Pollution, 6(1): 43-57, Winter 2020
DOI: 10.22059/poll.2019.279412.618
Print ISSN: 2383-451X Online ISSN: 2383-4501
Web Page: https://jpoll.ut.ac.ir, Email: jpoll@ut.ac.ir

Carbon Monoxide Prediction in the Atmosphere of Tehran Using Developed Support Vector Machine

Akbarzadeh, A.[1,2], Vesali Naseh, M. R.[3*] and NodeFarahani, M.[4]

1. Water Research Institute, Ministry of Energy, P.O. Box 16765-313, Tehran, Iran
2. The Institute for Energy and Hydro Technology, P.O. Box 14845-131Tehran, Iran
3. Department of Civil Engineering, Arak University, P.O. Box 38156-879, Arak, Iran
4. Department of Civil Engineering, Azad University South Tehran Branch, P.O. Box 15847-43311, Tehran, Iran

Received: 25.04.2019 Accepted: 17.10.2019

ABSTRACT: Air quality prediction is highly important in view of the health impacts caused by exposure to air pollutants in urban air. This work has presented a model based on support vector machine (SVM) technique to predict daily average carbon monoxide (CO) concentrations in the atmosphere of Tehran. Two types of SVM regression models, i.e. ε-SVM and υ-SVM techniques, were used to predict average daily CO concentration as a function of 12 input variables. Then, forward selection (FS) technique was applied to reduce the number of input variables. After converting 12 input variables to 7 using the FS, they were fed to SVM models (FS-(ε-SVM) and FS-(υ-SVM)). Finally, a comparison among SVM models operation and previously developed techniques, i.e. classical regression model and artificial intelligent methods such as ANN and adaptive neuro-fuzzy inference system (ANFIS) was carried out. Determination of coefficient (R^2) and mean absolute error (MAE) for ε-SVM (υ-SVM) were 0.87 (0.40) and 0.87 (0.41), respectively, while they were 0.90 (0.39) and 0.91 (0.35) for ANN and ANFIS, respectively. Results of developed SVM models indicated that both FS-(ε-SVM) and FS-(υ-SVM) regression techniques were superior. Furthermore, it was founded that the performance of FS-(ε-SVM) and FS-(υ-SVM) models were generally a bit better than the best FS-ANFIS and FS-ANN solutions for short term forecasting of CO concentrations.

Keywords: Air pollution, forward selection, carbon monoxide, artificial intelligent, Tehran.

INTRODUCTION

Predicting the next day's air pollution levels is the first and the most important step in urban air quality management to provide proper controlling strategies. Nowadays, air pollution forecasting has become a significant issue, especially in large cities like Tehran, Iran. Most of the air pollution problems in Tehran are related to carbon monoxide (CO) and particulate matter less than 10μm (PM_{10}). Among these pollutants, CO becomes more concerning given that it occupies more than 75% of air pollutants weight in the atmosphere of Tehran (Bayat, 2005). Therefore, predicting daily levels of this pollutant can play a significant role in

* Corresponding Author Email: m-vesalinaseh@araku.ac.ir

reducing possible risks. In past decades, different approaches such as classical regression methods (Noori et al., 2008; Sahin et al., 2004; Shakerkhatibi et al., 2015;), deterministic models (Jorquera, 2002; Gokhale *et al.*, 2007; Murena *et al.*, 2009; Elangasinghe et al., 2014) and artificial intelligent (AI) techniques (Pérez *et al.*, 2000; Perez-Roa *et al.*, 2006; Hrust et al., 2009; Wang et al., 2015) demonstrated that they can be used satisfactorily for air pollution forecasting.

Statistical approaches such as classical regression and AI methods have some advantages over deterministic techniques. They do not need data about emissions and their structure is often more familiar than deterministic models (Nunnari et al., 2004). On the other hand, most deterministic operative models for estimating the dispersion of gases and particles in the atmospheric boundary layer are based on the Gaussian approach. Such models concentrate on the hypothesis that states pollutants are dispersed in homogenous turbulence. Yet, considering the presence of the ground, turbulence is not generally homogenous along the vertical path (Pelliccioni et al., 2006).

Among the statistical approaches, classical regression methods are linear, model driven, and parametric in nature, assuming strong prior knowledge about the unknown dependency. However, in many real world problems, this underlying assumption is not always true. Further, such approaches are impractical in dealing with high-dimensional cases (Liong et al., 2002; Noori et al., 2008). Recently, AI models such as artificial neural network (ANN) and adaptive neuro-fuzzy inference systems (ANFIS) are being widely used to address the shortcomings of the parametric approach (Jalili Ghazi Zade & Noori, 2008; Noori et al., 2010b; Dehghani *et al.*, 2014). Yet, these methods lack underlying mathematical theory and are usually motivated by biological arguments.

A novel tool from the AI field known as support vector machine (SVM) has gained popularity in the machine learning community (Cristianini et al., 2000). It has a functional form (similar to classical regression models) and its complexity is determined by the available data to be "learned" (similar to AI approaches). It is found that the empirical performance of SVM is generally consistent the best ANN solutions (Hearst et al., 1998). It has been hypothesized that this is because there are fewer model parameters to be optimized in the SVM approach, reducing the possibility of over-fitting the training data and thus increasing the actual performance (Brown et al., 1999). Compared with traditional ANN, learning in SVM is very robust regarding the computations precision (Anguita et al., 1999). With the introduction of ε-insensitive loss function, the applications of SVM in non-linear regression estimation and time series prediction has been extended (Bray et al., 2004). Also, it has been applied successfully to environmental problems (Noori et al., 2009a; Noori et al., 2009b; Luna *et al.*, 2014). Lu and Wang (2005) examined the feasibility of applying SVM and classical radial basis function (RBF) network to predict air pollutant levels in advancing time series based on the monitored air pollutant database in Hong Kong downtown area. They reported that SVM outperforms the conventional RBF network in predicting air quality parameters with different time series and it has better generalization performance than the RBF model. Osowski & Garanty (2007) presented a method of daily air pollution forecasting by using SVM and wavelet decomposition based on the observed data of nitrogen dioxide (NO_2), CO, sulfur dioxide (SO_2) and dust, for the past years by considering actual meteorological parameters like wind, temperature, humidity and pressure. Salazar-Ruiz et al. (2008) offered a SVM model for prediction of maximum tropospheric ozone concentrations for the next day in the

Mexicalie-Calexico border area in US. Luna et al. (2014) investigated on the behavior of air pollution and meteorological variables (NO_2, nitrogen monoxide (NO), nitrogen oxides (NO_x), CO, ozone (O_3), scalar wind speed, global solar radiation, temperature, and moisture content in the air), using the method of principal component analysis (PCA) for exploratory data analysis. They proposed forecasts of O_3 levels applying nonlinear regression methods like ANN and SVM, from primary factors. The study concluded that the models' predictions were reliable, and PCA-ANN-SVM confirmed their robustness as a promissing method for modeling and analying of O_3 concentrations in the tropospheric levels. Moazami et al. (2016) developed an appropriate methodology for determination of uncertainty in support vector regression (SVR) as a well-known modeling approach in atmospheric science based on running SVR model many times to predict the next day CO concentrations in Tehran. Thereafter, they compared their results with ANFIS and ANN and showed that the SVR had less uncertainty in CO prediction than the ANN and ANFIS models. In another study, Azeez et al. (2018) developed a hybrid model based on the integration of three models, correlation-based feature selection (CFS), SVR and GIS, to predict vehicular emissions on roads in an urban areas of Kuala Lumpur, Malaysia. The suggested model was developed with seven road traffic CO predictors selected via CFS (sum of vehicles, sum of heavy vehicles, heavy vehicle ratio, sum of motorbikes, temperature, wind speed, and elevation). Their suggested model resulted high validation accuracy and correlation coefficient and introduced as a useful tool for traffic CO assessment on roads. Ghaemi et al. (2018) used a LaSVM-based online algorithm for air pollution prediction in Tehran. Pollutant concentration and meteorological data were fed to the developed online forecasting system. The authers evaluated performance of the system by comparing the prediction results of the air quality index (AQI) with those of a traditional SVM algorithm. Their results showed significant increase of speed by the online algorithm while preserving the accuracy of the SVM classifier.

Similar to statistical and mathematical models, SVM has also some disadvantages. The large number of input variables is one of the main problems for SVM planning. SVM models are not designed to eliminate superfluous inputs, thus solving the quadratic programming during the training procedure becomes difficult with standard solvers (Kecman, 2005). Moreover, high number of input variables may prevent SVM to find the optimized models. Therefore, if possible, it is recommended to reduce input variables even though this omits some parts of information.

In the present paper, two regression models, i.e. ε-SVM and υ-SVM are presented for average daily CO concentration prediction using forward selection (FS) method for input selection. Then the results of the models are compared and superior models for predicting the average daily CO concentration are reported. Finally, a comparison between superior models and some previously developed models, i.e. multiple linear regression (MLR) technique (Noori et al., 2008), ANN and ANFIS models (Noori et al., 2010a), is carried out.

MATERIAL AND METHODS

Tehran is the capital and the largest city of Iran which is located between 35^o 34' to 35^o 50' N and 51^o 02' to 51^o 36' E with the area about 570 km^2. It is surrounded by mountains to the north, west and east. Its current population includes about 8,000,000 (Bayat, 2005). There are 21 air quality measurement stations in Tehran (Fig. 1). The results of previous studies about air pollution of Tehran demonstrate that 90% of total air pollutants are generated from traffic and only 10% are

from other sources (Bayat, 2005). In comparison with other air pollutants in the atmosphere of Tehran, CO is more significant because it forms more than 75% of air pollutants (Bayat, 2005). For this study, the pollution and meteorological data during 2004-2005 are obtained from Gholhak station in the north of Tehran at 35^o 41' N and 51^o 19' E with the altitude of 1190.8 m above sea level (Fig. 1). To predict CO concentration in the future, (i.e. next 24 hours), the daily arithmetic averages of six air pollutants: PM_{10}, total hydrocarbons (THC), NO_x, methane (CH_4), SO_2 and O_3 and also six meteorological variables: pressure (Press), temperature (Temp), wind direction (WD), wind speed (WS) and relative humidity (Hum) are used. Statistical analysis of used data is listed in Table 1.

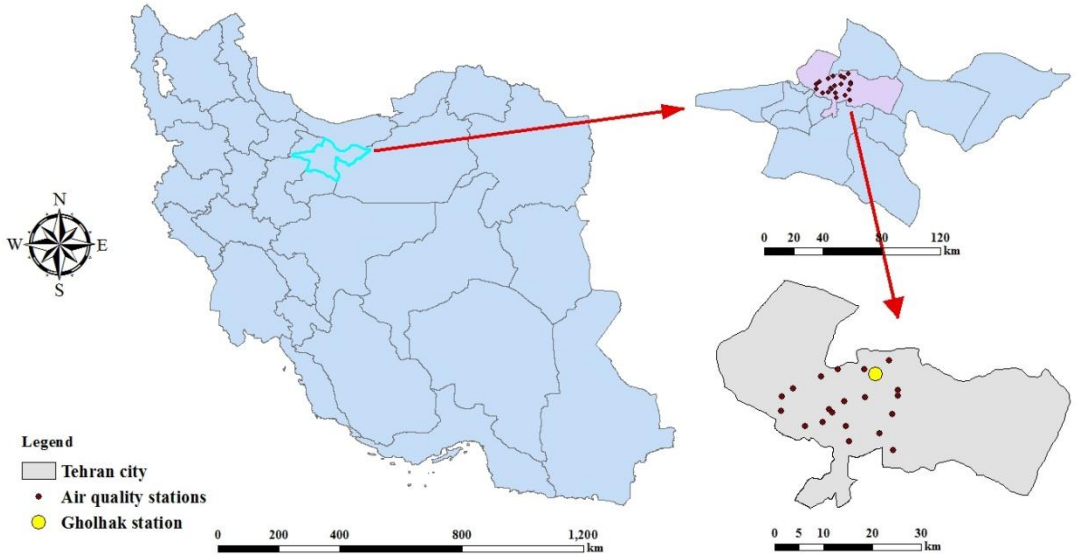

Fig. 1. Air quality stations in Tehran and the location of Gholhak station

Table ١. Statistical analysis of used data

Parameter Unit	PM_{10} µg/m^3	THC ppm	NO_x ppm	CH_4 ppm	SO_2 ppm	O_3 ppm	Press mBar	Temp oC	WD Deg	WS m/s	Hum %RH	Solar KW/M^2	CO ppm
Average	9.81	4.03	0.18	1.67	0.03	0.01	845.97	18.93	182.28	0.84	49.17	0.25	4.93
Max	190.27	8.73	0.64	4.73	0.07	0.04	903.30	38.29	282.73	20.71	94.29	0.59	9.67
Min	3.05	0.00	0.00	0.00	0.00	0.00	597.59	0.59	32.27	0.00	10.14	0.01	1.30
Median	8.92	4.34	0.12	1.82	0.03	0.01	850.71	20.87	181.61	0.80	47.64	0.27	4.81
St. Deviation	9.08	1.78	0.16	0.97	0.01	0.01	19.80	9.07	28.59	1.18	18.55	0.12	1.53

When the number of candidate covariates (*N*) is small, one can choose a prediction model by computing a reasonable criterion (e.g., root mean square error (RMSE), squared errors of prediction (SSE) or cross-validation error) for all possible subsets of the predictors. However, as *N* increases, the computational burden of this approach increases very quickly. This is one of the main reasons why step-by-step algorithms like FS are popular. FS has been successfully used by many researchers in order to build robust prediction models (Chen et al., 2004; Eksioglu et al., 2005; Wang *et al.*, 2006; Khan et al., 2007). In this approach, which is based on linear regression model, the first step is ordering the explanatory variables according to their correlation with the dependent variable (from the most to the least correlated variable). Then, the

explanatory variable which is highly correlated with the dependent variable is selected as the first input. All remained variables are then added one by one as the second input according to their correlation with the output and the variable which most significantly increases the determination coefficient (R^2) is selected as the second input. This step is repeated *N-1* times to evaluate the effect of each variable on model output. Finally, among *N* obtained subsets, the subset with optimum R^2 is selected as the model input subset. The optimum R^2 is integral to a set of variables after which adding new variable does not significantly increase the R^2 (Chen et al., 2004).

There are many papers and books, which provide a detailed description of the SVM theory (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000; Yu et al., 2006; Chen et al., 2007; Noori et al., 2009a, 2011, and 2015a), and hence only a brief description of SVM is given here. In a regression SVM, you have to estimate the functional dependence of the dependent variable y on a set of independent variables x. It assumes that, like other regression problems, the relationship between the independent and dependent variables is given by a deterministic function f plus the addition of some additive noise ($y = f(x) + noise$). The task is then to find a functional form for f that can correctly predict new cases that the SVM has not been presented before. This can be achieved by training the SVM model on a sample set, i.e., training set. This process involves the sequential optimization of an error function (Noori et al., 2015b). Depending on the definition of this error function, two types of SVM models can be recognized: regression SVM type 1 (also known as ε-SVM regression) and regression SVM type 2 (also known as υ-SVM regression).

Assume that training input and output data are defined as vectors $x_i \in R^n$ and $y_i \in R^k$ for i=(1, …, N) which are independent and identically distributed data with sample size *N*. ε-SVM and υ-SVM regression models map x_i to higher dimensional feature space R^k, where k and n ($k>>n$) represent the dimensions of feature space, using a function of $\phi(x_i)$ to linearize the nonlinear relationship between x_i and y_i. The estimation function of y_i is defined as $y = w.\phi(x) + b$, where w and b are the vectors of coefficients and a constant, respectively. These two parameters for each ε-SVM and υ-SVM regression model are derived by minimizing the error function Eqs. (1) and (2), respectively; subject to Eq. (3):

$$\frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i + C\sum_{i=1}^{N} \xi_i^{\bullet} \tag{1}$$

$$\frac{1}{2} w^T w + C\left(\upsilon\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\left(\xi_i + \sum_{i=1}^{N} \xi_i^{\bullet} \right) \right) \tag{2}$$

$$\begin{aligned} & w^T \phi(x_i) + b - y_i \le \varepsilon + \xi_i^{\bullet} \\ & y_i - w^T \phi(x_i) - b \le \varepsilon + \xi_i^{\bullet} \\ & \xi_i, \xi_i^{\bullet} \ge 0 \quad , \quad i = 1, ..., N \end{aligned} \tag{3}$$

where C = the capacity constant, ε = the accuracy demanded for the approximation, ϕ is the kernel function, and ξ_i and $\xi_i^{\bullet}$ are slack variables.

Different kernels can be selected to construct different types of SVM regression models. Typical examples include: polynomial kernels which have three tuning parameters; sigmoid kernel which has two tuning parameters; and RBF kernel which has one tuning parameter. The RBF kernel, given by γ in Eq. (4), has been reported as the best choice over other kernel functions (Dibike et al., 2001; Noori et al., 2009a). Parameter γ controls the amplitude of the Gaussian function and thus, controls the generalization ability of SVM.

$$K(x_i, x) = \exp\left(-\gamma \left| x_i - x \right|^2 \right) \tag{4}$$

RESULTS AND DISCUSSION

In this paper, the RBF is used as the kernel function of ε-SVM and υ-SVM regression models for the following reasons. First, unlike the linear kernel, the RBF kernel can handle the case when the relation between class labels and attributes is nonlinear. Besides, the linear kernel is a special case of RBF (Keerthi et al., 2003). Second, the number of tuning parameters influences the complexity of model selection. The RBF kernel has fewer tuning parameters than the polynomial and sigmoid kernels (Li et al., 2008). Another reason is that RBF kernels tend to give good performance under general smoothness assumptions (Noori et al., 2009a).

Building ε-SVM and υ-SVM regression models from training set requires values for *C* and ε (tuning parameters of ε-SVM model), *C* and υ (tuning parameters of υ-SVM model) and γ, while using the RBF kernel function. Fine tuning of these variables can greatly improve the generalization capacity of the prediction system. The γ value is important in RBF model and can lead to under-fitting and over-fitting in prediction. Under-fitting happens when the models are unable to predict the data that have been trained. Conversely, over-fitting occurs when the models tend to memorize all the training data but are unable to be generalized for unseen data; hence, only trained data points can be predicted. A massive increase in the γ will cause the risk of over-fitting because all the support vectors distances are taken into account; thus a complex model is built. On the other hand, when the γ value is changed to an extremely small value, the machine would ignore most of the support vectors and hence leads to a failure in the trained point prediction, known as under fitting (Han et al., 2007). The γ parameter has a default value in *Statistica* software equal to 1/*k*, where *k* is the number of input variables (in this step *k* is equal to 12). The best fitting γ value can be obtained by trial and error. In this research, γ parameter is set to several values. Parameter *C* is a regularization parameter that controls the trade-off between maximizing margin and minimizing training error. If *C* is too small, insufficient stress will be placed on fitting the training data. On the other hand, too large *C* values result in over-fitting of the training data. But, Wang et al. (2003) indicated that prediction error is scarcely influenced by *C*. The optimal value for ε and υ depends on the type of noise present in data, which is usually unknown. Even if enough knowledge of the noise is available for selecting an optimal value for ε and υ, there is a practical consideration of the number of resulting support vectors (Liu et al., 2006). In this research for finding optimal *C*, ε and υ values, a grid search in each model is performed as described by Hsu et al. (2003). The grid search algorithm works in simple way. It simply takes samples from the space of the independent variables. For each sample, the model predictions is computed and compared with the best value found from the previous iterations. If the newly found value is better than the previous one, the new results are stored. This process is repeated until the end of iterations is reached. The grid search algorithm performs the sampling by partitioning the space of the independent variables into a grid scheme, with the boundaries of the grid determined by the start and end values for the independent variables. The intensity of the computation is determined not only by the start and end values but also by the size of the steps used to forward the search from one location of the independent space to another. When using the grid search algorithm for response optimization, it is recommended that search is confined to those regions of independent space falling within the boundaries of the data set for which the models where trained. Making predictions in regions of independent space substantially laying

outside the boundaries of the data set used to train the model may lead to unreliable results. The grid search technique is unguided algorithms based on brute computing power. Hence, it can be computationally more expensive than other optimization techniques. For solving this problem, a two-step grid search method (Chen & Yu, 2007) with ten-fold cross-validation is used to derive the tuning parameters in the ε-SVM and υ-SVM regression models (C, ε, υ). First, a coarse grid search is used to determine the best region of these three-dimensional grids. Then, a finer grid search is conducted to find the optimal parameters. The training and the following forecasting work in this study are performed using the *Statistica* software. However, RMSE has been assessed on different γ values in the ε-SVM (Fig. 2) and υ-SVM (Fig. 3) regression models and then 22 models for each type of SVM regression technique is developed.

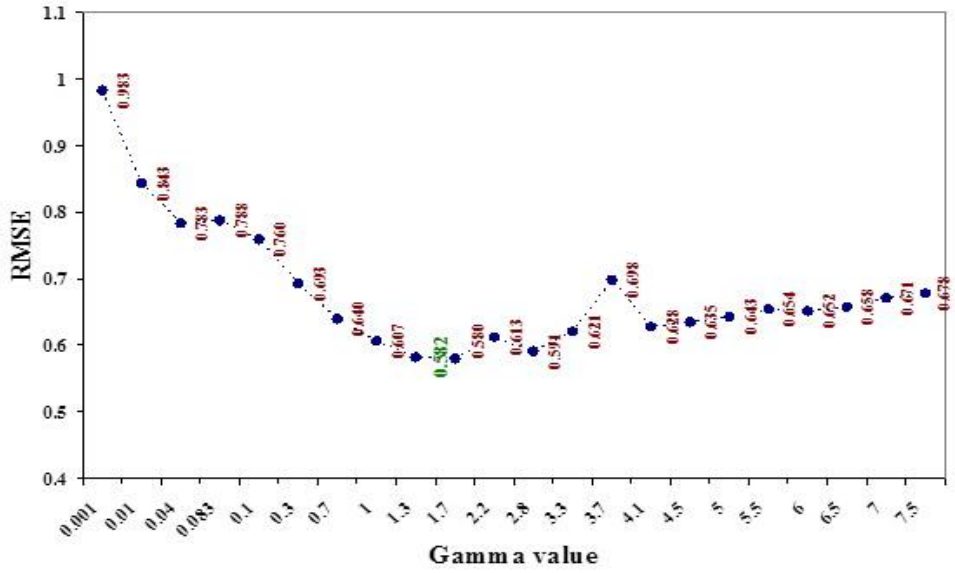

Fig. 2. The Gamma (γ) values versus RMSE index for the ε -SVM model

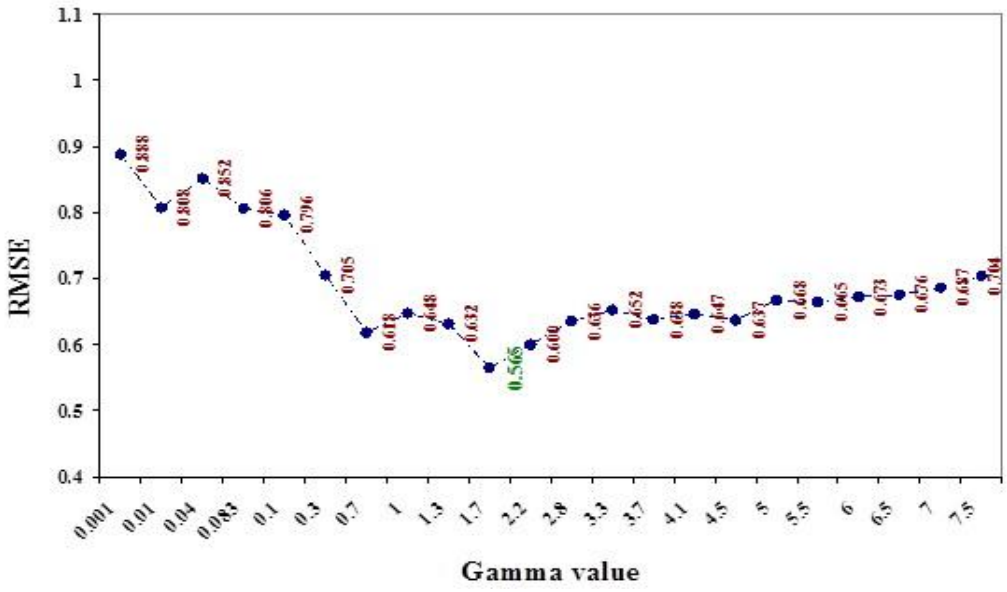

Fig. 3. The Gamma (γ) values versus RMSE index for the υ -SVM model

Among these 22 developed SVM models, the optimal parameters in the ε-SVM and υ-SVM regression models in the training phase are achieved as (C, ε, γ)=(6.00, 0.016, 1.7) and (C, υ, γ)=(6.00, 0.498, 1.7), respectively. The average daily CO concentrations are almost identical to the observed ones, indicating that the model is well trained, and that it can describe the relevant input-output relationship. The results of training and testing steps for ε-SVM regression and υ-SVM regression models are shown in Table 2 based on R^2 and mean absolute error (MAE). Also, the temporal variations of the observed and predicted average daily CO concentration using the ε-SVM and υ-SVM regression models for testing period are plotted in Figs. 4 and 5, respectively.

These figures and Table 2 indicate that the both models are comparable in terms of prediction accuracy and they have same performance in prediction of average daily CO concentration. Thus, type of SVM regression model does not have any effect on the SVM models operation. In addition, in the υ-SVM regression model, prediction error is scarcely influenced by υ values and for all of the 22 υ-SVM regression models, which are developed in this research, the υ value varied between 0.2 to 0.5. *C* and γ values, in the 22 υ-SVM regression models, varied between 5 to 250 and 0.001 to 7.5, respectively.

In the present investigation, the FS method is used as a linear input selection technique in order to select the best subset of 12 input candidates. First, correlation between each input variable and the desired output is evaluated as in Table 3.

Table 2. Results of training and testing of ε-SVM, υ-SVM, FS-(ε-SVM), and FS-(υ-SVM) models

Model	Training		Testing	
	R^2	MAE	R^2	MAE
ε-SVM	0.95	0.27	0.87	0.40
υ-SVM	0.94	0.30	0.87	0.41
FS-(ε-SVM)	0.94	0.29	0.90	0.34
FS-(υ-SVM)	0.93	0.29	0.90	0.35
PCA-MLR (Noori et al. (2008))	0.44	0.91	0.34	0.851
FS-ANFIS (Noori et al. (2009a))	0.90	0.36	0.91	0.35
FS-ANN (Noori et al. (2009a))	0.90	0.36	0.9	0.39

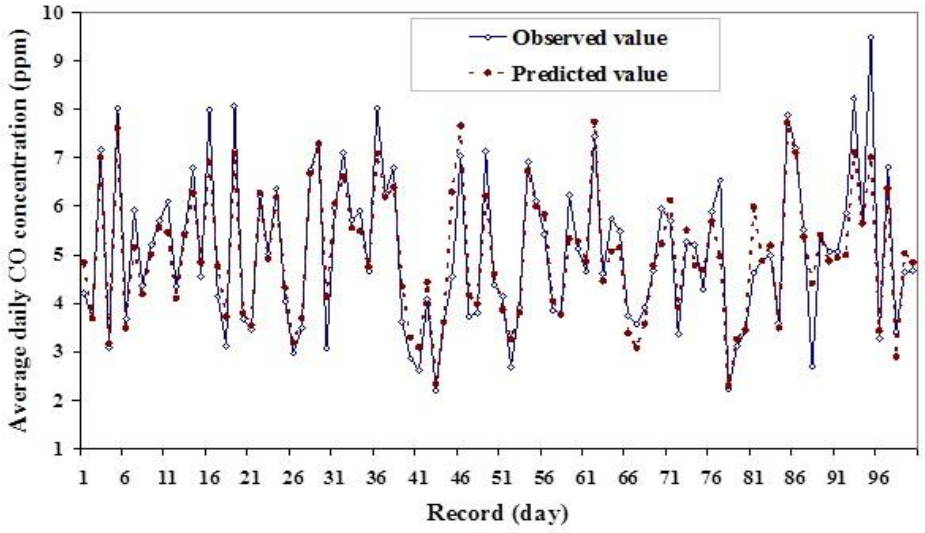

Fig. 4. Predicted and observed average daily CO concentration for testing of ε-SVM regression model

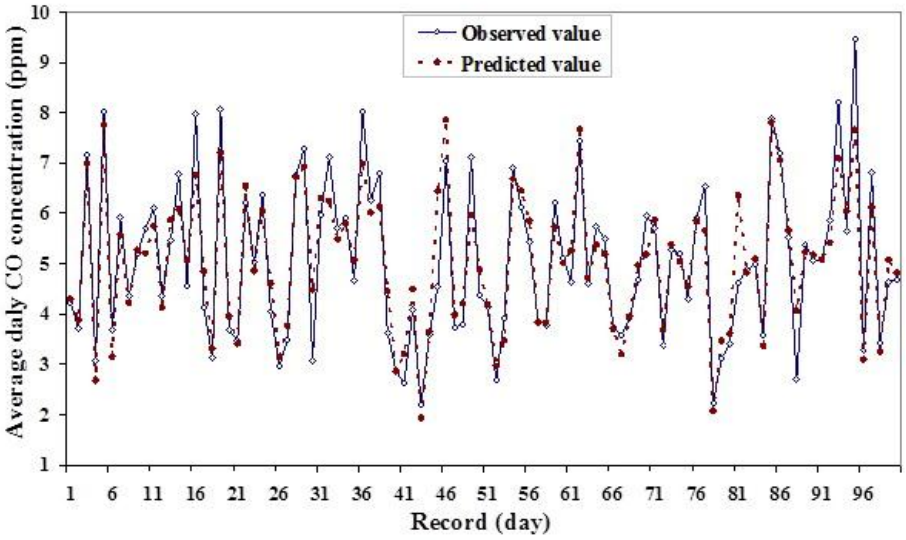

Fig. 5. Predicted and observed average daily CO concentration for testing of υ-SVM regression model

Table 3. Correlation between each input variable and the output variable CO

Variable	R^2	Variable	R^2
Temp	0.269864	O_3	0.026751
Hum	0.200260	WS	0.026637
Press	0.182048	NO_x	0.022939
Solar	0.090988	THC	0.019880
SO_2	0.078278	CH_4	0.014117
PM_{10}	0.070747	WD	0.007734

Table 4. Results of FS procedure

Input subset	R^2
Temp	0.2699
Temp,SO_2	0.3458
Temp, SO_2, THC	0.4594
Temp, SO_2, THC, Press	0.4615
Temp, SO_2, THC, Press, CH_4	0.5193
Temp, SO_2, THC, Press, CH_4, NO_x	0.6055
Temp, SO_2, THC, Press, CH_4, NO_x, O_3	0.6325*
Temp, SO_2, THC, Press, CH_4, NO_x, O_3, Solar	0.6337
Temp, SO_2, THC, Press, CH_4, NO_x, O_3, Solar, PM_{10}	0.6339
Temp, SO_2, THC, Press, CH_4, NO_x, O_3, Solar, PM_{10}, Hum	0.6345
Temp, SO_2, THC, Press, CH_4, NO_x, O_3, Solar, PM_{10}, Hum, WD	0.6347
Temp, SO_2, THC, Press, CH_4, NO_x, O_3, Solar, PM_{10}, Hum, WD, WS	0.6349

*After this value, variations of R^2 are negligible and thus, inputs related to this value are selected.

Second, the variable with highest correlation, i.e. Temp with R^2=0.26, is selected as the first and the most important input. Then, remained candidates are implemented into the model one by one and the new variable which provides the best modeling result is selected as new input and added to previous selected input (i.e. Temp). For evaluation of modeling goodness, R^2 is used. This step is repeated several times until adding new variable to inputs, has not significant improvement in the model output. In other words, if the increase in R^2 is more than 5%, the new variable is selected. Finally, input variables with most significant effect on output are selected and other variables are

removed. Results of FS are shown in Table 4 where seven candidates according to their importance are selected as input variables: Temp, SO_2, THC, Press, CH_4, NO_x, and O_3 and other candidates are removed.

In this section, to evaluate the effect of input selection on ε-SVM and υ-SVM regression models operation, the input variables resulting from FS method are considered as SVM inputs, i.e. 7 inputs (FS-(ε-SVM) and FS-(υ-SVM) models). The optimal parameters for FS-(ε-SVM) and FS-(υ-SVM) regression models are obtained same as the previous section (SVM models development). The best fitting γ value can be obtained by try and error. From Figs. 6 and 7 it is clear that between 26 developed models for each type of SVM regression model, the best fitting γ value for FS-(ε-SVM) and FS-(υ-SVM) regression models are 0.500 and 0.496, respectively.

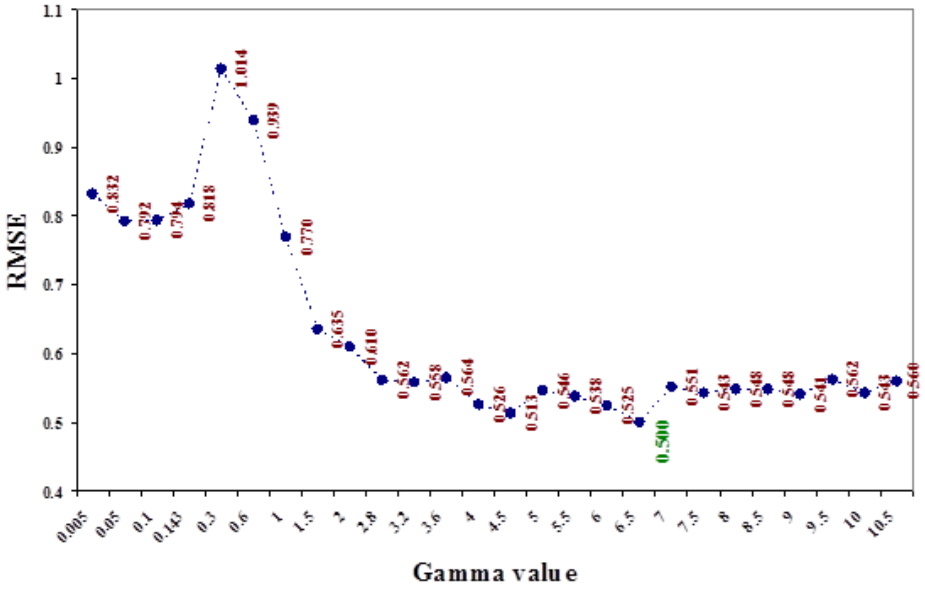

Fig. 6. The Gamma (γ) values versus RMSE index for the FS-(ε-SVM) regression model

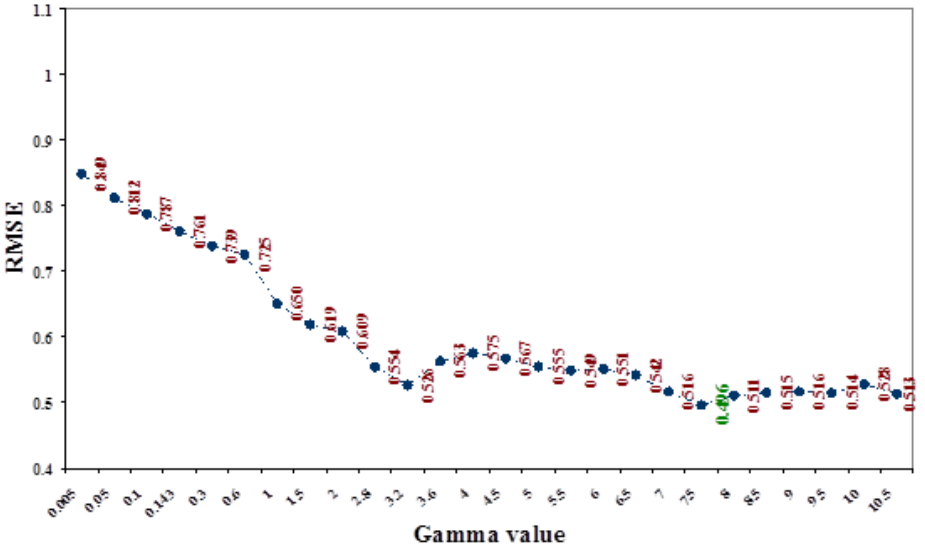

Fig. 7. The Gamma (γ) values versus RMSE index for the FS-(υ-SVM) regression model

Similar to previous section, a two-steps grid search method with ten-fold cross-validation is used to derive the tuning parameters in the models $(C, \varepsilon, \upsilon)$. However, the optimal values of C and ε in the FS-(ε-SVM) regression model are obtained as 3.200 and 0.045 by grid search algorithm, respectively, whereas for the FS-(υ-SVM) regression model, the best fitting C and υ values are 2.100 and 0.465, respectively. Results of these models for training and testing steps are listed in Table 2. Furthermore, observed and predicted average daily CO concentrations by both models are calculated and shown in Figs. 8 and 9, respectively. These figures and Table 2 indicated that both models have same performance in prediction of average daily CO concentrations. Thus, similar to previous section, type of SVM regression model does not have any effect on the SVM models operation.

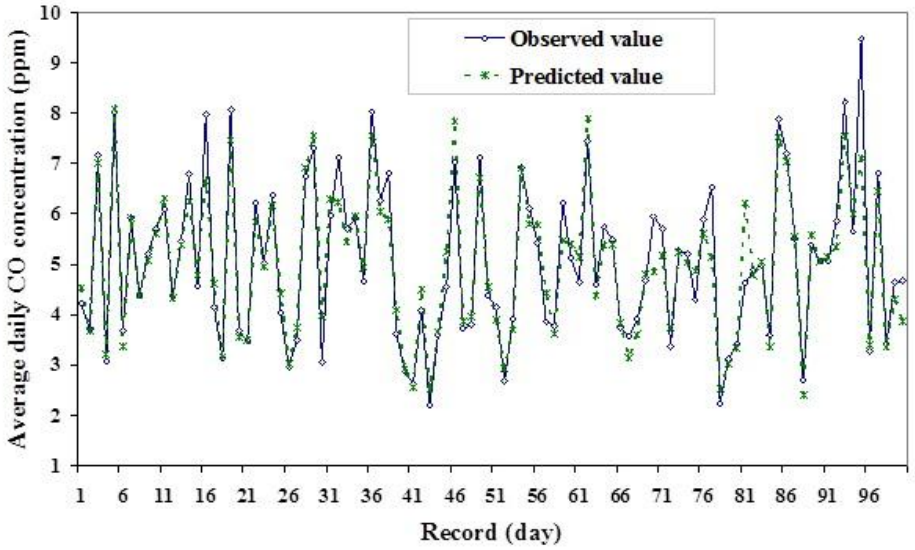

Fig. 8. Predicted and observed average daily CO concentration for testing of FS-(ε-SVM) regression model

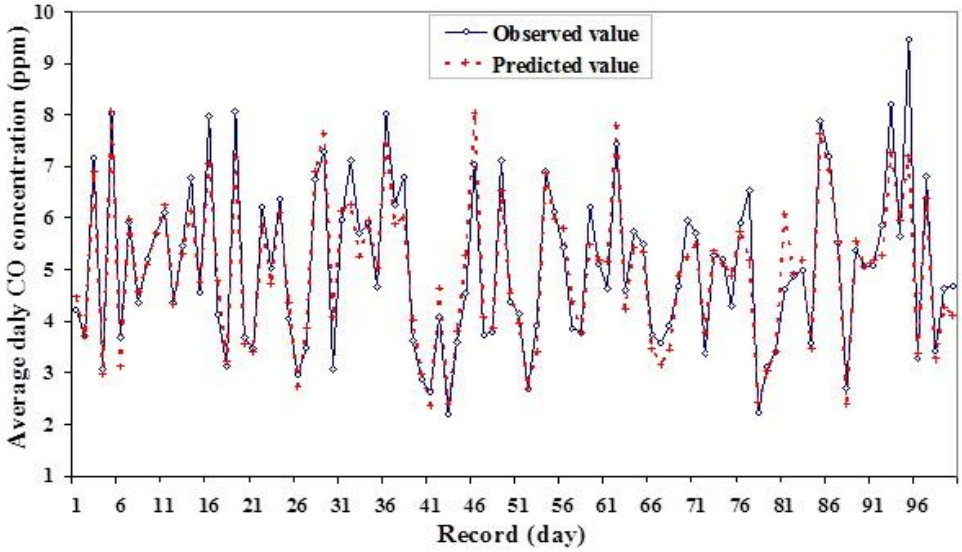

Fig. 9. Predicted and observed average daily CO concentration for testing of FS-(υ-SVM) regression model

In the past decades different statistical models such as MLR, ANN, ANFIS, and SVM models have been used by many researchers for short term air pollution forecasting in many parts of the world. The present work is the progress of tow previously papers published for daily average CO concentration in the atmosphere of Tehran. In the first work authors developed MLR model based on PCA (PCA-MLR model) to forecast daily CO concentration in the atmosphere of Tehran (Noori et al., 2008). The second research investigated the capability of ANN and ANFIS models developed using two data reduction techniques, i.e. FS and Gamma test methods. It concluded that FS technique was a proper solution to improve the performance of ANN and ANFIS models (Noori et al., 2010a). Results of PCA-MLR, FS-ANN, and FS-ANFIS models for training and testing steps are presented in Table 2.

According to Table 2, it is proved that the AI techniques have better performance than PCA-MLR model. In addition, it is founded that the performance of FS-(ε-SVM) and FS-(υ-SVM) models are generally similar to the best FS-ANFIS and FS-ANN solutions for short term forecasting of CO concentrations. Moreover, for better judgment, the developed discrepancy ratio (DDR), which is proposed by Noori et al. (2010a), is used to check the robustness of the models. The normalized value of DDR (Q_{DDR}) for all models in testing step is calculated and the standard normal distribution for each model is illustrated in Fig. 10. It should be noted that in error distribution graph, more tendencies to the centerline and also, larger value of the maximum Q_{DDR}, indicate more accuracy. The maximum Q_{DDR} values for models are plotted in this figure. The values for FS-ANFIS, FS-ANN, FS-(ε-SVM), and FS-(υ-SVM) models are 4.1, 4.2, 4.4, and 4.3, respectively. Considering the maximum Q_{DDR} values, all four models approximately have the same performance, although FS-(ε-SVM) and FS-(υ-SVM) regression models slightly have better performance than the others.

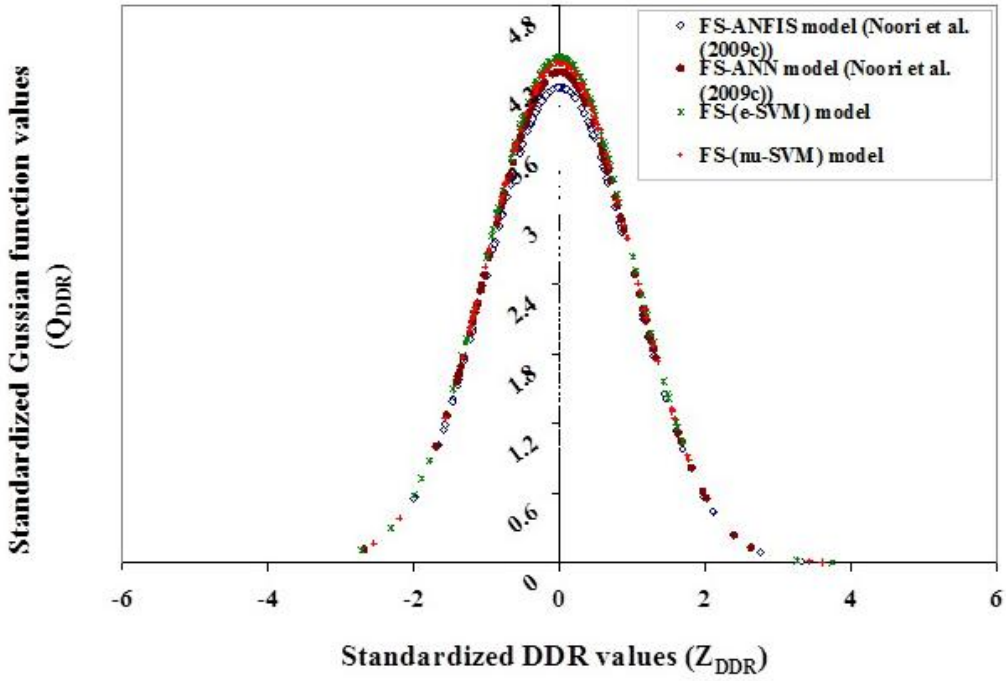

Fig. 10. Standardized normal distribution graph of DDR values for ε (e)-SVM, υ (nu)-SVM, FS-(ε (e)-SVM), and FS-(υ (nu)-SVM) regression models in the testing step (Noori et al., 2008)

CONCLUSION

Considering the importance of daily CO concentration in the atmosphere of Tehran, this research aimed to develop proper prediction models using SVM technique. Since input selection is a significant step in modeling, FS method is used and SVM models are developed. The goodness of each model is evaluated using R^2 and MAE statistics and also, DDR. The following conclusions could be founded from the present study: 1- Input selection improves prediction capability of both ε-SVM, υ-SVM regression models. It reduced the output error and improved the model performance. 2- The type of SVM regression models does not have any significant effect on SVM models operation. 3- Considering R^2 and MAE indices, FS-ANFIS, FS-ANN, FS-(ε-SVM), and FS-(υ-SVM) models approximately have the same performance. Also, the same result is obtained using DDR. 4- To reduce computational efforts, we recommend developing a method for SVM parameter optimization instead of grid search algorithm. In addition, since the result of SVM regression techniques are faced with uncertainty, we recommend that corrective measures such as utilization of Mont-Carlo simulation method should be considered. 5- Finally, input variables preprocessing by FS is recommended for increasing SVM regression models operation, especially in some cases which there is not a great deal of knowledge about the input variables.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Anguita, D., Boni, A. and Ridella, S. (1999). Learning algorithm for nonlinear support vector machines suited for digital VLSI. Electron. Lett., 35(16); 1349-1350.

Azeez, O., Pradhan, B. and Shafri, H. (2018). Vehicular CO emission prediction using support vector regression model and GIS. Sustainability-Basel., 10(10); 1-18.

Bayat, R. (2005). Source Apportionment of Tehran's Air Pollution. M. Sc thesis. Department of Civil and Environmental Engineering, Sharif University of Technology, Tehran, Iran,

Bray, M. and Han, D. (2004). Identification of support vector machines for runoff modelling. J. Hydroinform., 6(4); 265-280.

Brown, M., Gunn, S. R. and Lewis, H. G. (1999). Support vector machines for optimal classification and spectral unmixing. Ecol. Model., 120(2-3); 167-179.

Chen, S. T. and Yu, P. S. (2007). Real-time probabilistic forecasting of flood stages. J. Hydrol., 340(1-2); 63-77.

Chen, S., Hong, X., Harris, C. J. and Sharkey, P. M. (2004). Sparse modeling using orthogonal forward regression with PRESS statistic and regularization. IEEE T. Syst. Man. Cy. B., 34(2); 898-911.

Cristianini, N. and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge university press.

Dehghani, M., Saghafian, B., Nasiri Saleh, F., Farokhnia, A. and Noori, R. (2014). Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. Int. J. Climatol., 34(4); 1169-1180.

Dibike, Y. B., Velickov, S., Solomatine, D. and Abbott, M. B. (2001). Model induction with support vector machines: introduction and applications. J. Comput. Civil Eng., 15(3); 208-216.

Eksioglu, B., Demirer, R. and Capar, I. (2005). Subset selection in multiple linear regression: a new mathematical programming approach. Comput. Ind. Eng., 49(1); 155-167.

Elangasinghe, M., Dirks, K., Singhal, N., Costello, S., Longley, I. and Salmond, J. (2014). A simple semi-empirical technique for apportioning the impact of roadways on air quality in an urban neighbourhood. Atmos. Environ., 83; 99-108.

Ghaemi, Z., Alimohammadi, A. and Farnaghi, M. (2018). LaSVM-based big data learning system for

dynamic prediction of air pollution in Tehran. Environ. Monit. Assess., 190(300); 1-17.

Gokhale, S. and Pandian, S. (2007). A semi-empirical box modeling approach for predicting the carbon monoxide concentrations at an urban traffic intersection. Atmos. Environ., 41(36); 7940-7950.

Han, D., Chan, L. and Zhu, N. (2007). Flood forecasting using support vector machines. J. Hydroinform., 9(4); 267-276.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. and Scholkopf, B. (1998). Support vector machines. IEEE Intell. Sys., 13(4); 18-28.

Hrust, L., Klaić, Z. B., Križan, J., Antonić, O. and Hercog, P. (2009). Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. Atmos. Environ., 43(35); 5588-5596.

Hsu, C. W., Chang, C. C. and Lin, C. J. (2003). A practical guide to support vector classification. Tech. repurt. Department of Computer Science, National Taiwan University.

Jalili Ghazi Zade, M. and Noori, R. (2008). Prediction of municipal solid waste generation by use of artificial neural network: A case study of Mashhad. Int. J. Environ. Res., 2(1); 13-22.

Jorquera, H. (2002). Air quality at Santiago, Chile: a box modeling approach—I. Carbon monoxide, nitrogen oxides and sulfur dioxide. Atmos. Environ., 36(2); 315-330.

Kecman, V. (2005). Support vector machines–an introduction, in "Support Vector Machines: Theory and Applications (pp. 1-47)". New York: Springer.

Keerthi, S. S. and Lin, C. J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput., 15(7); 1667-1689.

Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007). Building a robust linear model with forward selection and stepwise procedures. Comput. Stat. Data An., 52(1); 239-248.

Li, X., Lord, D., Zhang, Y. and Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. Accident Anal. Prev., 40(4); 1611-1618.

Liong, S. Y. and Sivapragasam, C. (2002). Flood stage forecasting with support vector machines. J. Am. Water Resour. As., 38(1); 173-186.

Liu, H., Yao, X., Zhang, R., Liu, M., Hu, Z. and Fan, B. (2006). The accurate QSPR models to predict the bioconcentration factors of nonionic organic compounds based on the heuristic method and support vector machine. Chemosphere, 63(5); 722-733.

Lu, W. Z. and Wang, W. J. (2005). Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. Chemosphere, 59(5); 693-701.

Luna A.S., Paredes M. L. L., De Oliveira G. C. G. and Corrêa S. M. (2014). Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. Atmos. Environ., 98; 98–104.

Moazami, S., Noori, R., Amiri, B. J., Yeganeh, B., Partani, S. and Safavi, S. (2016). Reliable prediction of carbon monoxide using developed support vector machine. Atmos. Pollut. Res., 7(3); 412-418.

Murena, F., Favale, G., Vardoulakis, S. and Solazzo, E. (2009). Modelling dispersion of traffic pollution in a deep street canyon: application of CFD and operational models. Atmos. Environ., 43(14); 2303-2311.

Noori, R., Ashrafi, K. and Azhdarpour, A. (2008). Comparison of ANN and PCA based multivariate linear regression applied to predict the daily average concentration of CO: A case study of Tehran. J. Earth. Space. Phys., 34(1); 135-152.

Noori, R., Abdoli, M. A., Ghasrodashti, A. A. and Jalili Ghazizade, M. (2009a). Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: a case study of Mashhad. Environ. Prog. Sustain., 28(2); 249-258.

Noori, R., Karbassi, A., Farokhnia, A. and Dehghani, M. (2009b). Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. Environ. Eng. Sci., 26(10); 1503-1510.

Noori, R., Hoshyaripour, G., Ashrafi, K. and Araabi, B. N. (2010a). Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. Atmos. Environ., 44(4); 476-482.

Noori, R., Karbassi, A. and Sabahi, M. S. (2010b). Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste prediction. J. Environ. Manage., 91(3); 767-771.

Noori, R., Karbassi, A., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M., Farokhnia, A. and Gousheh, M. G. (2011). Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. J. Hydrol., 401(3-4); 177-189.

Noori, R., Yeh, H. D., Abbasi, M., Kachoosangi, F. T. and Moazami, S. (2015a). Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand. J. Hydrol., 527; 833-843.

Noori, R., Deng, Z., Kiaghadi, A. and Kachoosangi, F.T. (2015b). How reliable are ANN, ANFIS, and SVM techniques for predicting longitudinal dispersion coefficient in natural rivers?. J. Hydraul. Eng., 142(1); p.04015039.

Nunnari, G., Dorling, S., Schlink, U., Cawley, G., Foxall, R. and Chatterton, T. (2004). Modelling SO2 concentration at a point with statistical approaches. Environ. Modell. Softw., 19(10), 887-905.

Osowski, S. and Garanty, K. (2007). Forecasting of the daily meteorological pollution using wavelets and support vector machine. Eng. Appl. Artif. Intel., 20(6); 745-755.

Pelliccioni, A. and Tirabassi, T. (2006). Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. Environ. Modell. Softw., 21(4); 539-546.

Perez-Roa, R., Castro, J., Jorquera, H., Perez-Correa, J. and Vesovic, V. (2006). Air-pollution modelling in an urban area: Correlating turbulent diffusion coefficients by means of an artificial neural network approach. Atmos. Environ., 40(1); 109-125.

Pérez, P., Trier, A. and Reyes, J. (2000). Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ., 34(8); 1189-1196.

Sahin, U., Ucan, O. N., Soyhan, B. and Bayat, C. (2004). Modeling of CO distribution in Istanbul using artificial neural networks. Fresenius Environ. Bullet., 13(9); 839-845.

Salazar-Ruiz, E., Ordieres, J., Vergara, E. and Capuz-Rizo, S. F. (2008). Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US). Environ. Modell. Softw., 23(8); 1056-1069.

Shakerkhatibi, M., Mohammadi, N., Zoroufchi Benis, K., Behrooz Sarand, A., Fatehifar, E. and Asl Hashemi, A. (2015). Using ANN and EPR models to predict carbon monoxide concentrations in urban area of Tabriz. Environ. Health Eng. Manage. J., 2(3); 117-122.

Vapnik, V. (1998). Statistical learning theory. (Vol. 3) New York: Wiley.

Wang, W., Xu, Z., Lu, W. and Zhang, X. (2003). Determination of the spread parameter in the Gaussian kernel for classification and regression. Neurocomputing, 55(3-4); 643-663.

Wang, X., Chen, S., Lowe, D. and Harris, C. J. (2006). Sparse support vector regression based on orthogonal forward selection for the generalised kernel model. Neurocomputing, 70(1-3); 462-474.

Wang, Z., Lu, F., Lu, Q. C., Wang, D. and Peng, Z. R. (2015). Fine-scale estimation of carbon monoxide and fine particulate matter concentrations in proximity to a road intersection by using wavelet neural network with genetic algorithm. Atmos. Environ., 104; 264-272.

Yu, P. S., Chen, S. T. and Chang, I. F. (2006). Support vector regression for real-time flood stage forecasting. J. Hydrol., 328(3-4); 704-716.