Advances in Environmental Research, *Vol. 4, No. 4* (*2015*) *233-246*
DOI: http://dx.doi.org/10.12989/aer.2015.4.4.233

Air pollution study using factor analysis and univariate Box-Jenkins modeling for the northwest of Tehran

Gholamreza Asadollahfardi [*1], Mehran Zamanian [1], Mohsen Mirmohammadi [2], Mohsen Asadi [1] and Fatemeh Izadi Tameh [1]

[1] *Department of Civil Engineering, Kharazmi University, 43 Mofateh Ave, Tehran, Iran*
[2] *Department of Engineering, University of Tehran, Tehran, Iran*

(*Received July 01, 2015, Revised December 01, 2015, Accepted December 04, 2015*)

Abstract. High amounts of air pollution in crowded urban areas are always considered as one of the major environmental challenges especially in developing countries. Despite the errors in air pollution prediction, the forecasting of future data helps air quality management make decisions promptly and properly. We studied the air quality of the Aqdasiyeh location in Tehran using factor analysis and the Box-Jenkins time series methods. The Air Quality Control Company (AQCC) of the Municipality of Tehran monitors seven daily air quality parameters, including carbon monoxide (CO), Nitrogen Monoxide (NO), Nitrogen dioxide (NO_2), NO_X, ozone (O_3), particulate matter (PM_{10}) and sulfur dioxide (SO_2). We applied the AQCC data for our study. According to the results of the factor analysis, the air quality parameters were divided into two factors. The first factor included CO, NO_2, NO, NO_x, and O_3, and the second was SO_2 and PM_{10}. Subsequently, the Box- Jenkins time series was applied to the two mentioned factors. The results of the statistical testing and comparison of the factor data with the predicted data indicated Auto Regressive Integrated Moving Average (0, 0, 1) was appropriate for the first factor, and ARIMA (1, 0, 1) was proper for the second one. The coefficient of determination between the factor data and the predicted data for both models were 0.98 and 0.983 which may indicate the accuracy of the models. The application of these methods could be beneficial for the reduction of developing numbers of mathematical modeling.

Keywords: air quality; Aghdaseyah; Tehran; Iran

1. Introduction

The rapid growth of urbanization and uncontrolled urban development has created environmental problems for urban dwellers. One of the major difficulties of living in big cities now, especially in developing countries, and in small towns in the near future is air pollution. Air pollution is a permanent threat to densely populated cities and leads to deleterious effects on public health and will result in great economic losses. Urbanization and urban development together with a sharp increase in population growth, industrial development and indiscriminate use of fossil fuels, have drastically increased air pollution. A large amount of contaminants incompatible with the normal mechanisms will be released into the air. Air pollution means the

[*] Corresponding author, Research Associate, E-mail: reachquadri@yahoo.com

http://www.techno-press.org/?journal=aer&subpage=7 ISSN: 2234-1722 (Print), 2234-1730 (Online)

presence of one, several, or a mixture of different pollutants in ambient air to the extent that would be harmful to humans or cause harmful effects to animals, plants and properties; or which may induce immeasurable effects on humans, animals, crops and synthetic materials (Erfanmanesh and Afuni 2006). Therefore, air pollution control is very vital. In this regard, availability of accurate and plentiful data and interpretation are an imperative tool which can help the air quality management to decide properly. Several methods are available to analyze air quality data such as statistical and deterministic methods. We applied factor analysis and the Box-Jenkins time series to air quality data. The first reason for using both methods is when we apply the invert Box-Jenkins time series, developing a model for each air quality parameter is necessary. It consumes a lot of times for experts to develop several models. Secondly, several air quality parameters depend on each other. We grouped the air quality parameters to a few factors using factor analysis and then applied the Box-Jenkins time series of a few factors instead of a lot of air quality parameters. These types of air quality interpretations are rare; however, several studies have used the two methods separately.

Nowadays, Factor analysis is a statistical analysis technique that is used in various disciplines such as psychology, sociology, management, geography, and urban planning (Harvey and Todd 1983). The analysis of time series rapidly developed in a practical and theoretical way after the original work of Box and Jenkins (Box and Jenkins 1970). Polydoras *et al.* (1998) compared the Weather forecasts for the amount of CO and SO_2 in three different locations in the City of Athens using both distribute and Box nonlinear statistic models and predicted the maximum and the critical amount of pollutants. Sharma and Khare (2000) evaluated the Box-Jenkins time series based on the multivariate statistic models in predicting the CO concentration at a large crossroads in Delhi. Rodriguez-Rajo *et al.* (2005) developed pollen concentrations predictive models based on the ARIMA model in the North West of Spain. They found that the method of time series is an appropriate method for estimating the short-term effects of tree pollen in the atmosphere. Abbaspour and Rahmani (2005) and Sader Mousavi and Rahimi (2008) conducted some research about predicting carbon monoxide concentrations in the air of Tehran and Tabriz, respectively. Sharma *et al.* (2009) analyzed the air pollution parameters including SO_2, NO_2 and suspended particles using the Box-Jenkins method in Delhi India for a maximum of 24 hours per month. Vaseghi and Zibaie (2008) conducted air pollution predictions in Shiraz. Different methods of regression and non-regression were evaluated and the ARMA model was selected to predict the air pollution index based on the measured error rates. Kumar and Jain (2010) used ARIMA to predict CO, NO, NO_2 and O_3 in an urban traffic site in Delhi, India. They reported the suitability performance of ARIMA in short term air quality predictions.

Nakhaei *et al.* (2011) developed two ARIMA models with different coefficients for CO in the Towhid Eastern and Western tunnels, using Box-Jenkins time series model for the prediction of carbon monoxide. Sami *et al.* (2012) used seasonal ARIMA (SARIMA) to anticipate the rate of dust fall in Quetta, Pakistan. They reported the appropriate performance of SARIMA. Gocheva-Ilieva *et al.* (2014) applied time series analysis, including BOX-Jenkins approach and factor analysis to forecast air pollution in Blagoevgrad, Bulgaria. They considered NO, NO_2, NO_x, PM_{10}, SO_2 and ground level O_3, which were monitored for one year. They classified the pollutants in three factors and determined the factor to overall pollution. The results had a good agreement with the field monitoring. Kumar (2015) combined a signal extraction method SSA (Singular Spectrum Analysis) with ARIMA to predict the daily maximum ambient O_3 concentrations. They indicated the aptness of the SSA-ARIMA in predicting short term air quality.

The study area is located in northeast Tehran and the geographic coordinates of the city are

51°,2′ and 51°,36′ East longitude and 35°,34′ and 35°,50′ North latitude (Fig. 1). The elevations of Tehran are 2000, 1200 and 1050 meters in the north, in the center and in the south, respectively. The north and east of the city is surrounded by the Alborz Mountains. The main sources of precipitation are the Mediterranean and Atlantic winds that blow from the West, and the Alborz mountains act as a barrier to prevent the penetration of air masses. Tehran is also located in arid and semi-arid regions. The temperature variations are between 40 Celsius in summer and -5 Celsius in winter. The annual rainfall is about 250 millimeters.

To study air pollution emissions using deterministic modelling, the number and kinds of emission sources and environmental features such as wind direction and speed, temperature, stability of the atmosphere, topography of the region, solar elevation angle and surface heat flux need to be available. However, when few numbers of parameters such as air quality data with time are available using stochastic or probability modelling are applicable. In our study, the concentration of the pollutants was collected randomly. In the other words, three grab samples for each pollutant was collected. We referenced the mean value of the three samples in our study. For each air quality parameter a minimum 60, hourly or daily data are necessary and data should be measured in virtually equal time period. Because of lack of data for different mentioned parameters in our study area and air quality data was available which effect on air quality, we applied Box-Jenking time series models, which are a stochastic approach, to study air pollution of the north west of Tehran to predict air quality situation. Since, many numbers of air quality parameters have existed, we used factor analysis to reduce air quality parameters to a number of factors. Each factor includes number of air quality parameters with having similar characteristic. Instead of developing several numbers of univariate Box-Jenkin time series model for air quality parameters, we developed a model for each factor.

The first objective of the study was to apply the factor analysis technique to air quality parameters including carbon CO, NO_2, NO, NO_X, O_3, PM_{10} and SO_2 data. The second aim was to develop the Box-Jenkins time series model using the factors data which were the results of the first objective.

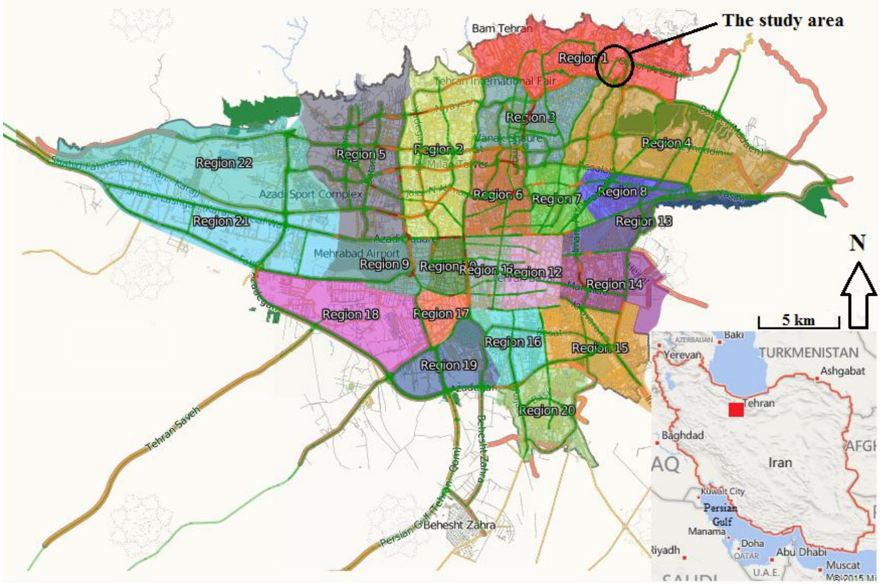


Fig. 1 The location of the study area

2. Materials and methods

Forecasting is a key element in management decisions, because every decision has its own consequences. Time series refers to the set of observations which are arranged according to time and at even intervals. Although the data may be sorted according to some other factors such as distance. The order of observations is an important factor in using a time series method, because the inherent nature of a time series is highly correlated with the dependence or interdependence of the observations (Tobias *et al*. 2004).

Factor analysis includes analysis of several variables simultaneously. Air quality parameters, which contain a high correlation coefficient between them, are placed into one single group. Factor analysis is applied to recognize the correct structure of collecting data and to identify the most significant factors contributing to the data structure (Buckley and Winter 1992, Padro *et al*. 1993). The Factor is a new independent variable, which cannot be directly measured and is in fact, immeasurable. Factor analysis is also used in establishing associations between parameters so that the number of parameters measured can be lessened. Eq. (1) describes the factor (Johnson and Wichem 2007).

$$F_j = \sum_{i=1}^{P} W_{ji} X_i = W_{j1}X_1 + W_{j2}X_2 + \cdots + W_{jP}X_P \tag{1}$$

Where F_j is factor; W_{ji} is factor *number*, and p is the number of variables. x_i is observed variables.

By using factor analysis, we manage air pollution parameters in a few groups which in the group (Factor) all parameters depend highly on each other. After that, instead of developing a model for each parameter separately, we developed a box Jenkins model for each group (Factor). Factor data are a combination of observed data according to Eq. (1).

Procedure of Factor Analysis: Implementation Performance of factor analysis involves the following steps (Asadollahfardi *et al*. 2012):

- Collecting all of the data and computing of a correlation coefficient matrix between all of the air quality parameters.
- Determining the factor loading according to correlation coefficients.
- Rotating the factors for simplicity and understandability of factor analysis.

For further information, refer to Mulaik (2009) and Brown (2014).

Box-Jenkins Methodology for Time Series Modeling: Decomposition of time series data into its components, while being instructive and revealing, is a difficult job. Moreover, it causes greater errors by accumulation of component errors (Asadollahfardi 2002). To avoid these problems, Box and Jenkins (1976) developed a new methodology, which in essence, performs the same job. We applied some transformations to remove simple and seasonal differences, trends, seasonal and cyclical components presented in the data. Then, a family of models is entertained for the transformed data, which is expected to be as simple as possible. The following section briefly explains the non-stationary Box-Jenkins method.

Classification of Time Series Models: The behaviors of the sample autocorrelation function (SACF) and the sample partial autocorrelation function (SPACF) are important in tentative identification of stationary time series models. For the values of a stationary time series *Zb*, Z_{b+1},..., Zn which may be the original time series values or the transformed time series values, SACF is defined as follows (Eq. (2)). The sample autocorrelation at lag *k* denoted by r_k is

$$r_k = \frac{\sum_{t=b}^{n-k}(z_t - z)(z_{t+k} - \overline{z})}{\sum_{t=b}^{n}(z_t - \overline{z})^2} \tag{2}$$

where

$$\overline{z} = \frac{\sum_{t=b}^{n} z_t}{(n - b + 1)} \tag{3}$$

Considering r_k is a function of lag *k*, for k = 1, 2,..., *K* is called the sample autocorrelation function (SACF). This quantity measures the linear relationship between a time series observation, separated by a lag of *k* time units. The r_k is a coefficient of correlation and it is always between -1 and +1. The standard error of r_k is defined by (Eq. (4))

$$s_{r_k} = \left[\frac{1 + 2\sum_{j=1}^{k-1} r_j^2}{n - b + 1}\right]^{0.5}, \quad k = 1,2,\dots \tag{4}$$

The t_{r_k} statistic is then computed as (Eq. (5))

$$t_{r_k} = \frac{r_k}{s_{r_k}} \tag{5}$$

Which is used to test the significance of r_k, for *k* = 1, 2, ...

Plotting r_k against *k* provides the SACF. The behavior of this function is a key tool for identification of the stationary of a time series and its order. To employ the Box-Jenkins approach, one must examine and try to classify the behavior of the SACF.

The sample partial autocorrelation function (SPACF) is another important tool for identifying time series models. The sample partial autocorrelation at lag *k* is defined by (Eq. (6))

$$r_{kk} = \begin{cases} r_1, & k = 1 \\ \dfrac{\left(r_k - \sum_{j=1}^{k-i} r_{k-1,j} \times r_{k-j}\right)}{\left(1 - \sum_{j=1}^{k-i} r_{k-1,j} \times r_{k-j}\right)}, & k = 2,3,\dots \end{cases} \tag{6}$$

Where, $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1} r_{k-j}$, $j = 1,2,\dots,k-1$

The standard error of r_{kk} is defined as (Eq. (7))

$$s_{r_k} = \left[\frac{1}{n - b + 1}\right]^{0.5} \tag{7}$$

The following equation indicates the student's t_{kk}-statistic

$$t_{r_{kk}} = \frac{r_{kk}}{s_{r_{kk}}} \tag{8}$$

The precise interpretation of the SPACF at lag *k* is rather complicated. However, this quantity can intuitively be thought of as the sample autocorrelation of time series observations, separated by a lag of *k* time units, with the effects of the intervening observations eliminated.

For a time series consisting of Z_b, Z_{b+1}, ..., Z_n, where Z_t is the original or transformed value of a time series, an autoregressive model of order p, AR(p), is defined as

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_r Z_{t-p} + a_t \tag{9}$$

Where ϕ_1, ..., ϕ_r are fixed coefficients and a_t, t = 1, 2, ..., n are independent random variables with zero mean and constant variance σ_a^2. They are usually assumed as normally distributed. Using the backward shift operator B, Eq. (9) can be written as

$$\phi p(B) Z_t = a_t \tag{10}$$

Where ϕ_p (B) =1 – $\phi_1 B$ – ... – $\phi_p B_p$; And BZ_t = $Z_{t\text{-}1}$,..., BpZ_t = $Z_{t\text{-}p}$.

A moving average model of order q, MA(q), is represented as

$$Z_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{11}$$

Or employing the backward shift operator B

$$Z_t = \theta q(B) a_t \tag{12}$$

where

$$\theta q(B) = 1 - \theta_1 B - \cdots - \theta_q B_q \tag{13}$$

The general non-seasonal autoregressive moving average model of order (p, q) is

$$Z_t = \delta + \phi_1 Z_{t-1} + \cdots + \phi p Z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{14}$$

This model utilizes a constant term δ. It has an autoregressive part which expresses the current value Z_t as a function of past values $Z_{t\text{-}1}$, $Z_{t\text{-}2}$, ..., $Z_{t\text{-}p}$ with unknown coefficients (parameters) ϕ_1, ..., ϕ_p. In addition, The model has a moving average part which is represented by a_t, $a_{t\text{-}1}$, ..., $a_{t\text{-}q}$ with unknown fixed parameters θ_1, ..., θ_q. The variable Z_t is also considered as a function of a random variable, as, $a_{t\text{-}1}$, ..., $a_{t\text{-}q}$.

In Eq. (14), the constant term δ can be shown as equal to $\mu\phi p(B)$, where μ is the mean of the stationary time series Z_t. In concise notation, Eq. (14) is presented as

$$\phi_p(B) zt = \delta + \theta q(B) a_t \tag{15}$$

The statistical tests are available, which can be used to decide whether to include δ in the model.If the stationary time series Z_b, Z_{b+1}, ..., Z_n is in the original series, then assuming μ is equal to zero, this implies that these original time series values are fluctuating around a zero mean, whereas $\mu \neq 0$ implies that these original values are fluctuating around a non-zero mean. In such a case one can use $(Z_t - \bar{Z})$ in place of Z_t. Then δ can be removed from the model. If the stationary time series Z_b, Z_{b+1}, ..., Z_n are different from those of the original time series values, where μ is not assumed to be zero, it can be assumed that a deterministic trend exists in those original values. Here the deterministic trend refers to a tendency of the original values to move persistently upward (if $\delta > 0$) or downward (if $\delta < 0$). If a time series does not exhibit a deterministic trend, then any trend (or failure of the series to fluctuate around a central value) is stochastic. The stochastic trend is more realistic in practical situations since it does not dictate a certain path to be taken by the future values (Asadollahfardi 2014).

The general seasonal autoregressive integrated moving average model of order (*P*, *Q*, *D*, *d*, *p*, *q*) is

$$\emptyset_P(B)\emptyset_P(B^L)\nabla_L^D\nabla^d y_t^* = \delta + \theta_q(B)\theta_q(B^L)\alpha_t \tag{16}$$

where

$$\emptyset_P(B) = (1 - Q_1B - Q_2B^2 - \cdots - Q_PB^P) \tag{17}$$

$$\emptyset_P(B^L) = (1 - \emptyset_{1,L}B^L - \emptyset_{2,L}B^{2L} - \cdots - \emptyset_{P,L}B^{PL} \tag{18}$$

$$\theta_q(B) = (1 - \theta_1B - \theta_2B^2 - \cdots - \theta_qB^q \tag{19}$$

$$\theta_Q(B^L) = \left(1 - \theta_{1,L}B^L - \cdots - \theta_{Q,L}B^{QL}\right) \tag{20}$$

Where $\emptyset_P(B)$, $\emptyset_P(B^L)$, $\theta_q(B)$ and $\theta_Q(B^L)$ are non-seasonal autoregressive operator of order *p*, seasonal autoregressive operator order *P*, non-seasonal moving average operator *q* and seasonal moving average operator order *Q*, respectively. The δ is a constant of the model which μ is the real meaning of stationary time series. The δ can be calculated by the Eq. (21)

$$\delta = \mu\emptyset_P(B)\emptyset_P(B^L) \tag{21}$$

B is the backward shift operator $B^k y_t = y_{t-k}$; ∇^d is equals to the backward difference operator, and $\emptyset\emptyset_1, \emptyset_2, \dots, \emptyset_P, \emptyset_{1,L}, \emptyset_{2,L}, \dots, \emptyset_{P,L}, \theta_1, \theta_2, \dots, \theta_q, \theta_{1,L}, \theta_{2,L}, \dots, \theta_{Q,L}$ are autoregressive and moving average orders which are unknown .We need to estimate using the sample data; $\alpha_t$is a random variable with mean zero and constant variance. The a_t is Independent and represent random error or random shocks.

The Box-Jenkins consists of four basic steps, in which Table 1 indicates the steps. For more details on the Box-Jenkins model structure and forecasting, refer to Box and Jenkins (1976) and Box *et al*. (2008).

Stages of Box-Jenkins modeling (Asadollahfardi *et al*. 2012)

1. Check the data for normality:
   a. No transformation.
   b. Square root transformation.
   c. Logarithmic transformation.
   d. Power transformation
2. Identification:
   a. Plot of the transformed series**.**
   b. Autocorrelation function (ACF)**.**
   c. Partial autocorrelation function (PACF)
3. Estimation
   a. Maximum likelihood estimate (MLE) for the model Parameters (Ansley algorithm).
4. Diagnostic checks
   a. Over fitting
   b. Examination of the residuals (modified Portmanteau test)
5. Model structure, selection criteria
   a. The Akaike Information Criterion (AIC) criteria
   b. PP criteria
   c. Bayesian criterion (BIC) criteria

To analyze, both factor analysis and Box-Jenkins time series, the SPSS software was applied. The Air Quality Control Company (AQCC) of Tehran Municipality monitors daily air quality data, including carbon monoxide (CO), Nitrogen Monoxide (NO), Nitrogen dioxide (NO_2), NO_X, ozone (O_3), particulate matter (PM_{10}) and sulfur dioxide (SO_2). The data used in our study were collected by the AQCC in Aghdaseyeh location in Tehran between late March 2012 and in late January 2013.

3. Results and discussion

Table 1 indicates the statistical summary of applied data in this study. These mentioned parameters in Table 1 were available for the study area. Data for THC and non-methane hydrocarbon compound (NMHC) and CH_4 was not available.

Table 2 indicates the correlation coefficient matrix for the observed air quality parameter data at the monitoring stations. Some of the parameters are positively correlated to each other, such as the correlation coefficient between NO_X and CO, r = 0.814, NO_X and NO, r = 0.993, PM_{10} and SO_2, r = 0.646, NO_X and SO_2, r = 0.377, NO and SO_2, r = 0.352, SO2 and CO, r = 0.346, SO_2 and NO2, r = 0.356 and SO2 and NO, r = 0.352. Some other air quality parameters are negatively correlated. Tabachnick and Fidell (1996) stated that if all of the correlation between parameters are less than 0.3 the factor analysis is not useable. We attempted to find out the dependent

Table 1 The statistical summary of air quality data in Aghdaseyeh located in Tehran

Air quality parameters	CO mg/l	NO_2 mg/l	NO mg/l	NO_X mg/l	O_3 mg/l	PM_{10} mg/l	SO_2 mg/l
Number	282	280	280	280	265	232	282
Mean	3.07	30.7	74.79	105.9	12.26	79.144	42.67
Standard error	0.07	0.45	3.47	3.66	0.50	3.3578	1.25
Mode	2.93	30.0	61.13	93.1	12.0	71.330	40.0
Standard deviation	1.17	7.54	58.06	61.2	8.23	51.145	21.1
Variance	1.39	56.9	3371.5	3752.8	67.8	2615.8	442.9
Range	8.20	40.2	244.3	278.3	68.5	385.68	163.3
Minimum	0.06	13.3	4.7	24.8	1.54	11.03	7.6
Maximum	8.26	53.6	249	303.1	70.04	396.71	170.9

Table 2 Correlation coefficient matrix between the air quality parameters

Parameters	CO	NO_2	NO	NO_X	O_3	PM_{10}	SO_2
CO	1.000	0.370	0.810	0.814	-0.282	0.254	0.346
NO_2	0.370	1.000	0.368	0.472	-0.071	0.229	0.356
NO	0.810	0.368	1.000	0.993	-0.203	0.259	0.352
NO_X	0.814	0.472	0.993	1.000	-0.201	0.274	0.377
O_3	-0.282	-0.071	-0.203	-0.201	1.000	-0.053	-0.078
PM_{10}	0.254	0.229	0.259	0.274	-0.053	1.000	0.646
SO_2	0.346	0.356	0.352	0.377	-0.078	0.646	1.000

parameters. Afterwards, the dependent parameters combined into a few groups. Then we applied the Box-Jenkins time series model to a few groups (factors) instead of all parameters. Other parameters have a much smaller correlation coefficient, which means that we must disregard them and suggests that they may affect the air quality independent from each other. As presented in Table 2, many of the correlation coefficients are above 0.3 which means the factor analysis is necessary. However, the correlation coefficient of O_3 is a minus value. Because ozone, as a secondary pollution concentration, is emanated from the others, and its value is evidently inversely proportional to their concentrations. Cryer (1986) also expressed that for confirmation of using factor analysis, the values of the KMO and the Bartlett's Test should be about 0.6 and less than 0.05, respectively. The value of the Kaiser-Meyer-Olkin (KMO) for the existing data was equal to 0.549 and the result of Bartlett's Test was significant ($a = 0.05 > p$-value = 0. 0).

One of the methods to figure out the number of factors is the estimation of eigenvalue and factor loadings for the correlation matrix, and each eigenvalue corresponded to an eigenvector that identifies the group of air quality parameters that are most highly correlated among them. The first eignfactor accounted for greatest variation among the observed air quality parameters, while each following eigenfactor was orthogonal to all former factors, and provided incrementally smaller contributions to the overall descriptive ability of the model. The eigenvalue of the correlation matrix of this study is described in Table 3. Since, a lower eigenvalue may contribute only modestly to the descriptive ability of the air quality data, only the first few factors were selected. Methods are present to figure out the number of factors that need to be considered and the number of those that can be safely neglected (Browne 1968, Linn 1968, Tucker *et al*. 1969, Hakstian *et al*. 1982). The method of Kaiser Criterion, which retains just those factors with eigenvalue larger than one, is the most widely used technique (Kaiser 1960). As indicated in Table 3 the first two factors, which have an eigenvalue bigger than one were selected because the first two factors contain 68.1% of air quality variables. The obtained *R* squares are valid for the two factors, which account for 68% of total variation of the data. According to Table 3, although the third eigenvalue was close to one, we selected the first two factors. Because the eigenvalue was drawn descending and sudden changes occurred in two points related to the other points. This indicates confirmation of two factors (Tabachnick and Fidell 1996). The disadvantage of the Kaiser's (1960) is often leads to over factoring and sometimes under factoring. However, a number of our variable was 7 and two factor could be used.

Factor loading reflects the correlation between the air quality parameters and the extracted factors. Factor loadings for the two retained eigenvalues are indicated in Table 4. Factor loading is described with rotation using the Quartimax method. The main function of the factor rotation

Table 3 Individual and cumulative eigenvalue of the air quality observations

Factors	*Eigenvalue*	Variance%	Cumulative variance%
1	3.478	49.68	49.6
2	1.294	18.49	68.1
3	0.929	13.27	81.4
4	0.724	10.33	91.7
5	0.338	4.82	96.6
6	0.236	3.37	100.0
7	1.497E-06	2.139E-05	100.0

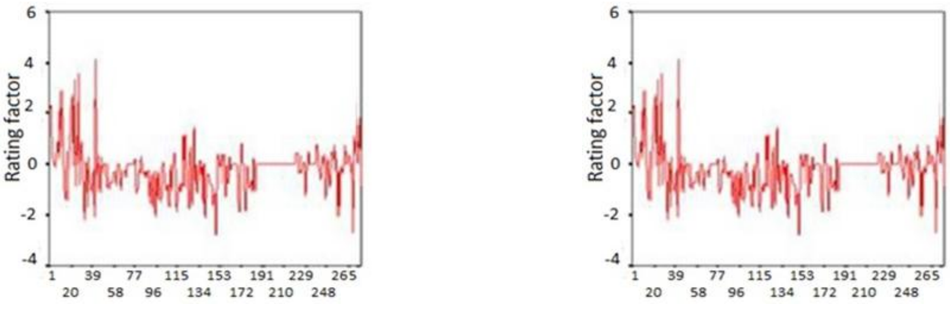
application is to facilitate interpretation by providing a simple factor structure. The factors were rotated in order that the observed axis were aligned with a dominant set of air quality parameters which assisted in understanding the relation of factors to the observed water quality parameter (Zeng and Rasmusson 2005). In this work the Quartimax rotation was used, another rotation such as biquartimax, equamax and varimax were also developed (Johnson and Wichern 2007, Kaiser 1960). Table 4 indicates the factors before and after rotation. In the factor loading after rotation the first factor has higher air quality parameters as compared to the second factor, with both being positive. The largest positive value 0.944 belongs to NO_2. The first factor included CO (0.89), NO_2 (0.453), NO (0.936), NO_X (0.944) and O_3 (-0.407), and the second factor included PM_{10} (0.877) and SO_2 (0.861).

Fig. 2 illustrates the time series for the first factor after normalization using logarithmic transformation.

In this stage, we had two time series, including the first and second factors. The results of normality of the first factor series are presented in Table 5. As described in the table, the first factor time series data was not normal since the significance level was less than 0.05 and a very strong presumption against a neutral hypothesis existed. The observed result would be highly unlikely under the null hypothesis. First, the series was transformed to normal series using logarithmic transformations. Normality was carried out for both factors time series data. Next and then stationary by the factor time series was performed.

Table 4 Factors loading for the air quality observations

Air quality parameters	Factor loading before rotation		Factor loading after rotation	
	1	2	1	2
CO	0.859		0.936	
NO_2	0.583		0.944	
NO	0.904		0.453	0.390
NO_X	0.929		0.890	
O_3		0.314		0.861
PM_{10}	0.506	0.724		-0.407
SO_2	0.615	0.649		0.861



(a) (b)

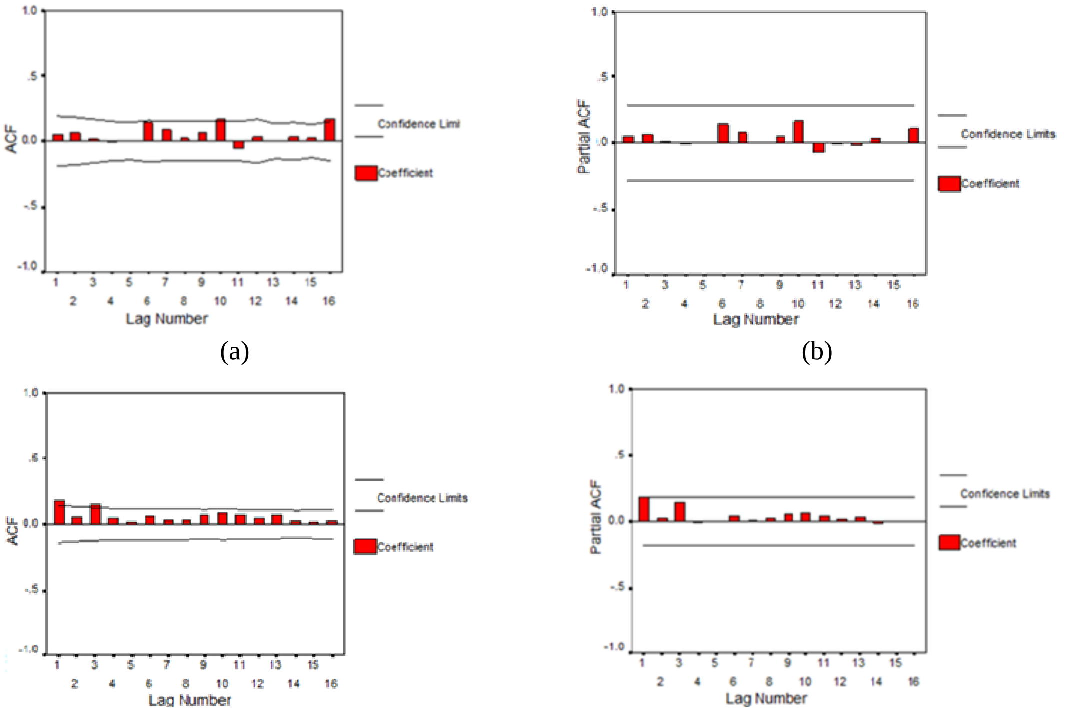
Fig. 2 (a) The time series for the first factor after normalization using a logarithmic transformation; (b) the time series for the second factor after normalization using logarithmic transformation

Table 5 The results of the normality of the first factor time series data

The first factor	Kolmogorov-Smirnov Test			Shapiro-Wilk test		
	Statistical test	Degree of freedom	Significant level	Statistical test	Degree of freedom	Significant level
	0.121	364	0.000	0.953	364	0.000

Table 6 The results of the statistical testing for the normality of the second factor data

The first factor	Kolmogorov-Smirnov Test			Shapiro-Wilk test		
	Statistical test	Degree of freedom	Significant level	Statistical test	Degree of freedom	Significant level
	0.159	364	0.000	0.785	364	0.000



(a)

(b)

(c)

(d)

Fig. 3 (a) the ACF of the transformed first factor time series data; (b) the PACF of the transformed first factor time series data; (c) the ACF of the transformed second factor time series; (d) the PACF of the transformed second factor time series data

Table 6 illustrates the results of both Kolmogorov-Smirnov and Shapiro-Wilk tests were less 0.05%, which describes that the data of the second factor was not normal. For the transformation of the second factor time series data the results of normality of the second factor series.

As indicated in Tabtle 6, the results of normality, logarithmic transformation was applied. Fig. 3 presents ACF and the PACF for the first and second factor time series. From the ACF and PACF we estimated the orders of the model. The first suggested model for the first factor time series model was ARIMA (0,0,2).

The parameter's value of Moving Average order one (MA (1)) model and its p-value were 0. 277 and 0.358, respectively. Therefore, no presumption against the neutral hypothesis are available. The MA (2) parameter value was 0.292. The constant value of the model and p-value were –0.0007185 and 0.004, respectively. Therefore, we selected ARMA (0,0,1) instead of ARMA (0,0,2) for the first factor.

For second Factor, Autoregressive order one (AR 1) parameters were 0.93 and the Moving Average order one (MA1) parameter was 0.567. Both parameters rejected the null hypothesis; the P-value was less than 0.05. However, the MA (2) parameters were estimated at 0.23and the p-value was 0.106 which was bigger than 0.05. Therefore, there is no presumption against the neutral hypothesis. The constant coefficient of the model was 0.35585991. Therefore, the ARIMA (1,0,1) was selected instead of the ARIMA (1,0,2).

As indicated in Table 7, ten data of the first and second factor were compared to the ten predicted data of the first and second factor. Table 7 indicates the lower and upper confidence interval. The coefficient of determination between the first and second factor data and predicted data were 0.98 and 0.983 which may describe the reliability of the model.

Nakhaei *et al.* (2011) developed a Box-Jenkins time series model for carbon monoxide of the Thohid tunnel in Tehran and the coefficient of determination between their observed and the predicted data was 0.7. Nevertheless, the coefficient of determination of this study of both factors were 0.98 and 0.983, respectively. All mentioned researchers used a Box-Jenkins time series model for each of the air quality parameters. Nonetheless, we developed two models for the first and second factor instead of seven air quality parameters. This caused time savings in developing models and air quality parameters.

Table 7 The comparisons between the 1st and 2nd factor data with the predicted data

No.	The factors series data		The values of transformed time series		The predicted values		Difference		Lower confidence interval		Upper confidence interval	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
1	1.39	0.46	0.38	-0.33	0.33	-0.37	0.05	0.03	0	0	0.75	0.57
2	1.29	0.61	0.34	-0.21	0.32	-0.24	0.02	0.03	0	0	0.73	0.66
3	0.96	0.55	0.33	-0.26	0.32	-0.28	0.01	0.02	0	0	0.73	0.67
4	0.94	0.52	0.33	-0.28	0.32	-0.28	0.01	0.00	0	0	0.73	0.62
5	1.19	0.25	0.28	-0.59	0.26	-0.59	0.02	0.00	0	0	0.62	0.31
6	2.34	0.73	0.61	-0.13	0.59	-0.11	0.02	-0.02	0.19	0	1.00	0.78
7	1.23	0.46	0.31	-0.33	0.31	-0.31	0.00	-0.02	0	0	0.73	0.60
8	1.42	0.33	0.39	-0.48	0.41	-0.45	0.02	-0.03	0.06	0	0.76	0.45
9	1.97	0.37	0.54	-0.43	0.57	-0.39	0.03	-0.03	0.23	0	0.91	0.52
10	0.93	0.90	0.58	-0.04	0.57	-0.01	0.01	-0.03	0.15	0	0.99	0.92

4. Conclusions

Considering the results of applying the factor analysis and the Box-Jenkins time series model to predict seven air qualities in Aghdaseyeh in the northeastern Tehran, we summarized the following conclusions:

- Seven air quality parameters were converted in two factors in which the first factor was CO, NO_2, NO, NO_X, and O_3, and the second factor was PM_{1O} and SO_2.
- We determined ARIMA (1,0,0) for the first factor and ARIMA (1,0,1) for the second factor.
- The coefficient of determination between the observed factor data and the predicted factor for both models were 0.98 and 0.983 which means the models may be reliable.

We reduced seven Univariate Box-Jenkins to two models which reduces time consumed for developing models.

References

Abbaspour, M. and Rahmani, A.M. (2005), "Carbon monoxide prediction using novel intelligent network", *Int. J. Environ. Sci. Technol.*, **1**(4), 257-264.

Asadollahfardi, G. (2002), "Analysis of surface water quality in Tehran", *Water Quality Res. J. Can.*, **37**(2), 489-511.

Asadollahfardi, G. (2014), *Water Quality Management*: *Assessment and Interpretation*, Springer Berlin Heidelberg, Germany.

Asadollahfardi, G., Khodadi, A. and Paykni, B. (2012), "Application of multivariate statistical analysis to define water quality in Jajrud River", *Asian J. Water Environ. Pollut.*, **9**(4), 1-10.

Box, G.E.P and Jenkins, G.M. (1970), *Time Series Analysis*: *Forecasting and Control*, Holden-Day, San Francisco, CA, USA, 562 p.

Box, G.E.P. and Jenkins, G.M. (1976), *Time Series Analysis, Forecasting and Control,* Holden-Day, San Francisco, CA, USA.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2008), *Time Series Analysis*: *Forecasting and Control*, (4th Edition), Wiley.

Brown, T.A. (2014), *Confirmatory Factor Analysis for Applied Research*, Gilford Press.

Browne, M.W. (1968), "A comparison of factor analytic techniques", *Pyschometrika,* **33**(3), 267-334.

Buckley, D.E. and Winter, G.A. (1992), "Geochemical characteristics of contaminated surflcial sediment in Halifax Harbour: Impact of waste discharge", *Can. J. Earth Sci.*, **29**(12), 2617-2639.

Cryer, J.D. (1986), *Time Series Analysis*, Duxbury Press, Boston, MA, USA.

Dimitriades, B. and Whisman, M. (1971), "Carbon monoxide in lower atmosphere reactions", *Environ. Sci. Technol.*, **5**(3), 213-222.

Erfanmanesh, M. and Afuni, M. (2006), *Pollution of Water, Soil, Air*, Arkan Publication, Isfahan, Iran. [In Persian]

Gocheva-Ilieva, S.G., Ivanov, A.V., Voynikova, D.S. and Boyadzhiev, D.T. (2014), "Time series analysis and forecasting of air pollution in small urban area: An SARIMA and factor analysis approach", *Stochastic Environ. Res. Risk Assess.*, **28**(4), 1045-1060.

Hakstian, A.P., Rogers, W.T. and Cattle, R.B. (1982), "The behavior of numbers of factors with simulated data", *Multivariate Behaviour Res. J.*, **17**(2), 193-219.

Harvey, A.C. and Todd, P.H. (1983), "Forecasting economic time series with structural and Box-Jenkins models (with comments)", *J. Business Economic Statistics*, **1**(4), 299-315.

Johnson, R.A. and Wichern, D.W. (2007), *Applied Multivariate Statistical Data Analysis*, (6th Edition), Pearson Prentice Hall Publisher.

Kaiser, H.F. (1960), *The Application of Electronic Computers to Factor Analysis*, Education Psychology Meas, **20**, 141-151.

Kumar, U. (2015), “An Integrated SSA-ARIMA approach to make multiple day ahead forecasts for the daily maximum ambient O_3 concentration”, *Aerosol Air Quality Res*., **15**(1), 208-219.

Kumar, U. and Jain, V.K. (2010), “ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO)”, *Stochastic Environ. Res. Risk Assess*., **24**(5), 751-760.

Linn, R.L. (1968), A Monte Carlo Approach to the Number of Factors Problem, *Sychometrika*, **33**, 37-71.

Maray, A.M. (2005), “Predection of airborne alnus pollen concentration by using arima models”, *Ann. Agricacut. Environ. Medicine*, **13**, 25-32.

Mulaik, S.A. (2009), *Foundation of Factor Analysis*, (2nd Edition), Chaoman/CRC.

Nakhae, A.M., Moravaji, G. and Mohammadipour, H. (2011), “Perdiction of of carbon monoxide in Thohid tunnel in Tehran using Box-Jenkins time series”, *Proceedings of the 11the National Conference of Transportation and Traffic*, Tehran, Iran, March. [In Persian]

Padro, R., Barrado, E., Cartilage, Y., Velasoco, M.A. and Vega, M. (1993), “Study of the contents and speciation of heavy metal in river sediments by factor analysis”, *Anal. Lett*., **26**(8), 1719-1739.

Polydoras, G.N., Anagnostopoulos, J.S. and Ch Bergeles, G. (1998), “Air quality predictions: Dispersion model vs Box-Jenkins stochastic models. An implementation and comparison for Athens, Greece”, *Appl. Therm. Eng*., **18**(11), 1037-1048.

Rodriguez-Rajo, F.J., Valencia-Barrera, R.M. and Vega-Murray, A.m. (2005), “Prediction of airborne on pollen concentration by using ARIMA”, *Ann. Agricultural Environ. Medicine*, **13**, 25-32.

Rumsey, D. (2003), *Statistics for Dummies*, John Wiley and Sons, 356 p.

Sadr Mousavi, M.S. and Rahimi, A. (2008), “The application of artificial neural networks in prediction of Co concentration: A case study of Tabriz”, *Iran. J. Natural Resour*., **61**(3), 681-691. [In Persian]

Sami, M., Waseem, A., Jafri, Y.Z., Shah, S.H., Ahmed, M., Khan, S.A., Akbar, S., Siddiqui, M. and Murtaza, G. (2012), “Prediction of the rate of dust fall in Quetta city, Pakistan uses seasonal ARIMA (SARIMA) modeling”, *Int. J. Phys. Sci*., **7**(10), 1713-1725.

Sharma, S. and Khare, M. (2000), “Real-time prediction of extreme ambient carbon monoxide concentrations due to vehicular exhaust emissions using univariate linear stochastic models”, *Transport Environ*., **5**(1), 59-69.

Sharma, P., Chandra, A. and Kaushik, S.C. (2009), “Forecasts using Box-Jenkins models for the ambient air quality data of Delhi City”, *Environ. Monitor. Assess*., **157**(1-4), 105-112.

Tabachnick, B.G. and Fidell, L.S. (1996), *Using Multivariate Statistics*, (3rd Edition), HarperCollin college Publisher.

Tobías, A., Sáez, M. and Galán, I. (2004), “Herramientasgráficaspara el análisisescriptivo de series temporales en la investigaciónmédica”, *Medicine Clínic*, **122**(18), 701-706.

Torstun, H. (1990), *Factor Analysis of Statistical*, John Wiley & Sons, New York, NY, USA.

Tucker, L.R., Koopman, R.F. and Linn, R.L. (1969), “Evaluation of factors analytic research procedures by means of simulated correlation matrix”, *Psychomertrica*, **34**, 421-459.

Vaseghi, A. and Zibaei, M. (2008), “Time series model of air pollution forecast in Shiraz”, *J. Environ. Studies* , **34**(47), 65-72. [In Persian]

Wark, K. and Warner, C.F. (1981), *Air pollution, its Origin and Control*, Harper & Row, New York, NY, USA.

Williams, P.L. and Burson, J.L, (1985), *Industrial Toxicology, Safety and Health Applications in the Workplace*, Van Nostrand Reinhold, New York, NY, USA.

Zeng, X. and Rasmussen, T.C. (2005), “Multivariate statistical characterization of water quality in Lake Lanier”, *J. Environ. Quality*, **34**(6), 1980-1991.

CC