Forecasting ambient air pollutants by box-Jenkins stochastic models in Tehran

J. Delaram and M. Khedmati*

Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran.

Received 9 February 2019; received in revised form 11 January 2020; accepted 22 February 2020

Abstract. This paper studies the behavior of six air pollutants (including PM$_{10}$, PM$_{2.5}$, O$_3$, SO$_2$, NO$_2$, and CO) in Tehran over a 6-year time span. In this paper, an iterative procedure based on the univariate Box-Jenkins stochastic models is applied to develop the most effective forecasting model for each air pollutant. Applying a number of widely used criteria, the best model for each air pollutant is selected and the results show that the proposed models perform accurately and satisfactorily for both fitting and predicting where the fitted and predicted values are close to the true values of the related data. Finally, factor analysis is conducted to investigate the relationships between the air pollutants where the results show that four factors account for 93.2704% of the total variance. In this regard, the factor containing PM$_{10}$ and PM$_{2.5}$ and the factor containing CO and NO$_2$ are, respectively, the most and the second most affecting factors with the proportion of 43.2594% and 21.6500% of the total variability. Since both of these factors stem from the large-scale use of fossil-fuel vehicles, reducing the number of vehicles or improving the quality of fossil fuels, may increase air quality by 60%.

1. Introduction

Urbanization, the ever-increasing development of cities, and the volatile speed of technologies have been intensified in recent decades. Despite all benefits, these technological developments and achievements have drastically affected the ecosystem and put them in danger. Air pollution is one of the serious environmental problems stemming from the development of these technologies, threatening public health, social welfare, and even economic success [1–3]. Air pollution is a consequence of one or several factors such as urbanization, rapid population growth, the inadequacy of public transportation systems, non-standard motor-vehicles and etc. The threat of air pollution and its catastrophic effects are serious especially for megacities and developing countries [4,5]. Accordingly, along with the utilization of effective solutions to hinder the negative effects of air pollution, monitoring of air pollution is a necessary matter in order to support municipal decision-making and management.

Air pollution can be interpreted as the presence of various pollutants in the ambient air. These pollutants are often detrimental to the health of humans, animals, plants, and living creatures [6–8]. There are many standards and regulations about air pollutants which define and determine the parameters and their acceptable level of health. The National Ambient Air Quality Standards (NAAQS) is one of the most prominent environmental standards which is developed

---

* Corresponding author.
E-mail addresses: jalal.delaram@ece.sharif.edu (J. Delaram);
Khedmati@sharif.edu (M. Khedmati)

doi: 10.24200/sci.2020.52890.2537
by the United States Environmental Protection Agency (EPA). NAAQS identifies the following parameters as
the criteria for air pollution: Particulate matter 10
\((PM_{10})\), particulate matter 2.5 \((PM_{2.5})\), ozone \((O_3)\),
sulfur dioxide \((SO_2)\), nitrogen dioxide \((NO_2)\), nitrogen
monoxide \((NO)\), and carbon monoxide \((CO)\) \cite{9}.

The problem of monitoring air pollutants has received
great attention among researchers and prac-
titioners for a long time and hence, various approaches
and tools have been developed to deal with this
problem. On one hand, the researches can be classified
according to the considered air pollutants. Considering
more detrimental effects of particulate matter \((PM_{10}
and PM_{2.5})\) and \(O_3\) parameters on the environment and
human health, they have received more attention in the
researches \cite{10-13}, however, other parameters are still
important and have been considered in some researches
including Aufferhammer and Carson \cite{14}, Olabemiwo
et al. \cite{15}, and Cabaneros et al. \cite{16}. On the other
hand, the researches can be classified according to the
exploited modeling approach. Considering the high
capability of Artificial Neural Networks (ANN) in
modeling non-linear relationships, some authors
developed an ANN-based forecasting model for air
pollutants \cite{17,18} while, due to the simplicity and
reliability of Multi Linear Regression (MLR), some
other authors proposed an MLR model \cite{19-21}. There
are also other researches in the field of forecasting air
pollutants that use hybrid models such as MLR and
ANN \cite{22}, support vector machine and ANN \cite{23},
and Artificial Fuzzy Neural Networks (AFNN) \cite{18}. In
addition, some researchers have used a Response Surface
Modeling (RSM)-based approach for the prediction of
air pollutants \cite{24}.

Time series modeling approaches have been
widely used for modeling and forecasting of time series
data in different applications (see for example \cite{25-36}).
Among them, there are many types of researches that
consider the air pollutants’ behavior as a stochastic
process and apply the univariate time series models for
analysis and forecasting of these processes. In fact,
the air quality is highly dependent on the weather
conditions, and air pollution is strongly governed by
meteorology \cite{37}. However, in the univariate time
series models, the concentration of air pollutants is
considered to be the final result of intricate interactions
between different actors including meteorology, chem-
istry, transportation, and etc. As a result, the process
of air pollution is modeled by univariate time series
models without the inclusion of other variables like me-
teorological ones which lead to simplifying the process
modeling and the related calculations. Considering
the high efficiency of univariate Box-Jenkins models,
exploitation of univariate models also helps to achieve
suitable results. Although there are some critics and
arguments about Box-Jenkins models \cite{38}, there are a
lot of research works that applied these models and
achieved invaluable results. Box-Jenkins models have
been popular in forecasting air pollution, and the high
capabilities of these models have been proved in dif-
ferent researches. Kumar and Jain \cite{39} proposed Box-
Jenkins models for forecasting ambient air pollutants
including \(O_3\), \(NO_2\), NO, and CO. Zhou and Goh \cite{40}
used the same approach for modeling \(PM_{2.5}\), and Jian
et al. \cite{41} applied the Box-Jenkins model for predicting
\(PM_{10}\) and submicron concentrations. The combination
of Box-Jenkins models with other statistical models has
been considered, also, for forecasting and analysis of
air pollution where ANN is more popular than others.
Díaz-Robles et al. \cite{42} applied a hybrid Box-Jenkins
and ANN model for forecasting \(PM_{10}\) and \(O_3\), and
Samia et al. \cite{43} used the same approach for \(PM_{10}\). A
summary of the related researches in the past decade
based on the type of model(s) for forecasting, the
timespan to develop the forecasting model, and the air
pollutants considered to be forecasted, is presented in
Table 1.

Based on the reports of the World Bank and
World Health Organization (WHO), Tehran has one
of the most polluted ambient air in the world, ranked
12th among 26 megacities in the world in 2016 \cite{51}.
Tehran as the capital and the biggest city of Iran
is a megacity in a developing country that is highly
endangered with harmful damages of air pollution.
The Air Quality Control Company (AQCC) of the
Municipality of Tehran is responsible for monitoring
the air quality in Tehran and its urban area. The
AQCC monitors the air pollution parameters based
on NAAQS, as mentioned above. The AQCC has
23 stations throughout the city, and systematically
collects data from these stations and processes them
to monitor air quality. In this study, we use the daily
reports and information of the AQCC for a timespan
of 6 years from 20 March 2012 (1 Farvardin 1391 in
Persian calendar) to 20 March 2018 (29 Esfand 1396
in Persian calendar).

This paper presents a statistical study of the air
pollutants parameters in Tehran. In this regard, a
stationary stochastic Box-Jenkins modeling approach
has been adapted to forecast the daily average ambient
air pollutants \((PM_{10}, PM_{2.5}, O_3, SO_2, NO_2,\) and \(CO)\)
concentrations in Tehran. The data for daily mean
air pollutants concentrations have been obtained from
AQCC under the supervision of the Municipality of
Tehran (http://air.tehran.ir/; accessed in June-2018).
This research is conducted, mainly, in order to provide
a good forecast for each of the air pollution parameters
and to present an effective short-term forecasting model
of air pollutants for Tehran, based on the statistical
and time series modeling techniques. The novelty of
this paper resides in providing a study over each of the six
air pollutants of Tehran’s ambient air and proposing
Table 1. Summary of the related researches in other countries and cities (sorted based on year).

| Author(s)            | Year  | Location              | Timespan | Model(s)                  | Index | PM$_{10}$ | PM$_{2.5}$ | O$_3$ | SO$_2$ | NO$_2$ | NO    | CO   |
|----------------------|-------|-----------------------|----------|---------------------------|-------|-----------|-----------|-------|--------|--------|-------|------|
| Díaz-Robles et al. [42] | 2008  | Temuco, Chile         | 7 years  | ANN and ARIMA             | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Kurt et al. [17]      | 2008  | Istanbul, Turkey      | 3 years  | ANN                       | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Hoj et al. [44]       | 2009  | Macau, Macau         | 5 years  | ANN                       | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Kumar and Jain [30]   | 2010  | Delhi, India          | 1 year   | ARIMA                     |     |           |           | ✓     | ✓      | ✓      | ✓     | ✓    |
| Genc et al. [45]      | 2010  | Ankara, Turkey        | 2 years  | MLR                       | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Vlahogianni et al. [22]| 2011  | Athens, Greece        | 1 year   | MLR and ANN               | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Poggi and Portier [46] | 2011  | Rouen, France         | 5 years  | CLR                       | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Samia et al. [43]     | 2012  | Sfax, Tunisia         | 5 years  | ANN and ARIMA             | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Jian et al. [41]      | 2012  | Hangzhou, China       | 1 year   | ARIMA                     | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Muñoz et al. [23]     | 2014  | Algeciras, Spain      | 7 years  | SVM and ANN               | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Gocheva-Blieva et al. [47] | 2014 | Blagoevgrad, Bulgaria | 1 year   | SARIMA                    | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Elbazgui et al. [19]  | 2014  | Gaza, Palestine       | 1 year   | MLR                       | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Asadollahhaﬁd et al. [72] | 2015 | Aghdavayeh, Iran      | 1 year   | ARIMA                     | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Cortina-Jamers et al. [46] | 2015 | Salamanca, Mexico     | 2 years  | ANN                       | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Jiang et al. [40]     | 2017  | Jingjinji, China      | 1 year   | AFNN                      |      | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Zhou and Goh [40]     | 2017  | Singapore, Singapore  | 1 year   | ARIMA                     | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| Abdollahimazdeh et al. [50] | 2018 | Tehran, Iran          | 2 years  | AFNN                      | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |
| This study            | 2021  | Tehran, Iran          | 6 years  | ARIMA                     | ✓     | ✓         | ✓         | ✓     | ✓      | ✓      | ✓     | ✓    |

The best forecasting model for each air pollutant. In addition, applying factor analysis, the relationships between these air pollutants are analyzed to determine the effect of air pollutants on air quality and group them based on their effect on the air quality. Accordingly, the main sources of the Tehran air pollution problem are studied empirically, and investigating the air pollutants that have the greatest impact on air quality, decision-makers can focus on the sources of air pollutants and take appropriate actions to improve air quality by reducing the emission of these air pollutants.

The rest of the paper is organized as follows. The study area and the related air pollutants data are described in Section 2. The Box-Jenkins models are generally introduced in Section 3. Then, in Section 4, the Autoregressive Integrated Moving Average (ARIMA) models are proposed for each of the air pollutants including PM$_{10}$, PM$_{2.5}$, O$_3$, SO$_2$, NO$_2$, and CO, based on the iterative procedure of time series modeling. The factor analysis is applied in Section 5 to illustrate the relationships among the variables and show the variables that have the greatest impact on air quality. Finally, concluding remarks are provided in Section 6.

2. Situation

2.1. Study area

The area under study in this paper is the city of Tehran, the capital of the Islamic Republic of Iran (IRI). Tehran is the biggest city of IRI with a mean population of 8.5 million which can reach over 12.5 million during the day, because of commuting people from nearby cities [52]. Tehran is located in the north of Iran and the south of the high altitudes of the Alborz Mountain Range. The geographic coordinates of the city are 51°.2' and 51°.36' East longitude and 35°.34' and 35°.50' North latitude as illustrated in Figure 1, and the altitudes of the city vary between 2000 to 1000 meters from the north to south above sea level.
level, respectively. Tehran has a semi-arid climate and the main source of precipitation is the Mediterranean and Atlantic winds which blow from the West. Also, the Alborz Mountain Range hinders the penetration of air masses from the Caspian Sea. The variations of Tehran’s temperature are in a range from 40°C (in summers) to −5°C (in winters) and the mean annual rainfall is about 250 millimeters [53].

Based on the reports provided by Hosseini and Shahbazi [54], Tehran has more than 17 million cars traveling every day, most of which are obsolete, and therefore become one of the main sources of air pollution in Tehran. On the other hand, Tehran is surrounded by the Alborz Mountain Range altitudes which trap polluted air, especially when the weather becomes cold and inhibits the pollutants to be diluted; a phenomenon called Temperature Inversion.

2.2. Data
The models, in this paper, are developed for data related to six air pollutants in a 6-year time span in Tehran from 20 March 2012 (1 Farvardin 1391 in Persian calendar) to 20 March 2018 (29 Esfand 1396 in Persian calendar), equal to 2192 days. The observed pollutants are concentrations of PM10, PM2.5, O3, SO2, NO2, and CO which are expressed in the unit of mass concentration of pollutants in microgram per cubic meter (µg/m³). The data related to the air pollutants have been collected and processed by 23 stations of AQCC across the city. It should be noted that there are no missing values in the data set.

The general properties of the data are illustrated and represented in Table 2 by descriptive statistics of the data including maximum, minimum, range, mean, standard deviation, variance, skewness, and kurtosis coefficients. The skewness and kurtosis are often employed to examine the properties of symmetry and flatness of the density function and the distribution of the data in the time series. According to Table 2, the maximum value of skewness and kurtosis are, respectively, 1.6537 and 13.6134 which is related to PM10. Also, the threshold limit for each of the six air pollutants is presented according to the standards of WHO and the European Commission for air quality and standard.

As the data are gathered in a daily routine, a seasonality behavior in the time series of the indices can be considered by days.

3. Box-Jenkins models
In 1970, Box and Jenkins introduced a general class of models in order to find the best fit for a time-series model of past observations, entitled ARIMA models [55]. ARIMA models are intrinsically a mixture of three processes of (a) Autoregressive (AR), (b) differencing, and (c) Moving Average (MA). Hence, the notation in order to distinguish an ARIMA model is suggested as ARIMA (p, d, q) where, p is a non-negative integer that describes the parameters of the AR process; d is a nonnegative integer that describes trend process (I), and q is also a nonnegative integer to introduce the parameters of MA process. Estimation of these parameters is usually determined by means of iterative procedures which seek to minimize the sum of squares for a non-linear regression model.

The general form of an ARMA model of order (p, q), a mixture of AR and MA models, can be represented as:

$$\Phi(B)x_t = \delta + \Theta(B)\varepsilon_t,$$

where $t = 1, 2, 3, \ldots, n$ denotes the time values and $n$ is the total number of observations in the time series, $x_t$ denotes the value of the time series variable $x$ at time $t$, and $B$ is the backshift operator. In addition, $\varepsilon_t$ represents the error term at time $t$ where, based on the Wold theorem [56], the error term should be white noise, that is, uncorrelated random shocks with mean zero and constant variance of $\sigma^2$ or equivalently $\varepsilon_t \sim WN(0, \sigma^2)$. Moreover, $\Phi(B)$ and $\Theta(B)$ are AR and MA operators of order, respectively, $p$ and $q$ are represented as:

$$\Phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i \quad \text{and} \quad \Theta(B) = 1 - \sum_{i=1}^{q} \theta_i B^i.$$
changeable to a stationary process by differencing, is called a homogenous nonstationary process. Therefore, a homogenous nonstationary ARMA \((p, q)\) process which is transformed into a stationary process using a differencing of order \(d\), is called an ARIMA process of orders \(p, d,\) and \(q\) or equivalently ARIMA \((p, d, q)\) and is represented as:

\[
\Phi(B)(1 - B)^d x_t = \delta + \Theta(B)z_t.
\]  

\(2\)

In most cases, first-order \((d = 1)\) or second-order \((d = 2)\) differencing is enough to achieve the stationarity condition.

If a periodic pattern or a seasonal behavior exists in the time series, a seasonal ARIMA or Seasonal Autoregressive Integrated Moving Average (SARIMA) model can be exploited to investigate the behavior of the process in which, the general form of SARIMA models is given as:

\[
\Phi'(B')\Phi(B)(1 - B)^d(1 - B')^D x_t = \delta + \Theta'(B')\Theta(B)z_t.
\]  

\(3\)

The notation of this model is ARIMA \((p, d, q)\times(P, D, Q)\) where, \(P\) is the parameter for the number of seasonal AR terms, \(D\) is the parameter for the order of seasonal differencing, \(Q\) is the parameter for the number of seasonal MA terms and \(s\) is the parameter for the number of periods in a season.

4. Building the forecasting model

There are several approaches to develop a Box-Jenkins forecasting model but, in essence, almost all of them are the same and only differ in lateral details. One of the common characteristics of these approaches is being iterative. Hence, it is needed to iterate the procedure successively to achieve a suitable forecasting model. The (iterative) procedure which is exploited in this paper to develop the forecasting model is as follows.

Step 1: Initial analysis. Plotting and analyzing the time series plot of the data can help to determine the general pattern and behavior of the corresponding process. By investigating the patterns and behavior of the process, one can take the appropriate action to stabilize data, reduce the variability and detrend data to achieve the stationarity conditions.

The time series plot of the six air pollutants presented in Appendix A reveals that the data have some patterns that make them nonstationary. For example, the time series of O\(_3\) in Figure A. 4 illustrates a strong seasonal behavior with a slight descending trend which is a sign of nonstationarity. Furthermore, Table 3 presents the results of applying the Augmented Dickey-Fuller (ADF) test on the time air pollutants time series.

| Variable | ADF statistic | Level of significance (%) | C-values (MacKinnon critical values) |
|----------|---------------|---------------------------|-------------------------------------|
| PM\(_{10}\) | -0.9801      | 10                        | 2.5060                              |
| PM\(_{1.0}\) | -4.8116      | 10                        | -1.6108                             |
| O\(_3\) | 4.9888        | 10                        | 2.5060                              |
| SO\(_2\) | -3.3806       | 10                        | -1.6108                             |
| NO\(_2\) | -3.2086       | 10                        | 2.5060                              |
| CO | -5.0748       | 10                        | -1.6108                             |

Dickey-Fuller (ADF) test on the time series of air pollutants in which, the results show that the values of the ADF statistic for all pollutants violate the critical values \((C\text{-value})\) at 1%, 5%, and 10% level of significance. Obviously, the null hypothesis of the unit root test should be rejected for all cases.

In addition, according to the results of the Kolmogorov-Smirnov (KS) test, presented previously in Table 2, a manipulation is required in order to reduce the variation and to achieve normality. Two appropriate techniques in this step are data transformation and differencing in order to stabilize the variation of the data and detrend the process, respectively. This paper exploits Yeo-Johnson power transformation [57] which is an improvement of the Box-Cox power transformation family [58]. The Yeo-Johnson power transformation is suitable for data with any sign:

\[
\Psi_{YJ}(\lambda, x) = \begin{cases} 
(x^{\lambda+1})^{\frac{1}{\lambda+1}} & x \geq 0, \lambda \neq 0 \\
\log(x + 1) & x \geq 0, \lambda = 0 \\
\frac{-(x^{\lambda+1})^{\frac{1}{\lambda+1}}}{\lambda} & x < 0, \lambda \neq 2 \\
\log(-x + 1) & x < 0, \lambda = 2 \\
\end{cases}
\]

\(0 \leq \lambda \leq 2.\)  

By considering all the possible values in the range of \([-2, -1.9, -1.8, \ldots, 1.8, 1.9, 2]\), the Yeo-Johnson power transformation coefficients for all the variables are obtained according to the KS test statistic in which the \(\lambda\) which provides transformed data with minimum KS statistic is selected as the best value. The values
for coefficient $\lambda$ and other descriptive statistics for transformed data are summarized in Table 4.

**Step 2: Development of tentative models.** In this step, a number of models which are expected to explain the behavior of the time series are proposed and examined. One of the effective tools to specify the potentially appropriate models for time series data is the sample Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). ACF is a function that provides a measure to show the correlation between $x_t$ and its value in another time period like $x_{t+k}$, and it is obtained as:

$$\rho_k = \frac{E[(x_t - \mu)(x_{t+k} - \mu)]}{\sqrt{E[(x_t - \mu)^2]E[(x_{t+k} - \mu)^2]}}$$

$$= \frac{Cov(x_t, x_{t+k})}{Var(x_t)} = \frac{\gamma_k}{\gamma_0}, \quad k = 0, 1, 2, ...$$

(5)

in which $\gamma_k$ is the autocovariance at lag $k$. To provide an estimation for $\rho_k$, according to time series $x_1, x_2, x_3, \ldots, x_T$, the following formula is presented to calculate the sample ACF:

$$r_k = \hat{\rho}_k = \frac{c_k}{c_0}, \quad k = 0, 1, 2, ..., K,$$

(6)

where $c_k = \hat{\gamma}_k = \frac{1}{T-k} \sum_{i=1}^{T-k} (x_t - \bar{x})(x_{t+k} - \bar{x})$ is an estimation for autocovariance at lag $k$.

However, the correlation between a variable and its lagged value does not always interpretable with only the autocorrelation coefficient. Therefore, the partial autocorrelation coefficient is presented to address this problem. The PACF gives the partial correlation coefficients of a time series with its own values at any lag. To obtain the PACF, consider the Yule-Walker equations set for the ACF of an AR ($p$) process:

$$\rho(j) = \sum_{i=1}^{k} \phi_{kj} \rho(j - i), \quad j = 1, 2, ..., k.$$  

(7)

Denoting $\phi_{kj}$ as the $j$th coefficient of an AR ($p$) process, the $\phi_{kk}$ for any given $k$ is called the partial autocorrelation coefficient at lag $k$ for the time series $x_t$.

According to the definition of PACF, in an AR ($p$) process if $k > p$, then $\phi_{kk} = 0$. Hence, the PACF should cut off after lag $p$ in an AR ($p$) process. This feature helps us to identify the order of the AR process, just like the ACF does in a moving average process. ACF and PACF are two useful tools in the model identification phase. As mentioned previously, AR and MA processes have some characteristics that affect the form of ACF and PACF plots. Therefore, despite the various forms of these diagrams according to different characteristics of AR, MA, and ARMA processes, their ACF and PACF can be categorized based on Table 5 [59].

Although the visual inspection of the time series, ACF, and PACF provide helpful insights about the model, they are very subjective and depend highly on the experience of the forecast experts. In order to have a more objective approach and comparing the proposed models in a quantitative manner to be able to select the best models among the tentative ones, the utilization of some criteria such as Akaike Information Criterion (AIC), Schwarz’s Bayesian Information Criterion (SBIC), and Hannan-Quinn Information Criterion (HQIC) are suggested [60,61]. These criteria measure the statistical model fitting performance and present the relative goodness of fit for potential models. AIC

| Variable | $\lambda$ | Maximum | Minimum | Range | Mean | Standard deviation | Variance | Skewness | Kurtosis | KS test |
|----------|-----------|---------|---------|-------|------|-------------------|----------|----------|----------|---------|
| trPM10   | 0.5       | 16.4552 | 4.2866  | 12.1685 | 8.9548 | 1.1942            | 1.4362   | 0.0757   | 5.5230   | 0.04    |
| trPM2.5  | 0.3       | 17.6970 | 6.9166  | 10.7804 | 12.4692 | 1.6438            | 2.7022   | 0.0064   | 3.1540   | 0.02    |
| trO3     | 0.4       | 7.9477  | 2.3521  | 5.5956  | 4.9808 | 0.8351            | 0.6974   | 0.0000   | 3.0307   | 0.02    |
| trSO2    | 0.7       | 26.4264 | 3.0933  | 23.3253 | 14.0137 | 3.6309            | 13.1839  | -0.0138  | 2.8896   | 0.03    |
| tr NO2   | 0.2       | 7.5254  | 4.7200  | 2.7903  | 6.2170 | 0.4588            | 0.2105   | -0.0046  | 2.6503   | 0.03    |
| tr CO    | -0.1      | 3.6230  | 2.5834  | 1.0395  | 3.0884 | 0.1703            | 0.0290   | 0.0030   | 2.5807   | 0.03    |

| Diagram   | AR ($p$)    | MA ($q$)  | ARMA ($p,q$) |
|-----------|-------------|-----------|---------------|
| ACF       | Tail off    | Cut off after lag $q$ | Tail off |
| PACF      | Cut off after lag $p$ | Tail off | Tail off |
is the first information criterion proposed by Akaike in 1974, which its logic is based on the Kullback-Leibler (KL) distance [62]. AIC is defined as an expected KL distance that calculates the Maximum Likelihood Estimation (MLE) with some corrections related to the number of the parameters in the model. By introducing AIC, other information criteria were developed with different mathematical and statistical properties. SBIC, like AIC, penalizes the complexity of the model and adding more parameters to the model in order to prevent overfitting, but the penalty term in SBIC is larger than AIC [63]. The corrected version of AIC or AICc is an efficient information criterion when the sample size is small and it prevents overfitting by introducing more penalty for parameters, compared to AIC [64]. HQIC is an alternate information criterion for AIC and SBIC which is developed based on the law of the iterated logarithm that states any strongly consistent method will miss its efficiency by at least one ln(ln(n)) and accordingly, HQIC has a very well behavior asymptotically [65]. Also, in contrast to AIC and AICc, SBIC and HQIC are two criteria that are not affected by increasing the sample size. As a result, the paper will use HQIC as the information criterion for model selection, although other information criteria will be calculated to be used if they are needed. The formulas for these criteria are given as:

\[ AIC = -2 \ln(\hat{L}_{max}) \frac{1}{n} + 2 \frac{k}{n} \]  

\[ AICc = -2 \ln(\hat{L}_{max}) \frac{1}{n} + 2k + 2 \frac{k(k+1)}{n-k+1} \]  

\[ BIC = -2 \ln(\hat{L}_{max}) \frac{1}{n} + k \ln(n) \frac{1}{n} \]  

\[ HQIC = -2 \ln(\hat{L}_{max}) \frac{1}{n} + 2k \ln(n) \frac{1}{n} \]  

where \( \hat{L}_{max} \) is the maximum likelihood of the model, \( k \) is the number of parameters, and \( n \) is the number of observations in the model.

**Step 3. Estimation and diagnosis of the models:** This step consists of estimating the parameters of the tentative models identified in the previous step and performing the diagnostic checking. After suggesting a number of eligible models, it is needed to estimate the parameters of these models. Estimation of the parameters of the models (\( \phi \) and \( \theta \)) can be obtained utilizing different methods such as MLE, Minimum Least Squares (MLS), or Conditional Least Squares (CLS) [66,67]. As the SARIMA/ARIMA models are almost nonlinear, it is needed to use the procedure of nonlinear model fitting. This procedure is usually performed by statistical software packages such as Minitab, JMP, and SAS. In this paper, JMP software has been exploited to develop the models. The best-fitted models among the tentative ones along with a number of important measures for selecting the best model are summarized in Table 6. The selected models are shown in bold font in Table 6 in which the best forecasting model is selected according to the combined criteria, with (1) maximum adjusted R² (R² Adj.), (2) minimum HQIC, (3) minimum Root Mean Square Error (RMSE), (4) minimum Mean Absolute Percentage Error (MAPE), and (5) minimum Mean Absolute Error (MAE). Also, in all cases, it has been considered that the conditions for stationarity and invertibility of AR and MA parameters (\( \phi \) and \( \theta \)) and white noise conditions of the residuals are satisfied.

**Step 4: Exploitation of the model.** After specifying the best model, the last step is to use the model to predict future data. To do this, the model selected in the previous step is employed to forecast the air pollutants for the first three months of the Persian calendar, from 21 March 2018 to 21 June 2018. In Figures (2)–(7), the fitted model for each of the six air pollutants over the six years timespan has been illustrated which shows a very good correspondence with the pattern of the real data. Also, in each figure, a red reference line distinguishes the prediction of the out-of-sample data from the rest of the data which are used for fitting.

![Figure 2. Fitting and predicting using ARIMA (3,1,2) model for CO.](image)
| Trans. variable | Box-jenkins models | Model fitting statistics |
|----------------|-------------------|-------------------------|
| trPM$_{10}$    |                   | R$^2$    R$^2$ Adj. RMSAE MAE MAPE AIC AICc SBIC HQIC |
| (1, 1, 3)      | 0.4221 0.4211 1.5218 1.1053 0.8844 8.065.322 8.011.288 8.065.267 |
| (1, 1, 5)      | 0.4209 0.4183 1.5245 1.2067 0.9155 8.014.173 8.016.200 8.054.011 8.000.188 |
| (3, 1, 2)      | 0.4225 0.4212 1.7210 1.1052 0.8845 8.005.900 8.007.900 8.040.601 8.053.922 |
| (3, 1, 5)      | 0.4231 0.4216 1.7218 1.1043 0.8820 8.009.574 8.011.674 8.060.803 8.051.592 |
| trPM$_{2.5}$   |                   | R$^2$    R$^2$ Adj. RMSAE MAE MAPE AIC AICc SBIC HQIC |
| (2, 1, 5)      | 0.4075 0.4058 0.7546 0.5627 0.5023 4.903.285 4.905.367 5.038.921 4.977.391 |
| (2, 1, 7)      | 0.4063 0.4060 0.7545 0.5625 0.5013 4.905.728 4.907.871 5.058.341 4.979.750 |
| (3, 1, 2)      | 0.4074 0.4062 0.7543 0.5628 0.5026 4.886.741 4.901.702 5.021.804 4.977.754 |
| trO$_3$        |                   | R$^2$    R$^2$ Adj. RMSAE MAE MAPE AIC AICc SBIC HQIC |
| (2, 1, 3)      | 0.7058 0.7054 0.7504 0.5033 4.6846 5.132.858 5.134.909 5.167.010 5.120.871 |
| (2, 1, 2)      | 0.7082 0.7078 0.7747 0.5040 4.6617 5.105.980 5.108.019 5.134.441 5.095.992 |
| (2, 1, 2) (2, 1, 1) | 0.7471 0.7461 0.8446 0.6959 5.5416 4.831.256 5.747.652 4.805.365 4.835.072 |
| (3, 1, 2) (2, 1, 0) | 0.7400 0.7390 0.8777 0.7042 5.6124 4.860.272 5.776.647 4.934.361 4.874.068 |
| trSO$_2$       |                   | R$^2$    R$^2$ Adj. RMSAE MAE MAPE AIC AICc SBIC HQIC |
| (1, 1, 1)      | 0.8113 0.8111 1.1820 0.8808 0.6016 6.056.761 6.058.770 6.073.837 6.050.768 |
| (2, 1, 3)      | 0.8122 0.8119 1.1801 0.8700 0.6735 6.048.555 6.050.583 6.071.924 6.049.565 |
| (3, 1, 2)      | 0.8124 0.8120 1.1790 0.8785 0.6360 6.049.983 6.052.035 6.084.126 6.097.906 |
| (2, 1, 3)      | 0.8124 0.8120 1.1800 0.8788 0.6380 6.050.182 6.052.233 6.084.335 6.098.105 |
| trNO$_2$       |                   | R$^2$    R$^2$ Adj. RMSAE MAE MAPE AIC AICc SBIC HQIC |
| (1, 1, 1)      | 0.5898 0.5894 0.3090 0.2384 1.4601 9.350.467 9.361.485 9.676.543 9.261.475 |
| (2, 1, 1)      | 0.5014 0.5008 0.3003 0.2381 1.4585 9.322.049 9.545.976 0.751.717 9.445.958 |
| (2, 1, 2)      | 0.5016 0.5008 0.3003 0.2380 1.4577 9.353.846 9.558.855 0.882.307 9.438.858 |
| (2, 1, 5)      | 0.5924 0.5911 0.3092 0.2374 1.4543 9.555.500 9.571.582 1.001.036 9.395.516 |
| trCO           |                   | R$^2$    R$^2$ Adj. RMSAE MAE MAPE AIC AICc SBIC HQIC |
| (1, 1, 3)      | 0.3543 0.3528 0.1031 0.0680 0.1287 2.759.805 2.750.767 2.760.345 2.708.704 |
| (2, 1, 4)      | 0.3546 0.3528 0.1031 0.0684 0.1287 2.755.733 2.753.067 2.715.888 2.709.718 |
| (2, 1, 5)      | 0.3574 0.3526 0.1030 0.0683 0.1287 2.754.141 2.752.058 2.760.604 2.770.124 |
| (3, 1, 2)      | 0.3543 0.3528 0.1031 0.0680 0.1287 2.756.803 2.754.752 2.722.651 2.708.700 |

Table 6. The best fitted models among the tentative ones with their selection measures.

---

**Figure 3.** Fitting and predicting using ARIMA (2,1,5) model for NO$_2$.

**Figure 4.** Fitting and predicting using ARIMA (3,1,2) model for SO$_2$. 
Based on the results, the models provide satisfactory performance where the predicted values are close to the data and follow correctly the trend of the related process. In addition, the three performance criteria including RMSE, MAE, and MAPE are calculated and reported in Table 7 for the model selected for each of the air pollutants which shows the high accuracy of the models and confirms the good performance of the models for predicting the future data.

### 5. Discussion

Grouping parameters and variables has always been one of the interesting ways to study air quality. Factor
analysis is one of the most widely used approaches for processing time-series data in atmospheric and environmental sciences. It provides a tool to classify variables based on a number of unknown resources (called factors). The aim of factor analysis is to determine the presence or absence of interactions between the variables or parameters [68]. In this approach, the presence of interactions is interpreted as existing a latent common source among the variables (the air pollutants in this study).

An advantage of the factor analysis approach is to permit the researchers to categorize the variables into distinct classes and recognizing the variables which are related to each other. The proposed classes by factor analysis include mutually dependent variables which are strongly correlated. After performing factor analysis, if a number of variables are recognized as (strongly) correlated, they will be considered to be affected by a latent variable. Although factor analysis can be used to classify variables and reduce the complexity of calculation, it will cause some information to be lost. Therefore, a number of factors should be chosen to have the least information loss. The procedure for applying factor analysis to the air pollutants considered in this study is carried out in the following steps:

Step 1. Calculating the correlation matrix;
Step 2. Testing the adequacy of the approach;
Step 3. Extracting the factors;
Step 4. Rotating the factors;
Step 5. Scoring the calculation of the variables factors.

The correlation matrix of the air pollutants is presented in Table 8. The existence of a large coefficient (greater than or equal to 0.5) in the correlation matrix indicates the singularity of the correlation matrix, which is interpreted as its determinant is near to zero. By calculating the determinant of the corresponding correlation matrix, the determinant is obtained as 0.0526 which is a small value but not equal to zero. As shown in Table 8, there are significant correlation coefficients between some variables.

![Figure 8. The scatter plot matrix of the six air pollutants.](image)

In addition, for visual inspection of the correlation structure between variables, the corresponding scatter plot matrix of the correlation structure is illustrated in Figure 8. The scatter plot matrix presents all pairwise combinations of variables to demonstrate the relationship between them. According to the scatter plot and correlation matrix, PM_{2.5} and PM_{10} are highly correlated and have a positive correlation coefficient of 0.8425. Also, CO and NO_{x} are positively correlated with a correlation coefficient of 0.6333, and the rest of the correlations between the variables are negligible.

The KMO test and Bartlett sphericity test can be used to measure the adequacy of the factor analysis approach. In the KMO test, the KMO statistic should be more than 0.5 and the Bartlett statistic should have a significance value less than 0.05 [69]. The KMO statistic is 0.601 and the Bartlett statistic is 0.0000. Accordingly, there is a relationship between air pollutants, and applying the factor analysis can be useful. Therefore, using Principal Component Analysis

|          | CO     | NO_{2} | SO_{2} | O_{3}  | PM_{2.5} | PM_{10} |
|----------|--------|--------|--------|--------|----------|---------|
| CO       | 1      | 0.6333 | 0.1913 | -0.0797| 0.4625   | 0.3317  |
| NO_{2}   | 0.6333 | 1      | -0.0935| 0.0651 | 0.4186   | 0.3141  |
| SO_{2}   | 0.1913 | -0.0935| 1      | -0.2035| 0.3627   | 0.1779  |
| O_{3}    | -0.0797| 0.0651 | -0.2035| 1      | -0.1739 | 0.0614  |
| PM_{2.5} | 0.4625 | 0.4186 | 0.3627 | -0.1739| 1        | 0.8425  |
| PM_{10}  | 0.3317 | 0.3141 | 0.1779 | 0.0614 | 0.8425   | 1       |
(PCA) method, four factors have been considered and the Promax method is used to rotate the factors. The results are presented in Table 9 in which, the values below 0.5 have been ignored.

Based on the results of factor analysis presented in Table 9, the air pollutants can be divided into four groups including $F1 = \{\text{PM}_{2.5}, \text{PM}_{10}\}$, $F2 = \{\text{CO}, \text{NO}_2\}$, $F3 = \{\text{SO}_2\}$, and $F4 = \{\text{O}_3\}$.

These four factors are considered to account for 93.2704% of the total variance, of which the minimum recommended value is 80% [70]. The results of factor analysis and grouping of the air pollutants are highly related to the resources of emanating air pollution; in this way, the unknown resources are considered as factors. According to these results for a 6-year timespan in Tehran, PM$_{2.5}$ and PM$_{10}$ have similar behavior as they decrease and increase simultaneously over time. Investigating the correlation structure of the air pollutants has resulted in different conclusions in the literature. For example, Kumar and Joseph [71] addressed the high correlation of PM$_{10}$, PM$_{2.5}$, and NO$_2$, while Asadolahiabadi et al. [72] addressed the correlation of PM$_{10}$ and SO$_2$. These different conclusions can be the consequence of high variability of weather conditions over time and various resources of pollution which vary from a location to the other. Especially for particle pollution (PM$_{2.5}$ and PM$_{10}$), location has a distinctive role in determining the sources of pollution. For Tehran, the main sources of these air pollutants are incomplete combustion, automobile emissions, and dust. Another useful result that can be obtained from factor analysis is to determine the air pollutant that accounts for the largest proportion of ambient air pollution. In this way, based on the results in Table 9, the partial contribution of $F1$ in the total variability of data is 43.2594% which is more than the partial contribution of other factors. This means that PM$_{2.5}$ and PM$_{10}$ have been the major air pollutants in Tehran over six recent years.

The presence of CO and NO$_2$ in one group ($F2$) can be a result of the high number of automobiles in Tehran. Considering the high number of automobiles in Tehran and the fact that CO and NO$_2$ are produced as a result of the combustion of fossil fuels, the presence of CO and NO$_2$ in one group ($F2$) is rational. According to the results of the factor analysis, O$_3$ is proposed to be grouped individually. O$_3$ or ozone is a colorless gas, formed in a series of complex reactions, in which the presence of sunlight and heat are the main variables. Because of the photochemical characteristic of this reaction, the level of O$_3$ has a seasonal behavior. As the temperature rises and the day gets longer, the level of O$_3$ becomes higher. Finally, SO$_2$ is the variable that is grouped as the least effective factor. SO$_2$ can be emanated from different sources but the sources that have the most proportion are electric power plants and refineries. Existing a lot of small and dispersed electric power plants across the city and Tehran Oil Refinery near the city are the main sources of SO$_2$ in air pollution.

### 6. Concluding remarks

While Tehran has one of the most polluted ambient air in the world and is endangered with harmful damages of air pollution, it has received less attention in the literature, and no one has considered determining the air pollutants that have the greatest impact on air quality. Hence, in this paper, univariate Box-Jenkins stochastic models along with factors analysis are used to predict environmental air pollutants in Tehran and analyze the relationship between air pollutants to determine the factors that have the greatest impact on air quality. In this regard, the behavior of six air pollutants including PM$_{10}$, PM$_{2.5}$, O$_3$, SO$_2$, NO$_2$, and CO in Tehran city over a 6-year timespan is studied. The data for this study are achieved from the Air Quality Control Company (AQCC) which is responsible for monitoring the air quality in Tehran city. Because of the high variability and non-normality of the data, a Yeo-Johnson power transformation is conducted to stabilize and normalize the data. Then, the univariate Box-
Jenkins stochastic models are applied in order to build forecasting models for each of the six air pollutants. The proposed Box-Jenkins models for the air pollutants have a relatively simple form and could be considered as fitted models. Then, the proposed models are used for forecasting the out-of-sample data from 21 March 2018 to 21 June 2018 in which the results reveal the good performance of the proposed methods in both fitting and forecasting the air pollutants.

Since air pollution can be a consequence of many factors, there is a need to study and analyze the origin of air pollutants and their relationships. Therefore, a factor analysis approach is used to categorize the air pollutants and determine the proportion of each of them in the total variability of the air quality. Based on the results of factor analysis, the variables are classified into four groups. The first group that has the largest proportion in air pollution includes PM$_{10}$ and PM$_{2.5}$ with a proportion of 43.2594% of the total variability. The second group includes CO and NO$_2$ with a proportion of 21.6500% of the total variability. Because of the geographical situation of the city and as the combustion of fossil fuels is the main source of emitting PM$_{2.5}$, PM$_{10}$, CO, and NO$_2$, the interpretation of the first and the second group indicates that the major concern of air pollution in Tehran city is related to a high number of automobiles and the quality of fossil fuels. Therefore, decreasing or controlling the number of automobiles and increasing the quality of fossil fuels, can resolve up to 60% of air pollution concerns.

As mentioned in the introduction, air quality modeling using univariate Box-Jenkins stochastic models have been one of the most effective and interesting approaches for researchers and practitioners. But, air quality is a result of many variables, especially weather conditions. Therefore, considering weather condition variables that affect air quality, such as temperature, humidity, precipitation, wind speed, and wind direction, can significantly improve the results of the prediction model. In this way, developing multivariate forecasting models can be an attractive subject for future research. In addition, other methods (such as ANN) can be applied to predict air pollutants and compare their performance with the proposed Box-Jenkins model.

**Nomenclature**

| Abbreviation | Description                                      |
|--------------|--------------------------------------------------|
| AIC          | Akaike Information Criterion                     |
| ADF          | Augmented Dickey-Fuller                          |
| AFNN         | Artificial Fuzzy Neural Networks                  |
| AQCC         | Air Quality Control Company                      |
| ANN          | Artificial Neural Networks                       |
| ARIMA        | Autoregressive Integrated Moving Average         |
| ACF          | Autocorrelation Function                         |
| CO           | Carbon Monoxide                                  |
| CLS          | Conditional Least Squares                         |
| EPA          | Environmental Protection Agency                  |
| HQIC         | Hannan-Quinn Information Criterion               |
| IMO          | Iran Meteorological Organization                 |
| IRI          | Islamic Republic of Iran                         |
| KMO          | Kaiser-Mayer-Oklin                               |
| KS           | Kolmogorov-Smirnov                               |
| MLE          | Maximum Likelihood Estimation                    |
| MAE          | Mean Absolute Error                              |
| MAPE         | Mean Absolute Percentage Error                   |
| MLS          | Minimum Least Squares                             |
| MLR          | Multi Linear Regression                          |
| NAAQS        | National Ambient Air Quality Standards            |
| NO$_2$       | Nitrogen dioxide                                 |
| NO           | Nitrogen monoxide                                |
| O$_3$        | Ozone                                            |
| PACF         | Partial Autocorrelation Function                  |
| PM$_{2.5}$   | Particulate Matter 2.5                           |
| PM$_{10}$    | Particulate Matter 10                            |
| PCA          | Principal Component Analysis                     |
| RMSE         | Root Mean Square Error                            |
| SBIC         | Schwarz’s Bayesian Information Criterion         |
| SARIMA       | Seasonal Autoregressive Integrated Moving Average |
| SO$_2$       | Sulfur dioxide                                   |
| WHO          | World Health Organization                        |

**Acknowledgment**

The authors appreciate Iran Meteorological Organization (IMO) and Air Quality Control Company (AQCC) for their valuable efforts and activities for Tehran air quality. Also, the authors thank the editors and reviewers for their comments, which contributed to the improvement and robustness of the paper.

**References**

1. Heger, M. and Sarraf, M. “Air pollution in Tehran: Health costs, sources, and policies”. *Environment and Natural Resources Global Practice*, World Bank, Washington, DC (2018).

2. Miri, M., Derakhshan, Z., Allahabadi, A., Almadi, E., Oliveri Conti, G., Ferrante, M., and Aval, H.E. “Mortality and morbidity due to exposure to outdoor
air pollution in Mashhad metropolis, Iran. The AirQ model approach”, Envir. Res., 151, pp. 451–457 (2016).

3. Karimzadehan, H., Rahmatian, M., Farland, D.D., and Yunesian, M. “Economic valuation of air pollution health impacts in the Tehran area, Iran”, Iran. J. Public Health, 37(1), pp. 20–30 (2008).

4. Chen, C., Li, C., Li, Y., Liu, J., Meng, C., Han, J., Zhang, Z., and Xu, D. “Short-term effects of ambient air pollution exposure on lung function: A longitudinal study among healthy primary school children in China”, Sci. Total Environ., 645, pp. 1014–1020 (2018).

5. Steinle, S., Reis, S., Sabel, C.E., Semple, S., Twigg, M.M., Brabban, C.F., Lessen, S.R., Heal, M.R., Harrison, D., Lin, C., and Wu, H. “Personal exposure monitoring of PM2.5 in indoor and outdoor microenvironments”, Sci. Total Environ., 508, pp. 383–394 (2015).

6. Iodice, P., Adamo, P., Capozzi, F., Di Palma, A., Senatore, A., Spagnuolo, V., and Giordano, S. “Air pollution monitoring using emission inventories combined with the moose bag approach”, Sci. Total Environ., 541, pp. 1410–1419 (2016).

7. Yousefian, F., Mahvi, A.H., Yunesian, M., Hassanvand, M.S., Khashi, H., and Amini, H. “Long-term exposure to ambient air pollution and autism spectrum disorder in children: A case-control study in Tehran, Iran”, Sci. Total Environ., 643, pp. 1216–1222 (2018).

8. Seifi, M., Nazi, S., Johnson, G., Nodehi, V., and Yunesian, M. “Exposure to ambient air pollution and risk of childhood cancers: A population-based study in Tehran, Iran”, Sci. Total Environ., 646, pp. 105–110 (2019).

9. Brunner, C.R. “National ambient air quality standards. In: Hazardous air emissions from incineration”, pp. 27–37, Springer, Boston, MA (1985).

10. Pope III, C.A., Ezzati, M., and Dockery, D.W. “Fine-particle air pollution and life expectancy in the United States”, New Engl. J. Med., 360(4), pp. 376–386 (2009).

11. Kulmola, J., Pohjola, M., Sokhi, R.S., Luhana, L., Kitwiroo, N., Funglou, L., Rantamäki, M., Berge, E., Ödegård, V., Slordal, L.H., and Denby, B. “Analysis and evaluation of selected local-scale PM2.5 air pollution episodes in four European cities: Helsinki, London, Milan, and Oslo”, Atmos. Environ., 39(15), pp. 2759–2773 (2005).

12. Barra-Berastegi, G., Elías, A., Barona, A., Suenz, J., Ezcurre, A., and Argandillo, J.D. “From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao”, Environ. Modell. Soft., 23(5), pp. 622–637 (2008).

13. Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., and Vitabile, S. “Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy”, Atmos. Environ., 41(14), pp. 2067–2095 (2007).

14. Auffhammer, M. and Carson, R.T. “Forecasting the path of China’s CO2 emissions using province-level information”, J. Environ. Econ. Manag., 55(3), pp. 220–217 (2008).

15. Olabeimiwo, F.A., Danamaliki, G.I. Oyehan, T.A., and Tawabini, B.S. “Forecasting CO2 emissions in the Persian Gulf States”, J. Environ. Sci. Manag., 3(1), pp. 1–10 (2017).

16. Cabaneros, S.M.S., Calautit, J.K.S., and Hughes, B.R. “Hybrid artificial neural network models for effective prediction and mitigation of urban roadside NO2 pollution”, Environ. Proced., 142, pp. 3524–3530 (2017).

17. Kurt, A., Gullaci, B., Karaca, F., and Alagha, O. “An online air pollution forecasting system using neural networks”, Environ. Int., 34(5), pp. 592–598 (2008).

18. Niska, H., Hiltunen, T., Karpinnen, A., Rusjunen, J., and Kolehmainen, M. “Evolving the neural network model for forecasting air pollution time series”, Eng. Appl. Artif. Intel., 17(2), pp. 159–167 (2004).

19. Elsayoumi, M., Ramli, N.A., Yusof, N.F.F.M., Yahaya, A.S.B., Madloun, W.A., and Ul-Saﬁe, A.Z. “Multivariate methods for indoor PM10 and PM2.5 modelling in naturally ventilated schools buildings”, Atmos. Environ., 94, pp. 11–21 (2014).

20. Amini, H., Taghavi-Shahri, S.M., Henderson, S.B., Naddafi, K., Nabizadeh, R., and Yunesian, M. “Land use regression models to estimate the annual and seasonal spatial variability of sulfur dioxide and particulate matter in Tehran, Iran”, Sci. Total Environ., 488, pp. 343–353 (2014).

21. Lee, M., Bruer, M., Hong, P., Tang, R., Tsui, T.H., Choi, C., Cheng, W., Lai, P.C., Tian, L., Thach, T.Q., Allen, R., and Barratt, B. “Land use regression modelling of air pollution in high density high rise cities: A case study in Hong Kong”, Sci. Total Environ., 592, pp. 306–315 (2017).

22. Vlachogianni, A., Kassomenos, P., Karpinnen, A., Karakitsios, S., and Kulmola, J. “Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki”, Sci. Total Environ., 409(8), pp. 1559–1571 (2011).

23. Muñoz, E., Martín, M.L., Turias, I.J., Jimenez-Comé, M.J., and Trujillo, F.J. “Prediction of PM10 and SO2 exceedances to control air pollution in the Bay of Algeciras, Spain”, Stoch. Env. Res. Risk A, 28(6), pp. 1409–1420 (2014).
24. Allamsetty, S. and Mohapatra, S. “Response surface methodology-based model for prediction of NO and NO₂ concentrations in nonthermal plasma-treated diesel exhaust”, *S.N. Appl. Sci.*, 1(2), p. 189 (2019).

25. Lee, C.P., Lin, W.C., and Yang, C.C. “A strategy for forecasting option prices using fuzzy time series and least square support vector regression with a bootstrap model”, *Sci. Iran.*, 21(3), pp. 815–825 (2014).

26. Syu, Y., Kuo, J.Y., and Fanjiang, Y.Y. “Time series forecasting for dynamic quality of web services: An empirical study”, *J. Syst. Soft.*, 134, pp. 279–303 (2017).

27. Ruby-Figueroa, R., Suavedra, J., Bahamonde, N., and Casasano, A. “Permeate flux prediction in the ultrafiltration of fruit juices by ARIMA models”, *J. Membrane Sci.*, 524, pp. 108–116 (2017).

28. Gao, Y., Shang, H.L., and Yang, Y. “High-dimensional functional time series forecasting: An application to age-specific mortality rates”, *J. Multivariate Anal.*, 170, pp. 232–243 (2018).

29. Cinar, Y.G., Mirissee, H., Goswami, P., Gaussen, E., and Ait-Bachir, A. “Period-aware content attention RNNs for time series forecasting with missing values”, *Neurocomputing*, 312 (2018).

30. Martínez, F., Frías, M.P., Perez-Godoy, M.D., and Rivera, A.J. “Dealing with seasonality by narrowing the training set in time series forecasting with LNN”, *Expert Syst. Appl.*, 103, pp. 38–48 (2018).

31. Huang, C.H., Yang, F.H., and Lee, C.P. “The strategy of investment in the stock market using modified support vector regression model”, *Sci. Iran.*, 25(3), pp. 1629–1640 (2018).

32. Sagheer, A. and Koth, M. “Time series forecasting of petroleum production using deep LSTM recurrent networks”, *Neurocomputing*, 323, pp. 203–213 (2018).

33. Suhermi, N., Suhartono, Prastyo, D.D., and Ali, B. “Roll motion prediction using a hybrid deep learning and ARIMA model”, *Procedia Comput. Sci.*, 144, pp. 251–258 (2018).

34. Entezami, A., Shariatmadar, H., and Karamodin, A. “Improving feature extraction via time series modeling for structural health monitoring based on unsupervised learning methods”, *Sci. Iran.*, 27(3), pp. 1001–1018 (2020). DOI: 10.24200/SCI.2018.30641.

35. Ohyver, M. and Pudiljastuti, H. “Arima model for forecasting the price of medium quality rice to anticipate price fluctuations”, *Procedia Comput. Sci.*, 135, pp. 707–711 (2018).

36. Singh, A.S.N. and Mohapatra, A. “Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting”, *Renew. Energ.*, 136, pp. 758–768 (2019).

37. Jacobson, M.Z. and Jacobson, M.Z., *Fundamentals of Atmospheric Modeling*, Cambridge university press (2005).

38. Pankratz, A., *Forecasting with univariate Box-Jenkins Models: Concepts and Cases*, John Wiley & Sons (2009).

39. Kumar, U. and Jain, V.K. “ARIMA forecasting of ambient air pollutants (O₃, NO, NO₂ and CO)”, *Stoch. Env. Res. Risk A.*, 24(5), pp. 751–760 (2010).

40. Zhou, M. and Goh, T.N. “Air quality modeling via PM2.5 Measurements”, *Theory and Practice of Quality and Reliability Engineering in Asia Industry*, pp. 197–210, Springer, Singapore (2017).

41. Jian, L., Zhaoa, Y., Zhu, Y.P., Zhang, M.B., and Bertolatti, D. “An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadway in Hangzhou, China”, *Sci. Total Environ.*, 426, pp. 336–345 (2012).

42. Díaz-Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., and Moncada-Herrera, J.A. “A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile”, *Atmos. Environ.*, 42(35), pp. 8331–8340 (2008).

43. Samia, A., Kaoutther, N., and Abdelwahed, T. “A hybrid ARIMA and artificial neural networks model to forecast air quality in urban areas: case of Tunisia”, *Adv. Mat. Res.*, 518, pp. 2900–2979 (2012).

44. Hoi, K.L., Yuen, K.V., and Mol, K.M. “Prediction of daily averaged PM₁₀ concentrations by statistical time-varying model”, *Atmos. Environ.*, 43(16), pp. 2579–2581 (2009).

45. Genc, D.D., Yesilyurt, C., and Tuncel, G. “Air pollution forecasting in Ankara, Turkey using air pollution index and its relation to assimilative capacity of the atmosphere”, *Environ. Monit. Assess.*, 166(1), pp. 11–27 (2010).

46. Poggi, J.M. and Portier, B. “PM₁₀ forecasting using clusterwise regression”, *Atmos Environ.*, 45(38), pp. 7005–7014 (2011).

47. Goecheva-Ilieva, S.G., Ivanov, A.V., Voynikova, D.S., and Boyadzhiev, D.T. “Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach”, *Stoch. Env. Res. Risk A.*, 28(4), pp. 1045–1060 (2014).

48. Cortina-Jaenuchs, M.G., Quintanilla-Dominguez, J., Vega-Corona, A., and Andina, D. “Development of a model for forecasting of PM₁₀ concentrations in Salamanca, Mexico”, *Atmos. Pollut. Res.*, 6(4), pp. 630–634 (2015).

49. Jiang, P., Dong, Q., and Li, P. “A novel hybrid strategy for PM₂.₅ concentration analysis and prediction”, *J. Environ. Manage.*, 196, pp. 443–457 (2017).
50. Abdollahimzadeh, L., Azadpour, M., and Zarandi, M.H.F. “Two hybrid expert system for diagnosis Air Quality Index (AQI)”, North American Fuzzy Information Processing Society Annual Conference, India, New Delhi, pp. 315–322 (2018).
51. https://www.worldbank.org/en/news/infographic/2016/09/08/death-in-the-air-air-pollution-costs-money-and-lives.
52. Shahlahzi, H., Reyhanian, M., Hosseini, V., and Afshin, H. “The relative contributions of mobile sources to air pollutant emissions in Tehran, Iran: an emission inventory approach”, Em. Cont. Sci. Tech., 2(1), pp. 41–56 (2016).
53. Shahlahzi, H., Ganjiazaad, R., Hosseini, V., and Hamedi, M. “Investigating the influence of traffic emission reduction plans on Tehran air quality using WRF/CAMx modeling tools”, Transport. Res. D-TR E, 57, pp. 484–495 (2017).
54. Hosseini, V. and Shahlahzi, H. “Urban air pollution in Iran”, Iran Stud-UK, 49(6), pp. 1020–1046 (2016).
55. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., and Ljung, G.M., Time Series Analysis: Forecasting and Control, John Wiley & Sons (2015).
56. Wold, H. “A study in the analysis of stationary time series”, Doctoral Dissertation, Almqvist & Wiksell (1938).
57. Yee, I.K. and Johnson, R.A. “A new family of power transformations to improve normality or symmetry”, Biometrika, 87(4), pp. 954–959 (2000).
58. Box, G.E.P. and Cox, D.R. “An Analysis of Transformations”, J. R. Stat. Soc. B., 26(2), pp. 211–252 (1964).
59. Box, G.E.P., Hunter, J.S., and Hunter, W.G., Statistics for Experimenters: Design, Innovation, and Discovery, Wiley-Interscience, New York (2005).
60. Bozdogan, H. “Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions”, Psychometrika, 52(3), pp. 345–370 (1987).
61. Sin, C.Y. and White, H. “Information criteria for selecting possibly misspecified parametric models”, J. Econometrics, 71(1), pp. 207–225 (1996).
62. Kullback, S. and Leibler, R.A. “On information and sufficiency”, Ann. Math. Stat., 22(1), pp. 79–86 (1951).
63. Schwarz, G. “Estimating the dimension of a model”, Ann. Stat., 6(2), pp. 461–464 (1978).
64. Hurvich, C.M. and Tsai, C.L. “Regression and time series model selection in small samples”, Biometrika, 76(2), pp. 297–307 (1989).
65. Hannan, E.J. and Quinn, B.G. “The determination of the order of an autoregression”, J. R. Stat. Soc. B., 41(2), pp. 190–195 (1979).
66. Brockwell, P.J. and Davis, R.A., Introduction to Time Series and Forecasting, Springer (2016).
67. Shumway, R.H. and Stoffer, D.S. “Time series analysis and its applications”, Stud. Inform. Control, 9(1), pp. 375–376 (2000).
68. Mulaik, S.A. “Foundations of factor analysis”, Chapman and Hall/CRC (2009).
69. Bartlett, M.S. “Tests of significance in factor analysis”, Brit. J. Statist. Psych., 3(2), pp. 77–85 (1950).
70. Jolliffe, I. “Principal component analysis”, International Encyclopedia of Statistical Science, pp. 1094–1096, Springer (2011).
71. Kumar, R. and Joseph, A.E. “Air pollution concentrations of PM$_{2.5}$, PM$_{10}$ and NO$_2$ at ambient and kerbsite and their correlation in metro city- Mumbai”, Environ. Monit. Assess., 119(1), pp. 191–190 (2006).
72. Asadollahfard, G., Zamanian, M., Mirmohammadi, M., and Asadi, M. “Air pollution study using factor analysis and univariate Box-Jenkins modeling for the northwest of Tehran”, Adv. Environ. Res., 4(4), pp. 233–246 (2015).

Appendix A

**Time series plots**

The time series plot of the air pollutants CO, NO$_2$, SO$_2$, O$_3$, PM$_{2.5}$, and PM$_{10}$ is presented in Figures A.1–A.6, respectively.

**ACF and PACF plots**

The ACF and PACF plots of the six air pollutants including CO, NO$_2$, SO$_2$, O$_3$, PM$_{2.5}$, and PM$_{10}$ are illustrated in Figures B.1–B.6 to specify the potentially appropriate models for each of the air pollutants.

![Figure A.1. The time series plot of CO (µg/m$^3$).](image-url)
Figure A.2. The time series plot of NO$_2$ (µg/m$^3$).

Figure A.3. The time series plot of SO$_2$ (µg/m$^3$).

Figure A.4. The time series plot of O$_3$ (µg/m$^3$).

Figure A.5. The time series plot of PM$_{2.5}$ (µg/m$^3$).

Figure A.6. The time series plot of PM$_{10}$ (µg/m$^3$).
Figure B.1. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of initial CO data.

Figure B.2. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of initial NO₂ data.

Figure B.3. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of initial SO₂ data.

Figure B.4. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of initial O₃ data.
Figure B.5. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of initial PM$_{2.5}$ data.

Figure B.6. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of initial PM$_{10}$ data.

Biographies

Jalal Delaram is a PhD candidate at the Department of Industrial Engineering at the Sharif University of Technology, Tehran, Iran. He received his BSc and MSc degrees in Industrial Engineering from the Sharif University of Technology. His major fields of research are cloud manufacturing, Computer Integrated Manufacturing (CIM), Manufacturing Operations Management (MOM), and statistical modeling and analysis.

Majid Khedmati received his PhD in Industrial Engineering from the Sharif University of Technology, Iran, in 2015. He is now an Assistant Professor of Industrial Engineering at the Sharif University of Technology. He received his BSc and MSc degrees both in Industrial Engineering from Iran University of Science and Technology and the Sharif University of Technology in 2010 and 2012, respectively. His research interests are in the areas of data science, machine learning, quality engineering, and applied statistics.