 pollutants indicates the presence of high multicollinearities. To resolve this problem and to classify the data, we use the well-known PCA method [16,11]. PCA allows the extraction of linearly independent principal components (PCs), equal in number to the number of output variables. Factor analysis is also used to identify patterns in data using the main extracted PCs, which group together the variables and account for the greater part of the total sample variance. Missing values are treated listwise.

In order to model ozone concentrations, the Principal Component Regression (PCR) [11] is applied resulting in an explicit dependence of the type:

$$[O3] = b_0 + b_1 X_1 + ... + b_m X_m\ . \tag{3}$$

where $[O3]$ is the transformed ozone variable, $b_0, b_1, ..., b_m$ are regression coefficients and $X_1, X_2, .., X_m$ are the predictors (independent variables, in our case – the PCs) in the model. In (3) the PCs obtained from the transformed variables for air pollutants and/or meteorological variables will be used as predictors. This type of dependence corresponds to the chemical processes which lead to ozone formation in urban areas, taking into account the influence of ozone precursors and meteorological conditions [1].

Table 2. Descriptive statistics of the transformed air pollutants*

Transformed variable	λ	Skewness	Kurtosis	JB test
[O3]	0.44	-0.011	0.049	8
[NOx]	-1.5	0.187	-0.963	2801
[NO]	-2	1.400	0.913	22743
[NO2]	-0.15	0.167	-0.710	1614
[CO]	-0.65	0.245	-0.828	2432
[SO2]	-2	0.374	-1.092	4597
[H2S]	-0.15	0.021	0.025	6
[NH3]	-0.4	0.391	-0.670	2781
[PM10]	-0.6	0.057	-0.514	726

**Std. error of Skewness for all variables is 0.010; Std. error of Kurtosis for all variables is 0.020*

Calculations are performed using the IBM SPSS 22 statistical software.

3. RESULTS AND DISCUSSION

3.1 Results from Principal Component Analysis for Air Pollutant Variables

Further, we will work with both initial and transformed variables to compare results. Within the first step of PCA the correlation matrices were calculated as presented in Table 3 and Table 4. All columns contain coefficients over 0.3, with the largest correlation coefficient being that between [NOx] and [NO2], equal to 0.957. All correlation coefficients of [O3] (ozone) with other variables are negative, which corresponds to the nature of chemical reactions since ozone is formed from other pollutants, i.e. its concentration is inversely proportional to their own. As expected, the highest negative correlations are those with nitrogen oxides, nitrogen dioxide and nitrogen oxide.

The relatively high absolute values of correlation coefficients in Tables 3 and 4, and the small values of the determinants indicate the presence of high multicollinearity. This result is weaker in the second case.

An adequacy test was also performed for PCA and factor analysis, with the KMO test yielding a value of KMO=0.815 > 0.5 and Bartlett's test significance, equal to .000 [11]. This shows that the these analyses are adequate.

The next step of the PCA method is to generate PCs resulting from the 9 transformed variables. Table 5 shows the calculated eigenvalues and the distribution of total variance. From Table 5 it can be observed that the last eigenvalue is very small and could be ignored [11,16]. With the presence of multicollinearity the number of PCs is less than 8.

To classify the air pollutants and discover the dominant patterns in the dataset we perform factor analysis. Varimax rotation did not yield well differentiated components. For this reason we applied Promax rotation. The optimal factor solution with all 9 pollutants contains 7 factors, which account for 96.024% of total variance. The resulting rotated solution with 7 factors is given in Table 6. We have to add, that all the variance inflation factors (VIF) of the PCs are less than 2.85, which indicate that the obtained PCs do not correlate one with another. Table 6 clearly shows that all PCs are well differentiated. PC1 groups together [NO2], [Nox], and [O3]. All other variables are individual factors.

Table 3. Pearson correlation table of the initial variables of air pollutants*

Variable	O3	NOx	NO	NO2	CO	SO2	H2S	NH3	PM10
O3	1	-0.418	-0.317	-0.508	-0.348	0.011	-0.267	-0.163	-0.326
NOx		1	0.957	0.813	0.755	0.260	0.572	0.437	0.656
NO			1	0.611	0.731	0.208	0.553	0.374	0.586
NO2				1	0.600	0.293	0.454	0.447	0.620
CO					1	0.304	0.587	0.323	0.672
SO2						1	0.329	0.098	0.336
H2S							1	0.255	0.545
NH3								1	0.373
PM10									1

**Significance (1-tailed) for all correlation coefficients is P=.001 in exception of (SO2, O3), which is P=.007. Determinant = 5.56E-06*

Analogically to the previous analysis, after excluding ozone, we apply PCA and get 7 uncorrelated PCs, which are used in the subsequent analyses. These PCs account for 99.548% of the total variance. The loadings of the rotated matrix are shown in Table 7.

3.2 Results from Principal Component Regression

The next goal is to establish the explicit dependence between ozone and the other pollutants, with and without the meteorological data.

Due to the multicollinearity of the variables, the direct application of multiple linear regression to non-transformed or transformed variables is not recommended and give unsatisfactory results [11,12]. This can be overcome using a well-known technique, namely the extraction of principal components which are not mutually correlated. The obtained new variables can be used to find regression models. This mixed regression approach is known as Principal Component Regression (PCR) [11].

In previous subsection 3.1, after excluding ozone, we obtained 7 uncorrelated PCs, which are used in subsequent regression analyses. In addition to these variables the six meteorological variables are also included as predictors to obtain regression equations. PCR is performed using the Stepwise method in SPSS. The statistical significance is established at level $\alpha = 0.05$.

The obtained standardized regression equation using the 7 extracted PCs, according the PCA (see Table 7) has the form

$$\begin{aligned}[O3] = &-0.502PC1 - 0.092PC2 \\ &+ 0.054PC3 + 0.202PC4 \\ &- 0.077PC5 - 0.008PC6 \\ &- 0.288PC7\end{aligned} \quad (4)$$

The coefficient of determination of (4) is R^2=0.551. All coefficients, as well as the ANOVA of the model are statistically significant. The relative influence of ozone precursors on the examined data is defined. The results show the strongest influence is that of $PC1 = \{NO2, NOx\}$, $PC7 = \{NO\}$, followed by $PC4 = \{SO2\}$. The remaining pollutants have weaker influence.

As was mentioned above, it is well-known that ozone concentration is strongly dependent on meteorological conditions which influences chemical reactions leading to its formation [1]. The next model is derived using the six meteorological variables. The resulting standardized equation has the following form:

$$\begin{aligned}[O3] = &-0.410\,HUMIDITY \\ &+ 0.295TEMP + 0.260WS \\ &- 0.115\,SIGMA \\ &+ 0.098\,PRESSURE \\ &+ 0.068\,GSR\end{aligned} \quad (5)$$

The corresponding coefficient of determination is R^2=0.635. It becomes clear from (5) that as a whole the main contribution in the examined interaction is that of low air humidity, air temperature and wind speed.

Finally, all 7 PCs from Table 7 and also 6 meteorological predictors are used to simultaneously take into account the precursors and the meteorological data. The resulting standardized regression equation is

$$\begin{aligned}[O3] = &-0.262PC1 - 0.020PC2 + 0.035PC3 + 0.159PC4 - 0.009PC5 \\ &+ 0.006PC6 - 0.265PC7 - 0.345HUMIDITY + 0.161TEMP + 0.103WS \\ &- 0.082SIGMA + 0.066PRESSURE + 0.080GSR\end{aligned} \quad (6)$$

The coefficient of determination of model (6) is R^2=0.783. The dominant part of the equation is due to the $PC1 = \{NO2, NOx\}$, $PC7 = \{NO\}$, $HUMIDITY$, $TEMP$, and $PC4 = \{SO2\}$.

3.3 Using the Model for a Short-time Forecasting

To demonstrate the predictive ability of the model for the case of mixed model (6) we removed the last 24 observations (24 hours measurements) from the transformed ozone series [O3]. The repeated procedure gives the same model (6), with only the coefficient of $PC4$ equal to 0.158 instead of 0.159. The forecast for the ozone after retransformation and re-standardization is compared with the measured values in Fig. 1 for 24 hours ahead. It is observed a good correspondence between the two series with R^2=0.82.

Table 4. Pearson correlation table of the transformed variables of air pollutants*

Transformed variable	[O3]	[NOx]	[NO]	[NO2]	[CO]	[SO2}	[H2S]	[NH3]	[PM10]
[O3]	1	-0.700	-0.644	-0.635	-0.399	-0.096	-0.308	-0.247	-0.411
[NOx]		1	0.756	0.949	0.450	0.348	0.336	0.367	0.595
[NO]			1	0.640	0.502	0.246	0.357	0.340	0.522
[NO2]				1	0.413	0.370	0.309	0.344	0.573
[CO]					1	0.332	0.297	0.215	0.424
[SO2]						1	0.242	0.086	0.420
[H2S]							1	0.258	0.337
[NH3]								1	0.337
[PM10]									1

**Significance (1-tailed) for all correlation coefficients is P=.001. Determinant = 0.004*

Table 5. Total variance explained*

Component	Initial eigenvalues		
	Total	% of variance	Cumulative %
PC1	4.492	49.907	49.907
PC2	1.043	11.586	61.493
PC3	0.891	9.897	71.390
PC4	0.760	8.442	79.832
PC5	0.655	7.282	87.114
PC6	0.462	5.136	92.250
PC7	0.340	3.774	96.024
PC8	0.322	3.581	99.605
PC9	0.036	0.395	100.000

**Extraction Method: Principal Component Analysis*

Table 6. Principal component pattern matrix for 9 air pollutants*

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
[NO2]	**1.052**	0.083	0.035	-0.009	-0.020	0.020	-0.175
[NOx]	**0.897**	0.072	0.039	-0.023	-0025	0.023	0.078
[O3]	**-0.636**	0.239	0.112	-0.053	-0.072	0.057	-0.358
[SO2]	0.020	**0.974**	-0.041	0.018	0.026	-0.020	0.077
[NH3]	0.011	-0.043	**0.985**	0.009	0.013	-0.010	0.037
[CO]	-0.001	0.020	0.009	**0.998**	-0.006	0.005	-0.013
[H2S]	-0.005	0.027	0.013	-0.005	**1.000**	0.007	-0.023
[PM10]	0.021	-0.021	-0.010	0.006	0.007	**0.985**	0.021
[NO]	0.053	0.113	0.053	-0.018	-0.032	0.029	**0.916**

**Extraction Method: Principal Component Analysis. Rotation Method: Promax with Kaiser Normalization. Rotation converged in 7 iterations*

Table 7. Factor analysis pattern matrix with 7 factors for 8 air pollutants*

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
[NO2]	**1.052**	0.005	0.000	0.009	0.001	0.001	-0.102
[NOx]	**0.890**	-0.005	0.001	-0.010	0.000	0.005	0.145
[CO]	0.003	**0.997**	0.000	0.000	0.000	0.001	0.003
[NH3]	0.001	0.000	**0.999**	0.000	0.000	0.000	0.000
[SO2]	0.002	0.000	0.000	**0.998**	0.000	0.001	0.002
[H2S]	0.001	0.000	0.000	0.000	**1.000**	0.000	0.000
[PM10]	0.014	0.001	0.001	0.002	0.000	**0.990**	0.001
[NO]	0.048	0.005	0.000	0.002	0.001	0.001	**0.963**

**Extraction Method: Principal Component Analysis. Rotation Method: Promax with Kaiser Normalization. Rotation converged in 6 iterations*

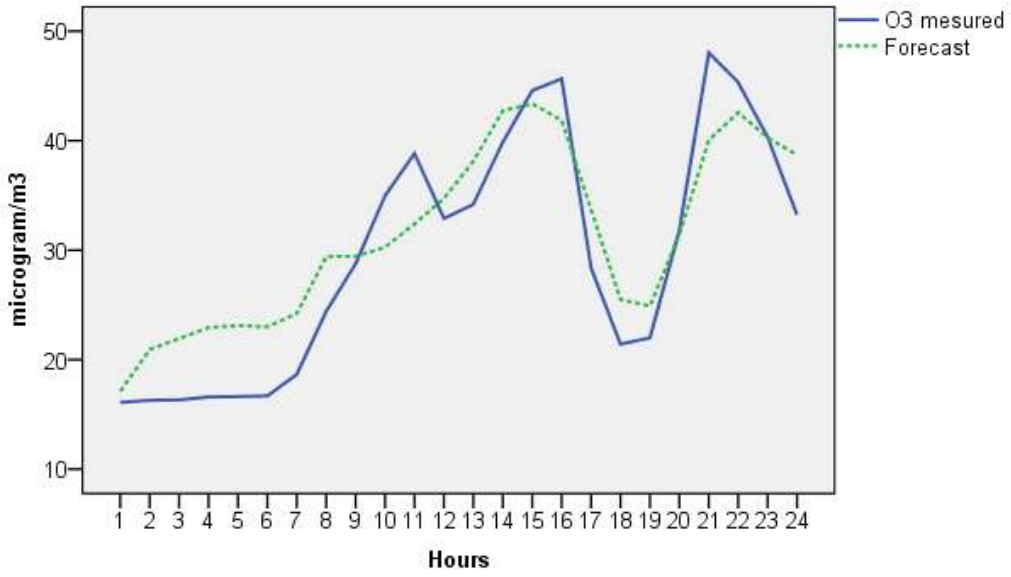

Fig. 1. Comparison between the measured and forecasted values of ozone (O3)

4. CONCLUSION

With the help of PCA and the correlation matrix, 8 PCs were derived and classified according to their relative contribution to the total level of air pollution. Nitrogen oxides (NOx, NO2, NO) play a dominant part. This result is explained by the presence of a large nitrogen fertilizer production plant in the town of Dimitrovgrad, which is the main source of industrial air pollution in the city, along with road traffic.

The obtained regression equation (6) shows that the combined contribution of ozone precursors and meteorological data account for up to 78% of the concentration of this pollutant. The other two equations (4) and (5) show lower values, but these are significant for the clear differentiation of the relative participation of each pollutant and each meteorological factor in overall ozone concentration. Example of using the mixed model (6) forecasting 24 hours ahead ozone concentrations shows good agreement with measured values.

The obtained results could be used as an assessment in all analyses of the air quality of the town Dimitrovgrad, due to the long period of investigation of over 7 years, including the official reports of the Environmental Agency. The results could be also used as an alternative to the official alerting, controlling and monitoring systems, by analyzing the detailed influence of the 9 primary pollutants and 6 meteorological variables in the overall air pollution and the evaluation of ozone concentration.

ACKNOWLEDGEMENT

The paper is partially supported by the Plovdiv University NPD, grant NI15-FMI-004.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. European Commission. Environment. Air Quality Standards; 2013. Available:http://ec.europa.eu/environment/air/quality/standards.htm (Accessed 20 December 2015)
2. European Parliament. Directive 2008/50/EC of the European Parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal of the European Union. L. 2008;152/1.
3. European Environmental Agency. Air quality in Europe- 2014 report. 2014. Available:http://www.eea.europa.eu/publications/air-quality-in-europe-2014 (Accessed 20 December 2015)
4. Kaplunovsky AS. Factor analysis in environmental studies. HAIT Journal Sci. Eng. B. 2005;2(1-2):54-94.

5. Shi JP, Harrison RM. Regression modelling of hourly NOx and NO2 concentrations in urban air in London. Atmos. Environ. 1997;31:4081-4094.
6. Lengyel A, Héberger K, Paksy L, Bánhidi O, Rajkó R. Prediction of ozone concentration in ambient air using multivariate methods. Chemosphere. 2004;57:889-896.
7. Gvozdić V, Kovač-Andrić E, Brana J. Influence of meteorological factors NO2, SO2, CO and PM10 on the concentration of O3 in the urban atmosphere of Eastern Croatia. Environ. Model. Asses. 2011;16 (5):491-501.
8. Chan TW, Mozurkewich M. Application of absolute principal component analysis to size distribution data: Identification of particle origins. Atmos. Chem. Phys. 2007;7:887-897.
9. Box GEP, Jenkins GM. Time series analysis, forecasting and control. Revized ed. San Francisco: Holden Day; 1976.
10. Azid A, Juahir H, Toriman ME et al. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. Water, Air, & Soil Poll. 2014;225:2063-2074.
11. Izenman A. Modern multivariate statistical techniques. New York: Springer; 2008.
12. Wilks DS. Statistical methods in the atmospheric sciences. 3th ed. Amsterdam: Elsevier; 2011.
13. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom. Intell. Lab. Syst. 1987;2:37-52.
14. Yeo IK, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000;87(4):954-959.
15. Jarque C, Bera A. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ. Lett. 1980;6:255-259.
16. Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer; 2002.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/13095