*Journal of Air Pollution and Health (Spring 2017); 2(2): 95-108*

**Original Article**

Available online at http://japh.tums.ac.ir

# $PM_{10}$ CONCENTRATION PREDICTION FOR AREAS WITH NO UP-DATING MONITORING SYSTEM USING AUTO – REGRESSIVE GROUP METHOD OF DATA HANDLING NEURAL NETWORK

*Fatemeh Kardel[1], Amirreza Lashkari[2], Manoochehr Babanezhad[3]**

[1] Department of Environmental Science, Faculty of Science, University of Mazandaran, Babolsar, Mazandaran, Iran

[2] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

[3] Department of Statistics, Faculty of Science, Golestan University, Gorgan, Golestan, Iran

**ARTICLE INFORMATION**

*Article Chronology:*

Received 20 February 2017
Revised 21 May 2017
Accepted 21 June 2017
Published 29 June 2017

*Keywords:*

$PM_{10}$ concentration; air pollutants; meteorological factors; AR-GMDH neural network; time series

**CORRESPONDING AUTHOR:**

m.babanezhad@gu.ac.ir
Tel: ( +98 17) 32245882
Fax: ( +98 17) 32245882

**ABSTRACT:**

**Introduction:** Predicting $PM_{10}$ concentration as a significant risk factor for anumber of pollution related diseases has been recently inevitable task for areas with high population density particularly for areas with no updating monitoring systems. This study aims to illustrate how $PM_{10}$ concentration level can be predicted by the prior information of the air pollutants and the meteorologicalfactors in urban areas.

**Materials and methods:** The data we used are measured from four monitoringstations in the city of Tehran between January 2012 and December 2014. We use the Auto-regressive group method of data handling (AR - GMDH) neuralnetwork approach which employees the prior stationary time series data setting.

**Results:** Our results demonstrate that $PM_{10}$ concentration level for a specific dayis more likely to be predictable by sulfur dioxide ($SO_2$) and nitrogen dioxide ($NO_2$) than the carbon monoxide (CO) concentrations, and also show that$PM_{10}$ concentration is positively associated with precipitation and wind speedand with high temperature. The accuracy of the predicted values of the $PM_{10}$ concentration is evaluated by inspecting the coefficient of determination, meansquared error, the square root of mean squared error, mean absolute deviation, and index of agreement.

**Conclusions:** The AR - GMDH algorithm can be proposedin comparison with the chemical and physical approaches due to its accuracyand simplicity, and its cost efficiency.

## INTRODUCTION

It is well known that urban air quality in many metropolitan cities is adversely affected by air pollutants such as particulate matter 10 □m or less in diameter ($PM_{10}$), sulfur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), and carbon monoxide (CO). According to the World Health Organization (WHO), air pollution is a main environmental risk to human health and causes annually 7 million premature deaths in the world [1, 2]. Among the air pollutionfactors listed, $PM_{10}$ is an important factor in the analysis of air quality. This

Please cite this article as : Kardel F, Lashkari A, Babanezhad M. $PM_{10}$ concentration prediction for areas with no updating monitoring system using auto – regressive group method of data handling neural network. Journal of Air Pollution and Health. 2017; 2(2): 95 - 108

may be because studies showed that particle pollution exposure to a variety of health problems [3 - 7]. Therefore, daily reporting of the $PM_{10}$ concentration level for a city with high population density such as Tehran is very important. Tehran has 20 air quality monitoring stations. Among these, there are stations where their data may not be able to be reported during one or several days due to technical reasons. Thus, predicting $PM_{10}$ concentration level for citizens who live around these stations seems to be a vital task. Predicting $PM_{10}$ concentration level in the atmosphere was always controversial subject, and often raises the question whether $PM_{10}$ concentration level is associated with the other air pollutants, or it is also associated with the meteorological factors. While studies have not directly addressed the association between $PM_{10}$ concentration and some other air pollutants and meteorological factors, some studies predicted the values of an air pollutant factor in Tehran, and also measured the air quality index and its temporal trend in Tehran [8 - 10]. Further, some studies investigated the influence of meteorological conditions and atmospheric circulation types on the $PM_{10}$ concentration, and also used a synoptic climatologically approach for geographical analysis to assess the $SO_2$ concentrations [5, 11]. As the major primary sources of $PM_{10}$ inurban areas are road traffic and chemical reactions such as atmospheric oxidation of $SO_2$, $NO_2$, CO, growing evidence indicates that $SO_2$, $NO_2$, CO seem to be associated with $PM_{10}$ concentration. Further, due to the combination effects of different factors such as meteorological, topographical, physical, and chemical factors [2, 5, 12, 13, 14], evidence indicates that meteorological variables such as temperature, precipitation, and wind speed are also seem to be associated with $PM_{10}$ concentration [3, 11, 15]. Investigating the association between the air pollution factors and the meteorological variables, studies often used the time series analysis and the graphical models due to the simplicity and the cost efficiency [8, 9, 16, 17, 18, 19, 20, 21]. For instance, [9] performed the auto-regressive integrated moving average (ARIMA) model to predict $SO_2$ in Tehran for the period of 2000 - 2005, and [9] used time series regression model to analyze air quality index in Tehran from the daily average data of the period of 2002 - 2012. The graphical multivariate time series model has also been employed to identify the inter - relationship between air pollutants such as $SO_2$, $NO_2$, and CO [16, 19]. Due to the ambiguous relationship between the air pollution factors with time series data setting, analyzing statistical models were extended to the neural network model [22] developed two artificial neural network models to predict hourly average $NO_2$, $O_3$, CO and $PM_{10}$ concentrations by considering different temporal averages of inputmeteorological parameters [23] used a PCA- based artificial neural network modeling approach to predict the daily mean concentrationsof $PM_{10}$ and $PM_{2.5}$ in Thessaloniki and Helsinki [24] developed various artificial neural network models for the Greater Athensarea for predicting hourly average $PM_{10}$ concentration. [5] used a complex time series analysis of $PM_{10}$ and $PM_{2.5}$ for a coastalsite using artificial neural network modeling and k-means clustering. [25] applied a neural network forecast for daily average $PM_{10}$ concentrations in Belgium. [14] used a neural network and data-pre-selection framework for accurately predicting the concentrations of $PM_{10}$ in the metropolitan region of Lisbon, Portugal. This study aims to predict $PM_{10}$ concentration level for a specific day through information of the $SO_2$,$NO_2$, CO, temperature, precipitation, and wind speed for areas with no updating monitoring station in the city of Tehran. We apply a new approachnamed auto-regressive group method of data handling (AR- GMDH) neural network which employees the prior stationary time series information of the $PM_{10}$ concentration and uses the prior information of the air pollutants and the meteorological factors listed. We used the data were measured from fourmonitoring stations between January 2012 and December 2014. Specifically,we seek to explore the following research questions: (1) What are the differences in the daily average of the $PM_{10}$ concentration between the four consideredmonitoring stations to figure out an exceed-

ance of federal $PM_{10}$ (t) standards( > 150 μ/$m^3$ ) ? (2) How the average of $PM_{10}$ (t) concentration for a specific dayt can be predicted by values of $SO_2$, $NO_2$, CO, temperature, precipitation,and wind speed by using the AR- GMDH algorithm ? (3) How accurate is the predicted values of the $PM_{10}$ concentration? in the city of Tehran.

## MATERIALS AND METHODS

### *Study areas*

Our study area is city of Tehran situated in the northern center of Iran. Tehran is Iran's capital city and is the most populated city of Iran with an estimated population of 8.7 million. Tehran is amongst a few capitals of the world, that is not located around a river or even close to the sea, and mountains surround the city from the north and east. Tehran is divided into 22 municipal regions and it has 20 air pollution monitoring stations. Due to differences between these regions, we considered four monitoring stations were located in the north,west, east, and south areas of Tehran which respectively numbered by 1, 5, 8, and 20 municipal regions with different levels of air pollution. Monitoring data used in this study were obtained from Tehran air quality database maintained by Tehran AQCC (Air Quality Control Company). The daily average concentrations of $PM_{10}$, $SO_2$, $NO_2$, CO, and also daily average temperature, total daily precipitation, and daily average wind speed in knots were measured from all monitoring stations between January 2012 and December 2014. During this period, the annual mean temperature was 18.2 °C and the highest and the lowest temperatures were 42.6 °C in Summer and -3.2 °C in Winter. The annual mean precipitation was about 231 mm; and the annual mean wind speed was 5.9 in Knots [8 - 10]. Note that, the meteorological monitoring stations are located in the same municipal regions of the four air pollutant monitoringstations.

### *Data setting*

Our data were measured from four monitoring stations located in the north, east, west, and south areas of Tehran between January 2012 and December 2014. Statistical time series method is often established based on the assumption that the time series can be rendered weakly stationary (mean, variance, and auto- correlation of a time series to be constant over the time t ). Therefore, we need to examine whether our time series data is weakly stationary or not. By plotting the original data, we observed that our data in four areas have clear seasonal patterns (variations in annual cycle of about 365 days). This means our data is potentially non- station in the variance. To make our data weakly stationary, we carried out two operations. First, we adopted the Box-Cox transformation by taking the natural logarithm of the data [26]. We then deseasonalized our data after taking the natural logarithm [27]. After deseasonalization, there was no evidence of non- stationary in data by observing the plots of auto- correlation (correlation of a variable between a given time tand a lagged version of itself) function. We called our weakly stationary dataset to smoothed data. Fig. 1 shows the smoothed daily average of $PM_{10}$, $SO_2$, $NO_2$, and CO concentrations measured from four considered monitoringstations between January 2012 and December 2014. We began by performinga linear time series regression model to figure out how daily average $PM_{10}$ (t) concentration at day t is explained by $SO_2$, $NO_2$, CO, temperature, precipitation, and wind speed both at the same day t and at the previous day $t - 1$. The coefficient of determination of the fitted time series regression model wassmall ($R^2 = 0.08$), and a nonrandom pattern for the residuals was observed. This indicates that nonlinear relationships between these factors may be required. Further, we found that $PM_{10}$ (t) on day t is likely more associated withthe listed factors on the previous day $t - 1$ than on the present day t for all fourmonitoring stations. Due to nonlinearity of the relationships between $PM_{10}$ and the listed factors, we used the Group method of data handling (GMDH) neural network algorithm [28, 29]. GMDH is a family of inductive algorithmfor analyzing multi-parametric datasets. This algorithm gives possibility tofind automatically in-

terrelations between several factors to select an optimalstructure of the relationship. We describe the basic features of the GMDH neural network algorithm in next section.

***Group method of data handling (GMDH)***

The GMDH neural network is in fact a model to estimate a high-order polynomial which relates the input vector x = [$x_1$, $x_2$, ..., $x_m$] to the output variable y [29, 30]. The form of this model is generally as follows, Eq. (1):

$$y=a+\sum_{i=1}^{m} b_i x_i+\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j+ \sum_{k=1}^{m} d_{ijk} x_i x_j+\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} d_{ijk} x_i x_j x_k+\ldots \quad (1)$$

Where coefficients a, $b_i$, $c_{ij}$ , $d_{ijk}$, ... are the parameters of the model (1). Thetraining process of the GMDH algorithm consists of the three following steps:

Step 1: For m features, there will be sets consisting of two features. For these sets, the least squares polynomial, which best fits the observations,will be calculated. The form of this polynomial is indicated as follows, Eq. (2):

$$\hat{y} = A + Bx_i + Cx_j + Dx^2_i + Ex^2_j + Fx_ix_j \quad (2)$$

Where $x_i$ and $x_j$ are the features that the polynomial is estimated upon, by isthe estimated output, and thecoefficients A, B, C, D, E, F are the parametersof the polynomial (2). In sum, polynomials should be estimated. Then, thepolynomials are evaluated at the n input data points, and the results wouldbe stored in matrix Z = . It can be seen that the input data $X_{n\times m}$ is transformed into Z. The goal of this step is to select a set of new features (columns of matrix Z) that best estimates the output variable y.

Step 2: In this step, the features from the matrix Z that are not effective will be reduced. First, vector d = [$d_1$, $d_2$,..., )] is formed from the least square errors ofeach column of Z which is shown in Eq. (3). Then, only the column of Z where it's corresponding least square error is lower than the predeterminedvalue k would be selected.

$$d^2_j=\sum_{i=1}^{n}(y_i - z_{ij})^2 \quad (3)$$

$$d_{min}=\min(dj) \quad (4)$$

Where j =1,2,…,

Step 3: In this step, the convergence of the model will be tested. In each iteration, steps 1 and 2 are repeated and the value dmin in Eq. (4), is computed. If $d_{min}$ of the current iteration is greater than the $d_{min}$ obtained from the previous iteration, the procedure will be stopped, and the modeling the previous step is assumed to be the best fitted model. It is commonto plot the values of $d_{min}$ of all the iterations to make a curve. Although the GMDH algorithm does not theoretically prove that the $d_{min}$ curve has a single minimum value, the experimental results indicate that the single minimum value of $d_{min}$ will always be achieved [11].

## RESULTS AND DISCUSSION

The daily average of $PM_{10}$ (t), $SO_2$ (t), $NO_2$ (t), and CO (t) concentration values plotted in Fig. 1 reveal that there is a nonlinear oscillating characteristicsin all four considered monitoring stations between January 2012 and December 2014. We also observed from Fig. 1 that the higher values of $PM_{10}$ (t) concentration occurred during Winter and Summer months and lower ones occurred during Autumn and Spring months. Further, the distribution of the daily average of $PM_{10}$ (t), $SO_2$ (t), $NO_2$ (t), and CO (t) concentration values from four monitoring stations are summarized in Table 1.

Table 1 shows that the highest values of $PM_{10}$ (t), $SO_2$ (t) and $NO_2$ (t) occurred in east area, and the lowest values occurred in west area. This implies that the high and lowvalues of $PM_{10}$ occurred in areas with high and low values of $SO_2$ and $NO_2$. This suggests that $PM_{10}$ concentration is strongly associated with $SO_2$, and $NO_2$. The highest value of CO (t) occurred in south area, and also the CO (t) concentration in east area is higher than the north and west areas. This may suggest that

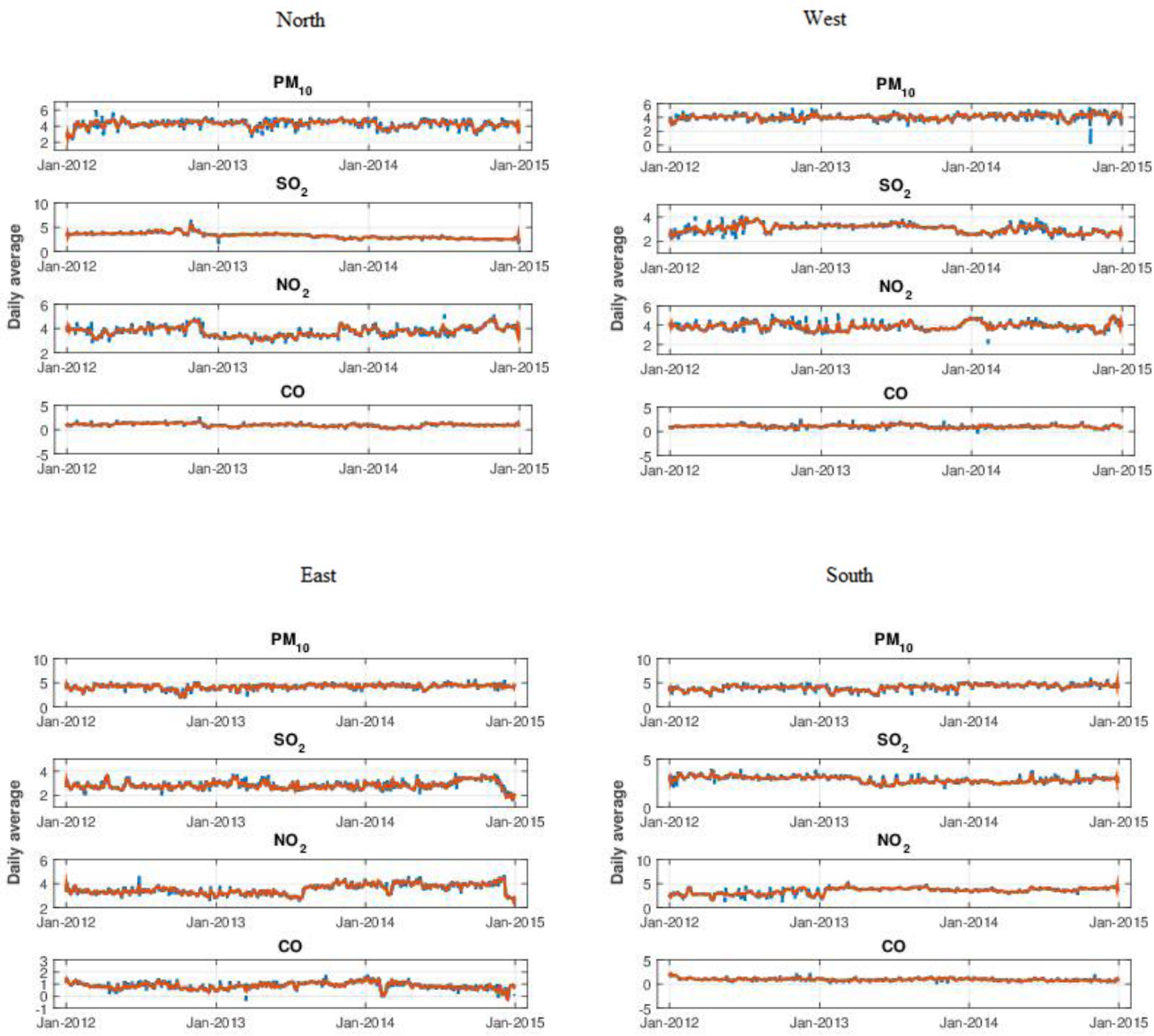


Fig. 1. Daily average of $PM_{10}$, $SO_2$, $NO_2$, and CO concentrations measured from the four monitoring stations located in the north, west, east, and south areas of Tehran between January 2012 and December 2014 after smoothing. Note that the dotted lines show the data after taking the logarithm, and the solid lines show the data after the deseasonalization.

$PM_{10}$ concentration is associated with CO concentration. Moreover, our datashowed that 8.2 %, 6.8 %, 5.9 %, and 5.6 % of days (24 - average) occurred an exceedance of federal $PM_{10}$ (t) standards ( > 150 μ/m$^3$) between January 2012 and December 2014 at east, south, north, and west areas respectively. In time series regression analysis, we found also that precipitation and wind speed were positively associated with $PM_{10}$ concentration. However, high $PM_{10}$ concentration values were observed at the high temperature (temperature more than 34 ◦C) days. To answer the question whether $PM_{10}$ (t) concentration at day t can be predicted by $SO_2$ (t − 1), $NO_2$ (t − 1), CO (t − 1), temperature (t - 1), precipitation (t - 1), and wind speed (t - 1) at the previous day t − 1, we performed the GMDH neural network. In this algorithm, we considered $PM_{10}$ (t) concentration at day t as output and $SO_2$ (t − 1), $NO_2$ (t − 1), CO (t − 1), temperature (t - 1), precipitation

Table 1. Distribution of the daily average of $PM_{10}$, $SO_2$, $NO_2$, and CO concentrations measured at the north, west, east, and south monitoring stations in Tehran between January 2012 and December 2014.

| Air pollutants | Station | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| PM10 | North | 8 | 354 | 50.5 | 34 |
| | West | 6 | 434 | 55.6 | 37.4 |
| | East | 7 | 416 | 77.1 | 45 |
| | South | 7 | 590 | 67.1 | 45 |
| SO2 | North | 5 | 347 | 40.5 | 35.9 |
| | West | 4 | 413 | 46 | 36.8 |
| | East | 4 | 491 | 58.7 | 58.9 |
| | South | 6 | 479 | 35 | 40.5 |
| NO2 | North | 5 | 214 | 30 | 22.2 |
| | West | 3 | 191 | 39 | 24.5 |
| | East | 4 | 322 | 43 | 26 |
| | South | 2 | 193 | 40 | 24.7 |
| CO | North | 0.8 | 203 | 17 | 24.2 |
| | West | 0.8 | 227 | 15 | 22.7 |
| | East | 0.8 | 210 | 19 | 26.2 |
| | South | 0.9 | 329 | 22 | 41.9 |

(t - 1), and wind speed (t - 1) at the previous day t − 1 as inputs. By this algorithm, the data set is divided into the training andthe test sets in order to examine the performance of the GMDH algorithm on unseen data. The training set contains 70 % of the original data and the test set contains 30 % of the original data. Finally, the prediction values of $PM_{10}$ (t) (outputs) are generated. The four columns of the left hand side of Fig. 3 shows the predicted and the observed values of the daily average of $PM_{10}$ concentration obtained by the GMDH algorithm for the north, west, east, andsouth areas respectively. The deviation between the predicted and the observedvalues on the left hand side of Fig. 3 yields that $PM_{10}$ (t) concentration at day t can not accurately be predicted by $SO_2$ (t − 1), $NO_2$ (t − 1), CO (t −1),temperature (t - 1), precipitation (t - 1), and wind speed (t - 1) at the previous day t − 1. To improve our algorithm, we focused on the additional significantinputs. Through auto-correlation analysis, we have found that the $PM_{10}$ (t − 1) and $PM_{10}$ (t − 2) values may be two significant inputs. To determine the association between $PM_{10}$ (t) and $PM_{10}$ (t − 1), and $PM_{10}$ (t − 2), we plotted the auto- correlation function (ACF) and the partial auto- correlation function (PACF) of $PM_{10}$ (t) against lags k=1, 2,..., 20. The resulting correlograms of $PM_{10}$ (t) of all four monitoring stations are shown in Fig. 2. We observedfrom Fig. 2 that ACF are significant up to lag 20, while PACF are significant only at the first lag. This suggests that $PM_{10}$ (t) is an Auto-regressiveprocess of order 1 (abbreviated to AR (1)), that is $PM_{10}$ (t) is strongly associated with $PM_{10}$ (t − 1). Note that the horizontal solid line indicates twoconfidence limits for the ACF and PACF in Fig. 2.We then performed againthe GMDH algorithm where $PM_{10}$ (t − 1), $SO_2$ (t − 1), $NO_2$ (t − 1), CO (t − 1), temperature (t - 1), precipitation (t - 1), and wind speed (t - 1) are considered astime series inputs and $PM_{10}$ (t) as a time series output. Therefore, the GMD Halgorithm is in fact improved by the Auto- regressive GMDH (AR-GMDH) algorithm. The term Autoregressive (AR) on the GMDH is used to describe theparadigm that brings together a number

of time series variables to provide theoutput and the inputs variables. In order to compare the observed values withthe predicted values obtained by the GMDH and AR- GMDH algorithms, we calculated the coefficient of determination ($R^2$), mean squared error (MSE), root- mean squared error (RMSE), mean absolute deviation (MAD), and index of agreement (IA). The listed statistical criterions are defined as follows,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi_{predicted} - yi_{observed})^2}{\sum_{i=1}^{n}(yi_{observed} - \bar{y}_{observed})^2} \quad (5)$$

$$MSE = \frac{\sum_{i=1}^{n}(yi_{predicted} - yi_{observed})^2}{n} \quad (6)$$

$$RMSE = \left[\frac{\sum_{i=1}^{n}(yi_{predicted} - yi_{observed})^2}{n}\right]^{\frac{1}{2}} \quad (7)$$

$$MAD = \frac{\sum_{i=1}^{n}|yi_{predicted} - yi_{observed}|}{n} \quad (8)$$

$$I A = 1 - \frac{\sum_{i=1}^{n}(yi_{predicted} - yi_{observed})^2}{\sum_{i=1}^{n}(|yi_{predicted} - \bar{y}_{observed}| + |yi_{observed} - \bar{y}_{observed}|)^2} \quad (9)$$

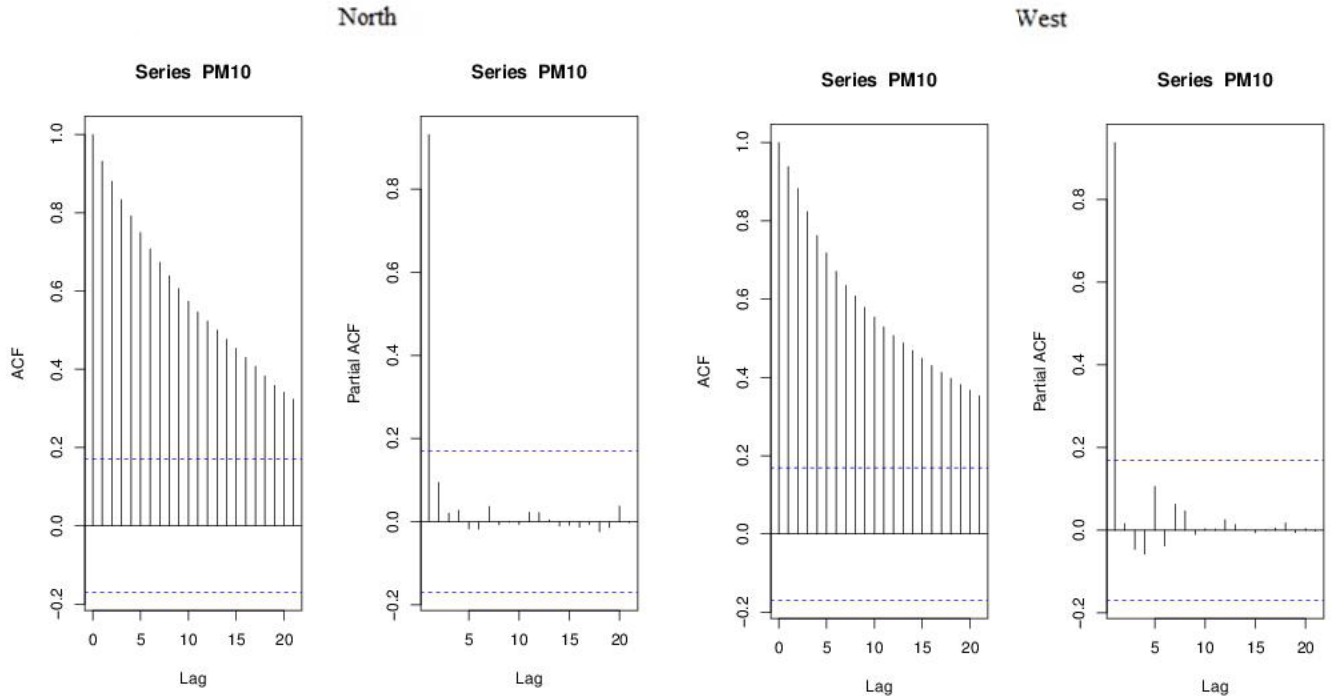


Fig.2. Correlograms of auto- correlations and partial auto- correlations of the smoothed daily average of $PM_{10}$ concentration measured at the north, west, east, and south monitoring stations of Tehran between January 2012 and December 2014.

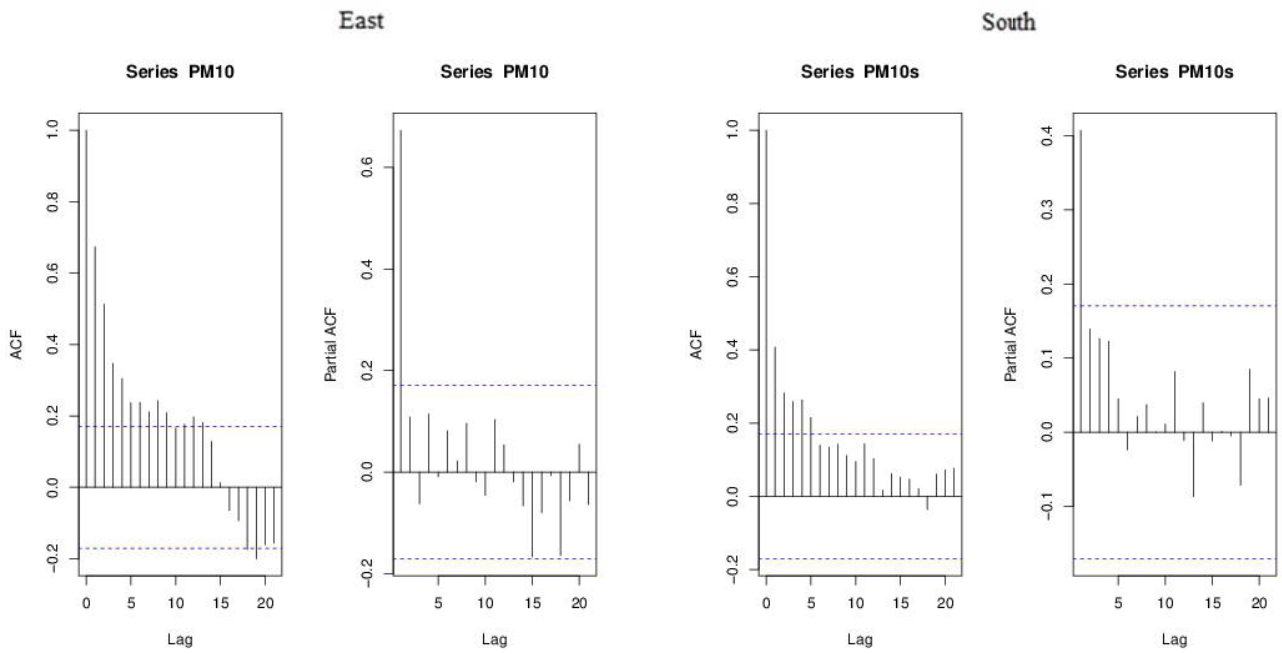


Fig.2. Correlograms of auto- correlations and partial auto- correlations of the smoothed daily average of $PM_{10}$ concentration measured at the north, west, east, and south monitoring stations of Tehran between January 2012 and December 2014.

Table 2. The values of the statistical criterions of the train and the test sets obtained by the AR-GMDH and the GMDH neural networks for the north, west, east, and south monitoring stations of Tehran between January 2012 and December 2014.

| Model | Station | Statistic | $R^2$ | RMSE | MSE | MAD | IA |
|---|---|---|---|---|---|---|---|
| GMDH | North | Train | 0.095 | 0.372 | 0.139 | 0.273 | 0.373 |
| | | Test | 0.077 | 0.375 | 0.141 | 0.276 | 0.372 |
| | West | Train | 0.120 | 0.298 | 0.089 | 0.239 | 0.426 |
| | | Test | 0.098 | 0.301 | 0.091 | 0.242 | 0.413 |
| | East | Train | 0.145 | 0.409 | 0.168 | 0.301 | 0.466 |
| | | Test | 0.134 | 0.412 | 0.171 | 0.304 | 0.460 |
| | South | Train | 0.066 | 0.519 | 0.270 | 0.404 | 0.306 |
| | | Test | 0.044 | 0.524 | 0.276 | 0.409 | 0.302 |
| AR-GMDH | North | Train | 0.949 | 0.087 | 0.008 | 0.066 | 0.987 |
| | | Test | 0.947 | 0.089 | 0.008 | 0.067 | 0.986 |
| | West | Train | 0.935 | 0.080 | 0.006 | 0.061 | 0.983 |
| | | Test | 0.933 | 0.082 | 0.007 | 0.062 | 0.983 |
| | East | Train | 0.954 | 0.094 | 0.009 | 0.071 | 0.988 |
| | | Test | 0.953 | 0.095 | 0.009 | 0.072 | 0.988 |
| | South | Train | 0.967 | 0.096 | 0.009 | 0.074 | 0.992 |
| | | Test | 0.966 | 0.097 | 0.010 | 0.074 | 0.991 |

The results are summarized in Table 2 and plotted in the four columns ofthe right hand side of Fig. 3 and the scatter plots of two algorithms performance on the train and the test data sets based on the predicted and theobserved correlation of the daily average of $PM_{10}$ concentration for north, west, east, and south areas in Fig. 4. Table 2 shows that the AR- GMD Halgorithm is performing better than the GMDH algorithm in all monitoringstations by calculating MSE,

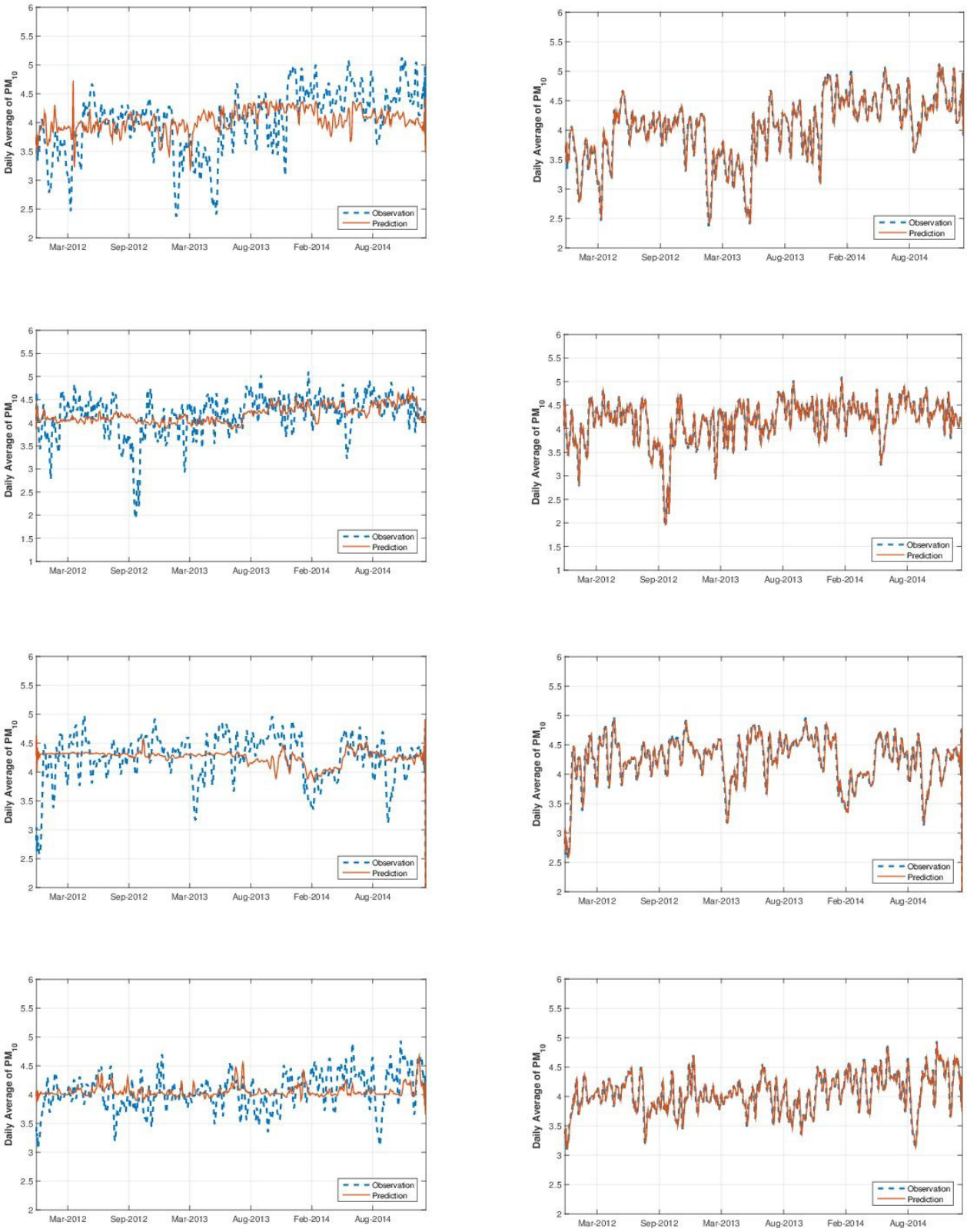


Fig. 3. The predicted and the observed values of daily average of $PM_{10}$ concentration obtained by the AR-GMDH model and the GMDH model for the north, west, east, and south monitoring stations in Tehran between January 2012 and December 2014.

RMSE, MAD, and IA values. The four columnsof the right hand side of Figure 3 also shows that there is a very good agreement between the predicted and the observed values of $PM_{10}$ concentrationobtained by the AR- GMDH model in all stations. That is, we see that thepredicted values estimated from the AR-GMDH algorithm have been able toidentify the observed values of $PM_{10}$ (t) on each day t in all four stations,while there is a big difference between the predicted values estimated fromthe GMDH algorithm and the observed values of $PM_{10}$ (t) on each day t in all four stations. Moreover, we plotted the predicted values versus the observedvalues for each data sample in the training and test sets in Fig. 4 to comparetwo algorithms performance. In this way, the better algorithm is the one with points closer to the ideal observed = predicted line. We regressed the predictedvalues on the observed values to estimate the regression line. Then, we compared the regression line to the ideal observed = predicted line. This process isperformed for all monitoring stations and the results are depicted in scatterplot (Fig. 4). It can be seen that the regression line between the observedand predicted values of the AR- GMDH algorithm is considerably closer to theideal observed = predicted line for both the training and test sets, comparedto the GMDH algorithm. For instance, the correlation between the observedand predicted values estimated by the GMDH algorithm for the test set inthe south station is 0.25, while this corresponding correlation obtained by the AR- GMDH algorithm is 0.98. Likewise, the correlation between the observedand predicted values estimated by the GMDH algorithm for the test set inthe west station is 0.34, while this corresponding correlation obtained by the AR- GMDH algorithm is 0.96. Note that the high concentration of $NO_2$ and CO is observed in the south and weststations respectively. The latter results can also be confirmed by observing the $R^2$ and IA values in Table 2.

## CONCLUSIONS

Predicting the daily level of the $PM_{10}$ concentration is the most important goalin urban air quality in some high populated city such as Tehran. Because it is a significant risk factor for a number of pollution related diseases. Modern and high quality monitoring systems in urban areas often allow predicting the daily $PM_{10}$ concentration level. During the recent years some monitoring stationsin Tehran have not been updated for some reasons and citizens have not been informed about the $PM_{10}$ concentration level for one or several days. Therefore in this situation, predicting $PM_{10}$ concentration level by a certain neural network approach seems to be necessary. For doing so, we first attempted to answer the question whether $PM_{10}$ concentration is associated with other air pollutant and meteorological factors in considered areas, or not. We found by a primary time series analysis that not only $PM_{10}$ is strongly associated with three pollutant factors ($SO_2$, $NO_2$, CO), and also it is associated with three meteorological variables such as temperature, precipitation, wind speed based on the data measured from four monitoring stations (north, east, west,and south areas). The positive correlation between $PM_{10}$ and other gaseouspollutants like $SO_2$ and $NO_2$ may in fact come from the contribution of the formation of fine sulfate and nitrate particles as part of the $PM_{10}$ concentration in the atmosphere. Our findings showed also that $PM_{10}$ is more likely tobe associated with $SO_2$ (pollutants related to road traffic emissions) than $NO_2$ and CO, which has also been observed in the time series analysis. This is maybe because four big passenger terminals are located at 13, 6, 5, and 16 municipal regions in which the east monitoring station is located at 13 and is nearto the 6 municipal regions. The west monitoring station is located at 5 municipal region but a bit far from the terminal location, and the north monitoringstation is located at the 20 municipal region which is near to the 5 municipal region These implied that the east area was more likely to have heavy traffic than the south and the west areas. We found also that $PM_{10}$ concentration for a specific day is also strongly associated with $PM_{10}$ concentration in the previous day. Further, apart from the air pollutant variables, the most significant meteorological variables in

predicting $PM_{10}$ (t) were the high temperatureval-ues, and the precipitation and wind speed values [7, 25]. On the ground of controversial problems such as the optimal input variables and the com-plexityof their nonlinear interaction pattern, we chose the GMDH neural networkalgorithm to predict $PM_{10}$ concentration for our study areas. The GMDH algorithm was applied in fact for prediction by training the network to output the next day value of $PM_{10}$ (t) concentration in terms

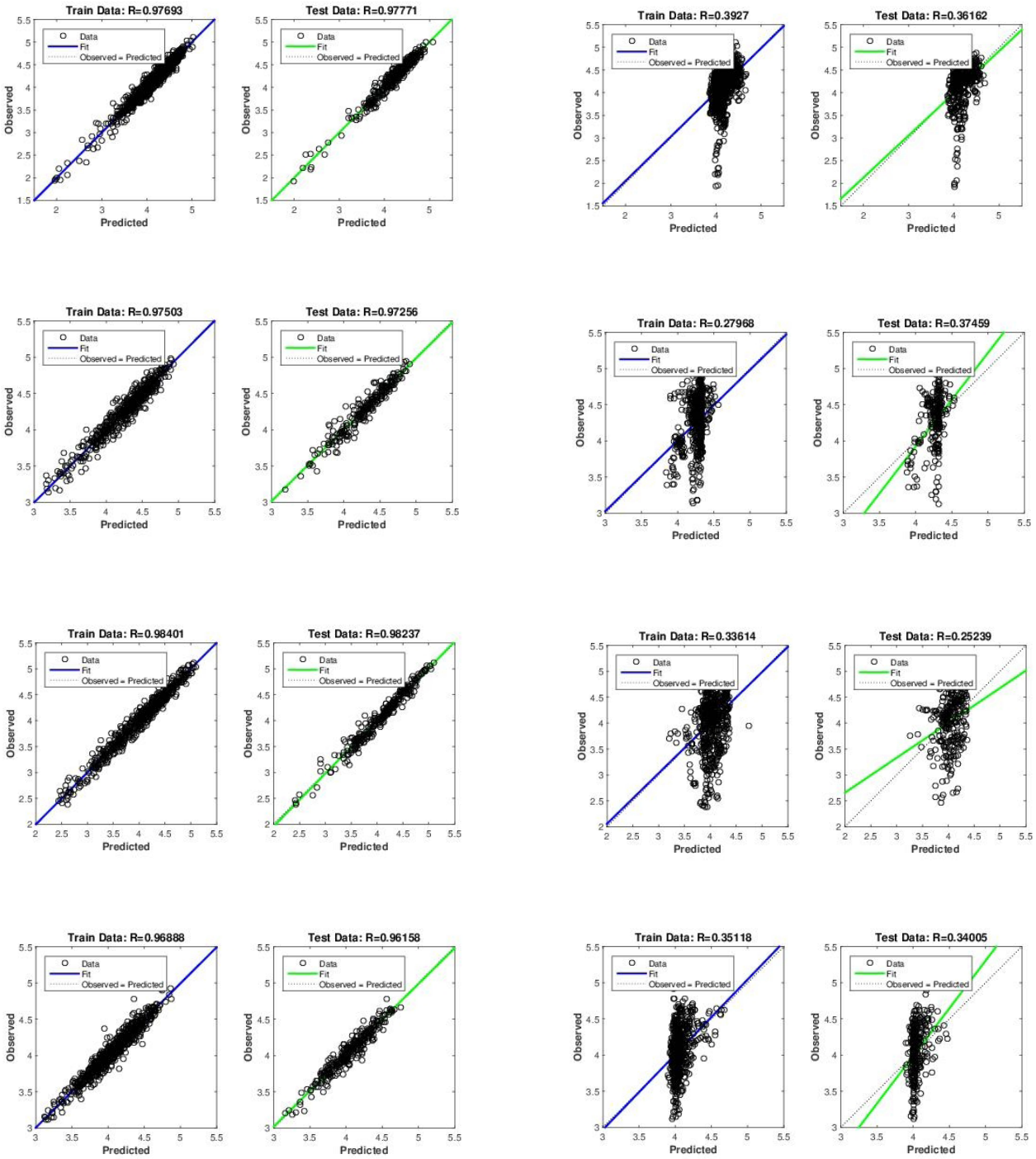


Fig. 4. Scatter plots of model performance on the train and the test data sets of the AR- GMDH model and the GMDH model based on the predicted and the observed correlation of the daily average of $PM_{10}$ concentration at the north, west, east, and south monitoring stations of Tehran between January 2012 and December 2014.

of the considered air pollutants and the meteorological variables on the present day values. Whilethe GMDH model ignores the prior information of the $PM_{10}$ concentrations ($PM_{10}$ (t − 1) on the present day), the AR- GMDH model showed that it canbe trained to approximate highly nonlinear function when $PM_{10}$ (t − 1) concentration on the present day is included in the algorithm. We also found thatadding time lags superior to present day, i.e. day t − 2, does not provide relevant additional information. Therefore, only the present day time lag for bothair pollutants and meteorological variables were taken into account in the AR- GMDH algorithm. Although we made time series data weakly stationary, ouranalysis gives evidence that the AR- GMDH algorithm is even better at picking up peak $SO_2$ (t − 1), $NO_2$ (t − 1), and CO (t − 1) compared to the GMDH algorithm. Moreover, the $NO_2$ (t − 1) and CO (t − 1) in some stations had aneffective role in predicting $PM_{10}$ (t). This may be due to road traffic influence, as road traffic behaves as a local source of $PM_{10}$ (t). The latter results were also observed in [13, 14]. The results of this study also reveal that more insightinto such predicted values can be obtained through studying the nonlinear relationship of the three listed meteorological variables. Because we observed that excluding precipitation and wind speed, the AR-GMDH algorithm wasnot able to best fit the prediction of $PM_{10}$ (t) concentration in all monitoringstations. Note also that, the coefficient of determination $R^2$ as absolute fraction of variance indicates the proportion of the variance in the output variablethat is predictable from the input variables. A disadvantage of the $R^2$ is thedifferences between observed and predicted are calculated as square values.Then, sometimes larger values in the time series data can be over estimated whereas smaller values are neglected. This insensitivity was overcome using the Willmott's IA [30].

As a result, our findings demonstrate that incorporating the prior information of the $PM_{10}$ concentration, ambient air pollutants,and the meteorological variables, the AR- GMDH neural network is the best fitted algorithm for predicting $PM_{10}$ concentration in time series data setting. Therefore, the AR- GMDH neural network can be proposed in predicting of $PM_{10}$ concentration due to its simplicity and in view of its cost efficiency, particularly for a country where air quality management is carried out on alimited budget. The limitation of this study was excluding the informationregarding atmospheric stability and circulation, which are important factors for the accumulation of $PM_{10}$ concentration [12, 14, 15, 25].

## FINANCIAL SUPPORTS

Golestan University financially supported the third author during his sabbatical leave 2015-2016 at the University of Waterloo, Canada**.**

## COMPETING INTERESTS

*The authors declare that there is no conflict of interest that would prejudice the impartiality of this scientific work.*

## AUTHOR CONTRIBUTIONS

It is certified that all of the authors have made the same contribution in the experiments and manuscript writing.

## ACKNOWLEDGEMENTS

The third author would like to thank Golestan University for the financial support during his sabbatical leave 2015- 2016 at the University of Waterloo, Canada. The authors also thank to Tehran air quality control Company and Iran meteorologicalorganization for availability of the air pollutants and meteorological data.

## ETHICAL CONSIDERATIONS

Authors are aware of, and have complied with, best practices in ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, and manipulation of figures, competing interests and compliance with policies on research ethics. Authors adhere to publication requirements that the submitted work is original

and has not been published elsewhere in any language.

## REFERENCES


[1] World Health Organization (WHO). Burden of disease from the joint effects of household and ambient air pollution for 2012. www. Who.int/phe/health topics/outdoor air/data bases.

[2] Zhang H, Wang Y, Hu J, Ying Q, Hu X-M. Relationships between meteorological parameters and criteria air pollutants in there megacities in China. Environmental Research. 2015; 140:242–54.

[3] Chelani AB, Devotta S. Nonlinear analysis and prediction of coarse particulate matter concentration in ambient air. Journal of the Air and Waste Management Association. 2006; 56:78- 84.

[4] de Kok TM, Driece HA, Hogervorst JG, Briedé JJ. Toxicological assessment of ambient and traffic-related particulate matter: a review of recent studies. Mutation Research/Reviews in Mutation Research. 2006;613(2):103-22.

[5] Elangasinghe MA, Singhal N, Dirks KN, Salmond JA, Samarasinghe S. Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modeling and k- means clustering. Atmospheric Environment. 2014; 94:106–16.

[6] Kelly FJ, Fussell JC. Size, source and chemical composition as determinants oftoxicity attributable to ambient particulate matter. Atmospheric environment. 2012. 60:504-26.

[7] Todorovic MN, Perisic MD, Kuzmanoski MM, Stojic AM, Sostarić AI, Mijic ZR, et al. Assessment of PM10 pollution level and required source emission reduction in Belgrade area. Journal of Environmental Science and Health, Part A. 2015;50(13):1351-9.

[8] Hassanzadeh S, Hosseinibalam F, Alizadeh R. Statistical models and time series forecasting of sulfur dioxide: a case study Tehran. Environmental Monitoring and Assessment. 2009; 155: 149–55.

[9] Saniei R, Zangiabadi A, Sharifikia M, Ghavidel Y. Air quality classificationand its temporal trend in Tehran, Iran, 2002–2012. Geospatial Health. 2016; 11:213–20.

[10] Taheri Shahraiyni H, Sodoudi S. Statistical modeling approaches for PM10 prediction in urban areas; a review of 21st-century studies. Atmosphere. 2016;7(2):15.

[11] Perez P, Reyes J. An integrated neural network model for PM10 forecasting. Atmospheric Environment. 2006; 40:2845–51.

[12] Dayan U, Levy I. The influence of meteorological conditions and atmosphericcirculation types on PM10 and visibility in Tel Aviv. Journal of Applied Meteorology. 2005; 44: 606–19.

[13] Russo A, Trigo RM, Martins H, Mendes MT. NO2, PM10 and O3 urban concentrations and its association with circulation weather types in Portugal. Atmospheric Environment. 2014; 89: 768–85.

[14] Russo A, Lind RG, Raischel F, Trigo R, Mendes M. Neural network forecast of dailypollution concentration using optimal meteorological data at synoptic and local scales. Atmospheric Pollution Research. 2015; 6:540–49.

[15] Choi W, Paulson SE, Casmassi J, Winer A. Evaluating meteorological comparability in air quality studies: Classification and regression trees for primary pollutants in California's South Coast Air Basin. Atmospheric Environment. 2013; 64:150–9.

[16] Eichler M. Graphical modelling of multivariate time series. Probability Theory and Related Fields. 2012; 153:233-68.

[17] Goyal P, Chan AT, Jaiswal N. Statistical models for the prediction of respirable suspended particulate matter in urban cities. Atmospheric Environment. 2006; 40: 2068-77.

[18] Hosseinpoor AR, Forouzanfar MH, Yunesian M, Asghari F, Naieni KH, Farhood D. Air pollution and hospitalization due to angina pectoris in Tehran, Iran: a time-series study. Environmental Research. 2005;99(1):126-31.

[19] Hu F, Lu Z, Wong H, Yuen TP. Analysis of air quality time series of Hong Kong with graphical modeling. Environmetrics. 2016;27(3):169-81.

[20] Liu PWG. Simulation of the daily average PM10 concentrations at Ta-Liao withBox-Jenkins time series models and multivariate analysis. Atmospheric Environment. 2009; 43: 2104-13.

[21] Stadlober E, Hormann S, Pfeiler B. Quality and performance of a PM10 daily forecasting model. Atmospheric Environment. 2008; 42:1098-109.

[22] Hrust L, Klaic ZB, Krizan J, Antonic O, Hercog P. Neural network forecastingof air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. Atmospheric Environment. 2009; 43: 5588–96.

[23] Voukantsis D, Karatzas K, Kukkonen J, Rsnen T, Karppinen A, Kolehmainen M. Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessalonikiand Helsinki. Science of the Total Environment. 2011; 409:1266–76.

[24] Grivas G, Chaloulakou A. Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. Atmospheric Environment. 2006; 40: 1216–29.

[25] Hooyberghs J, Mensink C, Dumont G, Fierens F, Brasseur O. A neural network forecast for daily average PM10 concentrations in Belgium. Atmospheric Environment. 2005; 39:3279–89.

[26] Sylvia Y, Santoso D. Transformation Box-Cox for stabilization of diversity in group random design. Journal of Computer Science. 2016; 11:18-29.

[27] Ian McLeod A, Gweon H. Optimal deseasonalization for monthly and daily geophysical time series. Journal

of Environmental statistics. 2012;4:1-11.
[28] Atashrouz S, Pazuki G, Alimoradi Y. Estimation of the viscosity of nine nanofluids using a hybrid GMDH- type neural network system. Fluid Phase Equilib. 2014;372: 43-48.
[29] Ebtehaj I, Bonakdari H, Zaji AH, Azimi H, Khoshbin F. GMDH-type neural network approach for modeling the discharge coefficient of rectangular sharp crested side weirs. Engineering Science and Technology, an International Journal. 2015; 18 (4):746-57.
[30] Acharya N, Shrivastava NA, Panigrahi B, Mohanty U. Development of an artificial neural network based multi-model ensemble to estimate the northeast monsoon rainfall over south peninsular India: an application of extreme learning machine. Climate Dynamics. 2014; 43( 5-6):1303–10.