Jundishapur J Health Sci. 2019 July; 11(3):e87978. doi: 10.5812/jjhs.87978.

Published online 2019 May 22. Research Article

Comparison of Statistical Methods, Neural Network, and Fuzzy Neural for Particular Matter Prediction: A Case Study in Mashhad, Iran

Elham Asrari [1,*] and Maryam Paydar [1]

[1] Department of Civil Engineering, Payame Noor University, Tehran, Iran

[*] *Corresponding author*: Department of Civil Engineering, Payame Noor University, P.O. Box. 19395-3697, Tehran, Iran. Email: e_asrari@pnu.ac.ir

Received 2018 December 19; **Revised** 2019 April 30; **Accepted** 2019 May 05.

Abstract

Background: In the recent era, air pollution is a major global concern that affects human health. The emission proportion of air pollutants is increased in many cities of Iran such as Mashhad. Particular matters (i.e. $PM_{2.5}$) are one of the five air pollutants known to be responsible for polluting the air in Mashhad. Nowadays, fuzzy neural intelligent systems, which are capable of solving nonlinear and complex problems, are widely used in the air pollution problem to determine the amount of the particles and dust in the air.
Methods: In the current study, the air quality data consisting of daily average concentrations of air pollutants and the meteorological data including the minimum temperature, precipitation, humidity, wind direction, and daily wind speed recorded by city monitoring stations from 2011 to 2017. The daily average pollutants concentration was used to study the relationship between $PM_{2.5}$ and the other air pollutants such SO_2, O_3, NO_2, PM_{10}, and CO. SPSS was used for data analysis. Linear regression, multilayer perceptron (MLP) neural network, and fuzzy neural network using MATLAB 2017 software were employed for modeling. Performance of the models was evaluated using root mean square error (RMSE) and coefficient of determination (R^2).
Results: Based on the obtained results, among the employed models, MLP neural network with $R^2 = 0.598$, RMSE = 0.088, and MSE = 0.0079 was better than linear regression, and the ANFIS model combining particle swarm optimization (PSO) algorithm with $R^2 = 0.804$, RMSE = 0.055, and MSE = 0.0031 had the best performance in the prediction of $PM_{2.5}$.
Conclusions: The ANFIS network correctly fitted more than 80% of total data; given that there were non-linear and complicated models in meteorological systems, this figure can indicate the high strength of ANFIS network through PSO-based combinational training methods for modeling nonlinear data.

Keywords: $PM_{2.5}$, Gas, Mashhad, Regression, Fuzzy and Neural Networks

1. Background

The industrialization activities, increasing the number of cities and their population density, and releasing different pollutants into the air make breathing the clean air an unattainable dream. A large portion of studies focus on air pollution in the big cities due to innumerable pollution sources caused by transportation, industrial, construction, and commercial activities. Air pollution problem can cause human mortality, and fauna and flora species extinction (1), along with numerous economic, social, cultural, and even political damages not only locally, but also nationally, regionally, and globally. According to the World Health Organization (WHO) research and the United Nations Environment Program (UNEP) in 50 countries, today, most of the world's population live in areas where the air pollutants exceed the permissible limits approved by WHO (2). In Iran, the emission of the air pollutants in many cities including Tehran, Mashhad, Ahvaz, Sanandaj, and Isfahan increase to dangerous levels. Studies show that particular matters (i.e. $PM_{2.5}$) have an association with various adverse human health effects such as premature mortality, exacerbation of asthma and other respiratory tract diseases, and cardiovascular diseases (3). Also, PM_{10} particles cause or intensify the cardiovascular or pulmonary diseases. It increases the risk of hospital referral, hospitalization, or even death in patients with congestive heart failure, asthma, or chronic pulmonary disease, especially the elderly (4).

The US Environmental Protection Agency (USEPA) classified air pollutants into two categories, primary (including carbon monoxide, nitrogen dioxide, sulfur dioxide, suspended particles, and lead) and secondary (including ozone peroxyacetyl nitrate). Pollutant particles known as PM include a mixture of solid and liquid droplets in the air. PM_1, $PM_{2.5}$, and PM_{10} are referred to suspended particles with an aerodynamic diameter of less than 1, 2.5, and 10 μm, respectively (5). All of them can remain in the atmosphere for several months. Particles larger than 2 μm have great importance; their sedimentation rate is low. Among different air pollutants, suspended particles are known as the most important air pollutants in the big cities, since for each 10 $\mu g/m^3$ increase in PM_{10}, daily mortality increases by 0.6% (6).

Regarding the adverse effects of air pollution on human health and environment, it is necessary to make correct decisions and plan to overcome this dilemma. In this regard, predicting significant concentration of pollutants and their correlation with meteorological parameters affect the decision making to deal with air pollution (7). This issue has more importance in metropolises. Mashhad is ranked among seven polluted cities of Iran and, after Tehran, it is the second most polluted city. There are many methods to predict the concentration of air pollutants. In recent years, significant advances are made to develop neural network models for air pollution prediction, which is more efficient than other methods. Neural and fuzzy networks are useful tools to identify and model the data (8) and predict unknown issues that there is no knowledge about their inputs and outputs (9-12). In the current study, three methods of multiple linear regression, artificial neural network multilayer perceptron (MLP), and fuzzy neural network were addressed based on the PSO (particle swarm optimization) algorithm to estimate $PM_{2.5}$ suspended particle concentration in terms of air pollution data and meteorological parameters; after the evaluation of the prediction accuracy, each model was estimated.

2. Objectives

The present study aimed at comparing statistical methods, neural network, and fuzzy neural network in suspended particles prediction ($PM_{2.5}$).

3. Methods

3.1. Study Area

Mashhad -the capital of Khorasan Razavi Province- with approximately 351 km^2 area is the second most polluted city in Iran after Tehran. Its geographic location is the Northeast of Iran in 36° 16' North latitude and 59° 37' East longitude (Figure 1).

3.2. Study Data

The data used in the current study included the minimum and daily average temperature (°C), minimum and maximum daily humidity (%), average wind speed (km/h), average daily precipitation (mm), wind angle (degree), average sunshine hours per day (h), average daily concentration of PM_{10} ($\mu g/m^3$), SO_2 (ppb), CO (ppm), NO_2 (ppb), and O_3 (ppb).

As shown in Figure 1, the hourly air pollutants concentration data were collected from three stations from 2011 to 2016. The metrological data were taken from Mashhad synoptic station from 2011 to 2017. After data attainment and validation and outliers' elimination, the daily average was calculated for each pollutant. SPSS version 16 was used to simulate multiple regression, and in order to simulate neural network models and Anfis, the MATLAB 2017 software was used. Network input data were divided into three categories to improve network prediction strength:

- Train data that contained 70% of data (from 2011 to 2014) were related to network training; the network weight was determined by them. This section data included 1555 items.
- Validation data that contained 15% of data (the year 2015) were in charge of network training monitoring; the decision to stop calculations was made through error consideration during training. This section data included 430 items.
- Test data that contained 15% of data (the year 2016) related to validation and network capabilities examination. This section data included 121 items.

3.3. Multiple Regression

Multiple regressions are a method for collective and individual participation of two or more independent variables in a dependent variable changes. In this method, the variables are entered one by one. Some conditions should be met before using the regression model; first, linear relationship between independent and dependent variables; second, normal distribution of error values, and third, independence of error values.

In the current study, after accuracy verification of the above conditions, variables that had a significant correlation with $PM_{2.5}$ (R^2 = 0.95) were extracted as a model through multiple linear regression by a step-by-step approach and the coefficients of each were obtained.

3.4. MLP Neural Network

Multi-layer feedforward networks are the most important structures of artificial neural networks. Generally, these networks include a set of sensory units (basic neurons) that form input layer, one or more hidden layers, and

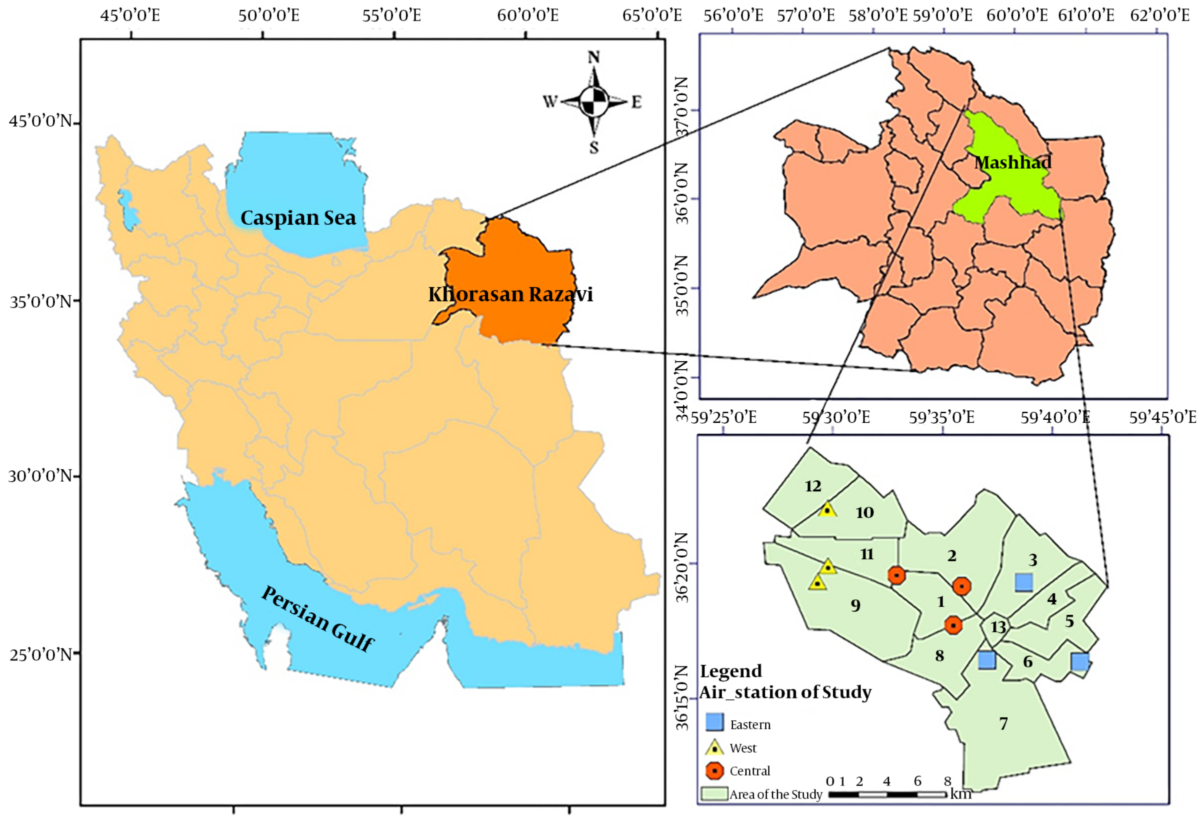

Figure 1. Study area and air quality measurement stations

an output layer. The input signal is spread layer-by-layer through the network in a forward direction. This kind of network is commonly referred to as MLP. The number of hidden layers should be as low as possible. Initially, the network is trained by a hidden layer; in case of inappropriate function, the number of hidden layers will increase (13). If possible, less hidden neurons are examined (14). The current study used an input and an output layer, a hidden layer consisting of 85 neurons, tansig conversion function, and trainbr training algorithm.

3.5. The PSO-Based ANFIS Network

In the current study, an adaptive type II neuro-fuzzy network was designed according to Sugeno algorithm using MATLAB 2017 fuzzy toolbox and genfis3 tool (using the FCM model for clustering). Network training was done by PSO algorithm to increase its efficiency. In order to optimize the particle swarm and find the most efficient state, the Kennedy and Eberhart algorithm was used due to its high efficiency. In the current study, 100 rounds for each network and 10 clusters were considered. These numbers were obtained by trial and error. In order to optimize membership function parameters in adaptive neuro-fuzzy inference system, ANFIS network and PSO algorithm parameters were presented; data are shown in Table 1.

3.6. Network Performance Assessment

Network performance was validated. Following equations were used to evaluate network performance.

$$Mean\ squared\ error\ \ (MSE) = \frac{\sum_1^n (obs - calc)^2}{N} \tag{1}$$

$$Root - mean - squared\,error\ \ (RMSE) \\ = \sqrt{\sum\nolimits_{i=1}^{n} \frac{(calc - obs)^2}{N}} \tag{2}$$

$$Coefficient\,of\,determination\ \ (R^2) \\ = \frac{\sum_1^n (calc - avg.obs)^2}{\sum_1^n (obs - avg.obs)^2} \tag{3}$$

The data were normalized by the below algorithm:
$X_{norm} = (X - X_{min})/(X_{max} - X_{min})$

Table 1. ANFIS Network Parameters

Parameter	Value
ANFIS network parameters	
Minimum improvement	1e-5
Number of clusters	10
Maximum iteration	100
Partition matrix exponent	2
Particle swarm optimization algorithms	
Iteration numbers	1000
Population	500
Inertia weight damping ratio	0.99
(C_1)	1
(C_2)	2
Inertia weight	1

4. Results

Concentrations of the daily average pollutants showed that $PM_{2.5}$ was the main pollutant in Mashhad. In about 40% of days, its concentration was upper than EPA (Environmental Protection Agency) limited standards. Annual average of five major air pollutants in different years is provided in Table 2.

4.1. Multiple Regression Results

Tables 3 and 4 show the multiple regression results through step-by-step method. In these tables, ADJ.R^2 indicates the $PM_{2.5}$ variance percentage that model predicts. In the best case, which four variables of PM_{10}, CO, NO_2, and wind angles are used to predict $PM_{2.5}$ suspended particles, 32% $PM_{2.5}$ variance is predicted. According to P values, these four variables can significantly predict $PM_{2.5}$ concentration. Beta values indicate that when a unit of PM_{10} increases, $PM_{2.5}$ level increases 0.45; also, when a unit of NO_2 increases $PM_{2.5}$ level increases 0.12, and if a unit of CO increases $PM_{2.5}$ level increases by 0.15. Ultimately, a one-degree increase in the wind angle reduces the $PM_{2.5}$ concentration by 0.06.

The regression model is:

$PM_{2.5} = 9.695 + 0.215\ PM_{10} + 5.250\ CO + 0.152\ NO_2 - 0.013$ windirect

This model shows positive relationships between PM_{10}, CO, NO_2, and negative impact of wind speed on $PM_{2.5}$ concentration prediction.

4.2. Estimation and Prediction Results of $PM_{2.5}$ Suspended Particles Using the MLP Neural Network

The MLP neural network has an input and output layer, a hidden layer consisting of 85 neurons, tansig conversion function and trainbr training algorithm, and presents the best model results. The estimated results through MLP neural network have less errors and better R^2 compared to multiple linear regression method. However, it does not provide a good prediction of $PM_{2.5}$ concentrations, especially in cases of maximum and minimum.

4.3. Estimation and Prediction Results of $PM_{2.5}$ Suspended Particles Using ANFIS Network

In this simulation, an adaptive type II neuro-fuzzy network designed according to Sugeno algorithm. Network training is done using the PSO optimization algorithm. After a few tries by trial, the best number of ANFIS network repetitions was 100 and the number of clusters was determined 10. In order to improve the network prediction ability, the data from 2011 to 2014 for ANFIS neural network training were separated from the beginning of the data, and then, the data for 2016 were used for validation, and afterwards, the network was tested with the data of 2017. Since the networks starting point is chosen based on random numbers, the ANFIS network was run more than 10 times to obtain the best output. The numbers, tables, and charts were the best state among these repetitions.

According to the obtained results, more than 80% of total data were correctly fitted in the ANFIS network; given that there were non-linear and complicated models in meteorological systems, this figure can indicate the high strength of ANFIS network through PSO-based combinational training methods for modeling nonlinear data. Table 5 summarizes the modeling results with three methods.

5. Discussion

Since Mashhad is the second largest metropolis in Iran, and according to high volume of tourists, the pollutants concentration prediction has a great importance to make necessary decisions to implement traffic constraints in the city in order to prevent and reduce the harmful effects. Therefore, the current study aimed at predicting the concentration of suspended particles through artificial neural networks in accordance with climatic conditions and other pollutants. The correlation analysis of suspended particles with other pollutants by regression indicated that $PM_{2.5}$ had a direct correlation with PM_{10}, NO_2, and CO; this result was consistent with those of Ehsanzadeh et al., study about factors affecting suspended particles in Tehran. The simulation results with the MLP neural network showed that in the best case, the training was done through Bayesian regularization (trainbr) and the correlation coefficient was R = 0.77 and the root mean square er-

Table 2. Average Daily Concentration of Pollutants from 2011 to 2016

Year	2011	2012	2014	2015	2016
PM_{10}, $\mu g/m^3$	84.38	78.46	81.98	75.57	83.9
CO, ppm	1.99	1.98	1.87	2.27	1.96
SO_2, ppm	0.01	0.01	0.01	0.02	0.01
NO_2, ppb	34.62	29.78	30.29	22.75	21.95
O_3, ppb	16.7	17.19	8.19	11.39	20.92
$PM_{2.5}$, $\mu g/m^3$	43.29	34.52	29.11	23.38	30.22

Table 3. The Results of Multiple Regression[a]

Predictive Variable	T	Beta	SE	B	P Value
Constant	6.214		1.560	9.695	0.000
PM_{10}	18.095	0.456	0.012	0.215	0.000
CO	5.421	0.150	0.874	5.250	0.000
NO_2	4.531	0.123	0.033	0.152	0.000
Wind direction	-2.461	-0.061	0.005	-0.013	0.0014

[a] $R = 0.568$, $R^2 = 0.323$, $ADJ.R^2 = 0.316$

Table 4. Result of Step-wise Coefficients[a]

Model	Unstandardized Coefficient (B)	Standardized Coefficient (Std. Error)	Beta	T	P Value
(Constant)	16.596	1.113		14.905	0.000
PM_{10}	0.245	0.012	0.518	20.647	0.000
(Constant)	10.011	1.397		7.165	0.000
PM_{10}	0.223	0.012	0.473	18.695	0.000
CO	5.385	0.717	0.190	7.511	0.000
(Constant)	8.208	1.441		5.694	0.000
PM_{10}	0.216	0.012	0.458	18.149	0.000
CO	3.963	0.777	0.140	5.100	0.000
NO_2	0.152	0.034	0.123	4.536	0.000
(Constant)	9.695	1.560		6.214	0.000
PM_{10}	0.215	0.012	0.456	18.095	0.000
CO	4.250	0.784	0.150	5.421	0.000
NO_2	0.152	0.033	0.123	4.531	0.000
Wind direct	-0.013	0.005	-0.061	-2.461	0.014

[a] Dependent variable: $PM_{2.5}$

ror (RMSE) was 0.088, which was better than that of step-by-step regression method with R = 0.56 and RMSE = 0.106. Taghavi in a part of her master's thesis modeled CO pollutant in Mashhad using MLP artificial neural network and compared it with linear regression method (15). The execution results of these two models showed that the artificial neural network model had more ability than linear regression to predict the CO concentration. The correlation coefficient (R) and the root mean square error (RMSE) in neural network based on the regression model were 0.61 and 0.069, and 0.61 and 0.1, respectively. Koorani developed models using the neural network and linear regression to predict the daily average concentrations of ozone and PM_{10} in Milan, Italy. The results of this study also indicated the superiority of the neural network to linear regression. Based on the neural network results, the corre-

Table 5. Results of Modeling with Regression, MLP, and ANFIS

Model	Data Type	R^2	RMSE	MSE
Regression	Stepwise	0.2335	0.10653	0.01135
MLP	Train data	0.6649	0.08170	0.0066
	Test data	0.4286	0.1072	0.0114
	Validation data	0.4764	0.10075	0.0101
	Total	0.5986	0.0888	0.0079
ANFIS	Train data	0.8012	0.05583	0.003109
	Test data	0.8307	0.05649	0.003192
	Validation data	0.706	0.06287	0.003953
	Total	0.804	0.05598	0.003134

lation coefficient between the results of the model, and the actual data for ozone and suspended particles was 0.85 and 0.9, respectively (16). To predict the concentration of suspended particles of less than 2.5 μm using an adaptive PSO-based neuro-fuzzy algorithm, given RMSE and R statistics, the current study results showed that the PSO-based neuro-fuzzy network had a relatively high accuracy and efficiency. The best performance was related to the neural fuzzy network with R = 0.914 and RMSE = 0.05598 at the validation stage. Also, the designed network could predict data for the year 2017 with over 70% accuracy. The model presented by Yildirim and Bayramoglu in Zonguldak City using an adaptive neuro-fuzzy network to estimate the meteorological effect on sulfur dioxide and suspended particles showed the high efficiency of ANFIS model in air pollutants issue, which was consistent with the results of the current study (17). Aliari Shorehdel et al., suggested short-term air pollution prediction using multilayer perceptron neural networks delayed memory line, gamma, and ANFIS though PSO-based combinational training, and showed that the combined proposed method based on PSO and Kalman filter had a good capability to improve prediction performance for ANFIS network (18). In the current study, the PSO optimization results were more accurate. Saadabadi et al., suggested a model to predict $PM_{2.5}$ concentration in the air of Mashhad using one-year data from a hybrid model including wavelet transform and MLP neural network, the values of R^2 =0.778 and RMSE = 0.7059 represented that the proposed model in this study had a high accuracy in the suspended particles prediction, maybe due to the longer period of time. The ANFIS network correctly fitted more than 80% of total data; given that there were non-linear and complicated models in meteorological systems, this figure can indicate the high strength of ANFIS network through PSO-based combinational training methods for modeling nonlinear data.

Acknowledgments

The authors are extremely grateful to the Department of Environment in Mashhad for providing information to conduct and complete the study.

Footnotes

Authors' Contribution: Maryam Paydar: the research conduction; Elham Asrari: the study supervision.

Conflict of Interests: The authors declared no conflict of interest.

Ethical Approval: It was not declared by the author.

Funding/Support: No grants were received from any funding agencies in the public, commercial, or profit sectors.

References

1. Miri M, Derakhshan Z, Allahabadi A, Ahmadi E, Oliveri Conti G, Ferrante M, et al. Mortality and morbidity due to exposure to outdoor air pollution in Mashhad metropolis, Iran. The AirQ model approach. *Environ Res*. 2016;**151**:451–7. doi: 10.1016/j.envres.2016.07.039. [PubMed: 27565880].
2. Colls J. *Air pollution*. London: Spon Press; 2002. doi: 10.4324/9780203476024.
3. Dockery DW. Epidemiologic evidence of cardiovascular effects of particulate air pollution. *Environ Health Perspect*. 2001;**109 Suppl 4**:483–6. doi: 10.1289/ehp.01109s4483. [PubMed: 11544151]. [PubMed Central: PMC1240569].
4. Chen YS, Sheen PC, Chen ER, Liu YK, Wu TN, Yang CY. Effects of Asian dust storm events on daily mortality in Taipei, Taiwan. *Environ Res*. 2004;**95**(2):151–5. doi: 10.1016/j.envres.2003.08.008. [PubMed: 15147920].
5. EPAU. *Ecological effects test guidelines*. Gammarid Acute Toxicity Test OPPTS; 2007.
6. Brunekreef B, Holgate ST. Air pollution and health. *The Lancet*. 2002;**360**(9341):1233–42. doi: 10.1016/s0140-6736(02)11274-8.
7. Lira TS, Barrozo MA, Assis AJ. Air quality prediction in Uberlândia, Brazil, using linear models and neural networks. *Comput Aided Chem Eng*. 2007;**24**:51–6. doi: 10.1016/s1570-7946(07)80032-0.

8. Razi M, Athappilly K. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert System Application*. 2005;**29**(1):65–74. doi: 10.1016/j.eswa.2005.01.006.
9. Murat YS, Ceylan H. Use of artificial neural networks for transport energy demand modeling. *Energ Policy*. 2006;**34**(17):3165–72. doi: 10.1016/j.enpol.2005.02.010.
10. Hájek P, Olej V. Ozone prediction on the basis of neural networks, support vector regression and methods with uncertainty. *Ecol Inform*. 2012;**12**:31–42. doi: 10.1016/j.ecoinf.2012.09.001.
11. Mao X, Shen T, Feng X. Prediction of hourly ground-level PM 2.5 concentrations 3 days in advance using neural networks with satellite data in eastern China. *Atmos Pollut Res*. 2017;**8**(6):1005–15. doi: 10.1016/j.apr.2017.04.002.
12. Olvera-García MÁ, Carbajal-Hernández JJ, Sánchez-Fernández LP, Hernández-Bautista I. Air quality assessment using a weighted Fuzzy Inference System. *Ecol Inform*. 2016;**33**:57–74. doi: 10.1016/j.ecoinf.2016.04.005.
13. Salehi H, Zeinali-Heris S, Esfandyari M, Koolivand M. Nero-fuzzy modeling of the convection heat transfer coefficient for the nanofluid. *Heat Mass Transfer*. 2012;**49**(4):575–83. doi: 10.1007/s00231-012-1104-9.
14. Menhaj M. *Basics of neural networks [dissertation]*. Tehran: Amir Kabir University of Technology; 2005.
15. Taghavi H. *Temporal distribution and spatial distribution of air pollution index contamination and effective factors on ander Mashhad [dissertation]*. Iran: Ferdowsi University of Mashhad; 2012.
16. Comrie AC, Diem JE. Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona. *Atmos Env*. 1999;**33**(30):5023–36. doi: 10.1016/s1352-2310(99)00314-3.
17. Yildirim Y, Bayramoglu M. Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak. *Chemosphere*. 2006;**63**(9):1575–82. doi: 10.1016/j.chemosphere.2005.08.070. [PubMed: 16310825].
18. Aliari Shorehdel M, Teshnehlab M, Khaki Sedigh A. Short term prediction of air pollution using MLP, GAMMA, ANFIS, and mixed training methods based on PSO. *J Control*. 2008;**2**(1):1–19.