Available online at www.ap.iauardabil.ac.ir
Islamic Azad University, Ardabil Branch
Anthropogenic Pollution Journal, Vol 5 (1), 2021: 81-97
DOI: 10.22034/ap.2021.1923264.1091
ISSN: 2588-4646

Estimation of Concentration of Air Pollutants in Shazand Thermal Power Plant with Support Vector Machine Model Based on Selection of Effective Input Variables with Partial Mutual Information (PMI) Algorithm of Distribution of Air Pollutants

Ghodratollah Siahpour [1], Seyed Ali Jozi[2*], Neda Orak[1], Hossein Fathian[3], Solmaz Dashti[1]
1. Department of Environment, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran
2*. Department of Environment, North Tehran Branch, Islamic Azad University, Iran
3. Department of Water Resources Engineering, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran
*Correspondence author: sajozi@yahoo.com

Received: 11 February 2021/ **Accepted**: 26 March 2021/ **Published**: 28 March 2021

Abstract: Due to the difficulty of estimating the pollutant gas concentration in power plants, this study aimed to estimate the concentration of the air pollutants in a thermal power plant using the support vector machine model (SVM). The concentration of environmental pollutants in the thermal power plant, Shazand, Iran, at different distances from the chimney was estimated using SVM. The effective input variables in the SVM model were selected using the Partial Mutual Information (PMI) algorithm. The modeling period was weekly from December 2018 to December 2019. Effective input variables for estimating the concentration of CO2, SO2, NOX, and CO pollutants at different distances are the the same gas concentration at the chimney of the power plant. Effective input variable for estimating CO gas concentration at a distance of 5 km was average wind speed and CO gas concentration were obtained at the chimney location and the maximum gas concentration among air pollutants was CO2 (2811.63 *µg/m3)*, occurring at a distance of 5 km from the plant chimney Co (5.5 *µg/m3*, occurring at a distance of 20 km from the power plant chimney. The polynomial kernel function is the best kernel function of the SVM model for estimating SO_2 and NO_X concentrations at different distances and the best kernel function in the SVM model for estimating CO_2 and CO concentrations. The SVM model has good accuracy and performance in estimating the pollutant concentrations, and selecting effective input variables in the SVM model with the PMI algorithm increases the model accuracy.

Keywords: Concentration of gaseous pollutants, Chimney, Hampel test, SVM model.

1. Introduction

The air pollution crisis has become a human catastrophe due to improper urban planning. The energy produced in a thermal power plant by combustion of fuel, such as fossil fuel energy, constitutes the highest air pollution in the world, emitting 85% of the pollution of particles inhaled in the air (Awasthi, et al. 2006; Conti, et al. 2016; Perera 2018; Pope III 2007), and several parameters affect it (Hasenfratz, et al. 2012; Zheng, et al. 2013). This study aimed to estimate the concentration of environmental pollutants: carbon monoxide (CO), *carbon dioxide* (CO_2), *sulfur dioxide* (SO_2) and nitrogen dioxide (NO_X) from the fuel of a thermal power plant with the support vector machine (SVM) model at different distances from the power plant. Also, the effective input variables in the SVM model are determined for 4 types of pollutants and at different distances by the partial mutual information algorithm (PMI).

The results of a 2020 study in South Korea showed that teaching non-default models improves the accuracy of the test phase by at least 10% and 40% with efficiency coefficients of 0.8925 and 0.9904 for SOx and NOx, respectively, and provides a maximum reduction in computational time of 46.67% to predict NOx emissions (Adams, et al. 2020). In a study, Vahid et al. used neural networks to solve the problem of improper distribution of pollution monitoring stations (Wahid, et al. 2013). In a study conducted by Reikard, the efficiency of neural networks and vector machines in predicting air pollution was compared, and the results showed that the support vector machine performed better than the neural network (Reikard 2012). Due to their greater ability to model nonlinear relationships, they have attracted the attention of researchers in the field of energy prediction. These techniques include various types of neural networks such as radial basis function (RBF), SVM, error backpropagation (BP) learning, and fuzzy logic

(Alexiadis, et al. 1998; Hosseinnezhad and Babaei 2013; Kariniotakis, et al. 1996; Zhou, et al. 2011). Suspended particle concentrations are more influenced by human factors. On the other hand, the volume of suspended particles in urban areas and foam factories and the use of various fuels in residential areas are effective (Lippmann, et al. 2003). The electricity sector is one of the most important sources of emissions, especially greenhouse gases in the world. According to estimates, about 37.5% of global carbon emissions are due to thermal power plant operations. Reducing greenhouse gas emissions from the electricity sector requires the use of different energy patterns to generate electricity. Comparing the amount of greenhouse gas emissions from different electrical technologies helps to select effective methods to reduce greenhouse gas emissions (Liu, et al. 2007). Recently, based on statistical methods, algorithms have been proposed for determining effective input variables in models based on data processing. The only nonlinear algorithm for determining effective input variables in data-based models is the partial mutual information (PMI). Due to population growth and the environmental consequences of industrialization, the ability to accurately predict the concentration of air pollutants has become increasingly important. The chimney of the power plant is one of the main sources of air pollution that emits large amounts of NOx, nitrogen dioxide (NO_2) (Lu, et al. 2004), and the volume of NOx and NO_2 in many countries as one of the most important air pollutants mentioned (Arain, et al. 2007). The purpose of this study is to properly estimate the concentration of environmental pollutants at different distances from the power plant, and to find the pollutants from the production process of Shazand, Arak thermal power plant, located in Markazi Province, Iran, and the impact of each pollutant on the environment and health of people working in the power plant and its inhabitants using a nonlinear algorithm based on vector machine of design (modeling).

2. Materials and Methods

Introduction of Shazand thermal power plant in Arak

Shazand thermal power plant in a geographical position of 49 degrees and 25 minutes and 50 seconds east longitude and 34 degrees and 3 minutes and 41 seconds north latitude, which is located in the southwestern region of the city of Arak in Markazi Province. Located at km 15 from the Arak-Boroujerd road and east of Shazand refinery. It is one of the thermal power plants of Iran, its electricity generation capacity is 1300 MW and includes 4 steam units of 325 MW. Given the proximity of this power plant to residential areas, it is important to study the pollutants of this power plant and their impact on human communities. The main fuel of this power plant is natural gas and fuel oil, which emits pollutants such as SO_2 and NO_X into the environment when using fuel oil. The characteristics of the chimney in the Shazand power plant are shown in Table 1. To use the support vector machine (SVM) to model the dispersion of air pollutants from Shazand thermal power plant in Arak, Shazand synoptic station's meteorological data were used. The longitude and latitude of Shazand synoptic station are 49 25 north and 33 57 east, respectively, and its height is equal to 1913 meters above sea level (Central Province Meteorological Department-2019).

Table1: Chimney characteristics in Shazand power plant

Number of chimneys	Chimney Height (m)	Inner diameter of chimney end (m)	Exhaust gas speed (m/s)	Production of each unit (MW)	Amount of fuel consumed (m3/h)	Fuel Type	Period of use of any fuel per year (month)	Output Temperature (Kelvin)
2	200	8.75	9.1	300	65900	Gas	8	410
				300	59	Fuel oil	4	415

Source: Shazand Power Generation Management Co: 2015, see at http://news.moe.gov.ir/Home (Kohandel, 2018)

Introduction of PMI algorithm

The only nonlinear algorithm for selecting input variables to determine effective input variables in data-based models is the PMI algorithm. The PMI-based input selection (*PMIS*) algorithm was first developed by Sharma (2000) to identify effective input variables in hydrological models. The statistical concept that PMI estimates for Cs is based on confidence bounds determined from a distribution formed by a bootstrap loop. If the input is significant, Cs is added to S (set of selected input variables) and the selection continues, until no significant input remains, then the algorithm is subsequently stopped.

Estimating Partial Mutual Information (PMI)

According to a random output variable Y, there is some uncertainty about an observation y that is a member of the Y, which can be defined according to Shannon H entropy (Shannon 1948). But assuming a random input variable X to which Y depends, cross-observations (x, y) reduce this uncertainty. According to the definition of mutual information I (X; Y), the decrease in the uncertainty of the variable Y is due to the observation of X (Cover 1991). Mutual information (MI) can be calculated directly from the following *Relation* (May, et al. 2008).

$$I(X;Y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dx\,dy$$

(Relation 1)

Where p (y) and p (x) are the marginal probability density functions (*PDFs*) X and Y, respectively, and p (x, y) is the joint probability density functions. In practice, however, in *Relation 1*, the correct form of the probability density functions is unknown. Therefore, estimation of probability densities is used instead. By placing the density estimates, we will have the probability with the integral numerical approximation in *Relation 1* (Shannon 1948).

$$I(X;Y) \approx \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{f(x_i,y_i)}{f(x_i)f(y_i)}\right]$$

(Relation 2)

Where *f* represents the density estimated based on a sample of n observations of (x, y). Different bases for the logarithm in *Relation 2* can be used, but usually 2 or *e*. Assuming *Relation 2*, it can be said that accurate and effective estimation of mutual-information (MI) is highly dependent on the method used in estimating marginal and co-occurrence probability density functions. In total, there are three criteria for stopping the PMI algorithm: 1) Critical tabulated values, estimation of marginal and co-occurrence probability density functions; 2) Criteria based on Akaike information criterion (AIC) (Akaike 1974); 3) Hampel test criterion. Tables of critical values of the correlation coefficient (R) are readily available, based on the analytical formula for the error distribution of an estimate for the assumed sample size.

In the case of a linear correlation coefficient (R), the sample estimate distribution follows a t-distribution. Tables of critical values of the correlation coefficient (R) are based on the distribution of t (David 1938), which provide the critical value for R for the number of samples and a certain level of confidence. However, unlike the linear correlation coefficient, for calculating *I*, an equivalent analytical definition cannot be inferred from *Relation 2* (May, et al. 2008). Therefore, researchers for calculating *I*, should seek help from self-starters. Instead of using analytic values, one method for constructing tables of critical values calculated for *I* is to use the Monte Carlo simulation (David 1938). For each sample size, first a series ~ N (0,1) is constructed and then the marginal probability density functions are calculated. Two alternative stop-formulation criteria are formulated by which, in each iteration, *I'CsY.S* is compared with the corresponding critical values I (95) and I (99) obtained from the table of critical values estimated by the mutual-information, which of these two criteria Instead of direct calculations, the boot system is used to determine which variable should be selected or which algorithm should be stopped. The computational elimination of the boot system loop makes the selection of input variables much faster.

Hampel test criterion (Z)

Outlier determination methods, as a robust statistical method for determining whether a given x value is significantly different from other values in a set of X values by more than 3 times the deviation. This test compares the deviation of a single observation from the average of all observations. An observational value with a Z score greater than 3 is usually considered as outlier data by 3α law for normal distribution. Hence, a modified Z value, which uses Hampel spacing, is used instead of increasing the efficiency of the method (Goebel, et al. 2005). The Hampel distance is based on the middle of the set of inputs. The failure point is the Hampel test and it is known as one of the most powerful tests for detecting outliers despite several outliers (Davies and Gather 1993; Pearson 2002). The Hampel test begins with the calculation of the absolute deviation from the median PMI for all inputs as follows:

$$d_j = \left|I_{C_jY.S} - I_{C_jY.S}^{(50)}\right|$$

(Relation 3)

In this relation, d_j represents the absolute deviation, $I_{C_jY.S}$ is equal to the PMI value for the variable x and $I_{C_jY.S}^{(50)}$ is the median PMI for the set of C inputs. Then the hemp distance can be determined as follows:

$$Z_j = \frac{d_j}{1.4826\,d_j^{(50)}}$$

(Relation 4)

Which Z_j represents the Hampel distance for the input set Cj and $d_j^{(50)}$ indicates the median absolute deviation 25 (MAD), dj. The coefficient of 1.4826 changed the distance in such a way that the law Z> 3 could be applied, as used in the conventional Z-test (Pearson 2002). Using this stop criterion, the input-based PMI algorithm no longer contains the bootstrapping loop, and the PMI is not compared to any critical value of *I*.

SVM model

Support vector machine is one of the supervised learning methods used for classification and regression. The support vector machine is a non-statistical binary classifier that has received much attention in recent years. SVM features include: 1) Maximum generalization in the design of classifiers; 2) Ability to find the optimal answer of the function; 3) Automatic preparation of the optimal structure and mechanism in solving classification problems; 4) Modelling of nonlinear functions using nonlinear kernels and the internal multiplication method in Hilbert spaces (May, et al. 2006). The main difference between this method and conventional statistical classifiers is the ability to process hyperspectral data classification and reduce the number of bands in different modelling processes. If the data are linear and separate,

SVM uses linear machines to achieve an optimal level with minimal error and separates and teaches the maximum distance between the page and the nearest training points (backup vectors) (May, et al. 2006).

If the training points are conditioned with $x_i \in R^n$ in the form of $[x_i, y_j]$and input vector, then the value of each class is defined as $y_i \in \{-1, 1\}\ i = 1, \ldots, i$. It can be expressed as *Relation 5*:

$$Y = sgn(\sum_{i=1}^{N} y_i a_i (X \times X_i) + b)$$

(Relation 5)

Where the Y output is the sample class value of X_i, a_i, and b, and the parameters that determine the hyper plane. If it is not possible to separate the data linearly, then *Relation 5* is changed as follows:

$$Y = sgn(\sum_{i=1}^{N} y_i a_i K(X \times X_i) + b)$$

(Relation 6)

In this *Relation*, $K(X \times X_i)$ is a kernel function that generates internal beats to create support vector machines with different states from nonlinear decision levels in the data space, and for this purpose, the line equation needs to be defined. The line equation in two-dimensional space is calculated using *Relation 6*, the plane equation from *Relation 7*, and the display equation from *Relation 8* (Davies and Gather 1993).

$$w_1 x_1 + w_2 x_2 + b = 0$$

(Relation 7)

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + b = 0$$

(Relation 8)

In cases where the data are not linearly separable, they can be separable linearly by mapping the data to a feature space. In general, the dimensions of this space are infinite, so performing calculations in the feature space can be costly. To overcome this problem, kernel functions are used, so the equation of the separator plane for the nonlinear state with the intervention of the kernel function of $K(X)$is as follows (Davies and Gather 1993).

$$\mathbf{w^T K(X) + b = 0}$$

(Relation 9)

Where the function maps the data from a nonlinear space to a linear space. This function can also be defined as $K(X_i, X_j)$ and used to explore more complex spaces. This process is called the kernel trick. In the SVM model, the choice of kernel function is very important and in different problems, depending on the nature of the problem, different cases can be considered. Different types of functions are presented in Table 2. In the polynomial kernel function, the parameter d must be computed by trial and error or by optimization algorithms. The Gaussian or radial kernel function (RBF) is another type of kernel function and is usually used in cases where no knowledge of the type and nature of the data is available. In most studies using the backup vector machine, the effect of data periodicity and the type of kernel function used has not been much discussed, and in a few studies, the role and determination of the optimal kernel function have been addressed (Pearson 2002).

Table 2. Kernel functions used in backup vector machines

Function Type	Kernel function
Linear	$K(X_i, X_j) = X_i^T \times X_j$
Polynomial	$K(X_i, X_j) = (\gamma X_i^T \times X_j + C)^d$
RBF	$K(X_i, X_j) = -\exp(-\gamma \lvert X_i - X_j \rvert^2)$
Hyperbolic Tangent	$K(Xi, Xj) = tanh(\gamma XiT \times Xj + C)$

Source: (Hamel 2011)

Construction of educational patterns of SVM model

In order to estimate the concentration of gases, the meteorological variables of Shazand synoptic station were considered as inputs in the potential set of input variables. Also, the concentration of each gas at the chimney site (0 km) was considered as input. The concentration of each gas at distances of 5, 10, 15, and 20 km was considered as output variables. The time intervals of all variables were selected weekly from January 2019 to December 2019.

PMI algorithm was used to determine the variables affecting the output variable (concentration of each gas at distances of 5, 10, 15, and 20 km). This algorithm, which is added as an add-on 1 to software Excel 2007 and above, it should be noted that this algorithm considers by default that input variables have zero skewness and follow a normal distribution, so to use this algorithm, data should be normalized first. To normalize the data, logarithmic conversion was used and the converted data were introduced to algorithms.

efficiency coefficient (NSE) and root mean square error, and the mean absolute value of error were used which is stated in *Relations 10 to 12*, respectively. The Nish-Sutcliffe coefficient indicates the efficiency of the model which has been widely used, in *Relations 10 to 12*, n is equal to the number of data, $O_i$and S_in is are the estimated and observational data in the time step of i_M and $\overline{O}$ is the mean of observational data.

Table 3. Potential set of input and output variables resulting from the construction of educational patterns

Row	Variable Name	Introducing variable	Variable Type

1	Vmax (t)	Maximum wind speed	Input
2	dd_max (t)	Dominant wind direction	Input
3	Vm (t)	Average wind speed	Input
4	Tmax (t)	Maximum air temperature	Input
5	Tmin (t)	Minimum air temperature	Input
6	Tm (t)	Average air temperature	Input
7	Umax (t)	Maximum relative humidity	Input
8	Umin (t)	Minimum relative humidity	Input
9	Um (t)	Moderate relative humidity	Input
10	p0max (t)	Maximum air pressure	Input
11	p0min (t)	Minimum air pressure	Input
12	p0m (t)	Medium Air Pressure	Input
13	So2-km0 (t)	So2 gas concentration at chimney sit	Input
14	CO2-km0 (t)	CO2 gas concentration at chimney sit	Input
15	Co-km0 (t)	Co gas concentration at the chimney site	Input
16	Nox-km0 (t)	Nox gas concentration at chimney site	Input
17	So2-km5 (t)	So2 gas concentration at 5 km	Output
18	So2-km10(t)	So2 gas concentration at 10 km	Output
19	So2-km15(t)	So2 gas concentration at 15 km	Output
20	So2-km20(t)	So2 gas concentration at 20 km	Output
21	CO2-km5(t)	CO2 concentration at 5 km	Output
22	CO2-km10(t)	CO2 concentration at 10 km	Output
23	CO2-km15(t)	CO2 concentration at 15 km	Output
24	CO2-km20(t)	CO2 concentration at 20 km	Output
25	Co-km5(t)	Co gas concentration at 5 km	Output
26	Co-km10(t)	Co concentration at 10 km	Output
27	Co-km15(t)	Co gas concentration at 15 km	Output
28	Co-km20(t)	Co concentration at 20 km	Output
29	Nox-km5(t)	Nox gas concentration at 5 km	Output
30	Nox-km10(t)	Nox gas concentration at 10 km	Output
31	Nox-km15(t)	Nox gas concentration at 15 km	Output
32	Nox-km20(t)	Nox gas concentration at 20 km	Output

Accuracy Assessment Indicators of SVM Model

In order to compare the concentration of gas observed and estimated with SVM model in training stages and tests, the statistical indices of the *Nash*–Sutcliffe model

$$NSE = 1 - \left[\frac{\sum_i^n (S_i - O_i)^2}{\sum_i^n (O_i - \bar{O})^2}\right] \quad (10)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2} \quad (11)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|S_i - O_i| \quad (12)$$

3. Results

Selection of effective input variables with PMI algorithm

Table 4 shows results of the PMI algorithm for SO_2 gas concentrations at 5, 10, 15, and 20 km for 5 effective variables, respectively.

Table 4. Results of PMI algorithm for So2 and Co2 concentration at distances of 5, 10, 15, and 20 km

Iteration	Variable	I(x;y)	MC-I*(95)	MC-I*(99)	Hampel	km
0	logSo2-km0(t)	3.155	-8.28E+06	-1.08E+07	39.7	
1	tmax(t)	0.219	-8.28E+06	-1.08E+07	1.2	5
2	dd_max(t)	0.149	-8.28E+06	-1.08E+07	2.3	
3	p0min(t)	0.125	-8.28E+06	-1.08E+07	1.2	
4	logVm(t)	0.154	-8.28E+06	-1.08E+07	1.5	
0	logSo2-km0(t)	3.180	-8.28E+06	-1.08E+07	41.15	
1	logVm(t)	0.232	-8.28E+06	-1.08E+07	1.22	10
0	logSo2-km0(t)	3.180	-8.28E+06	-1.08E+07	41.15	
1	logVm(t)	0.232	-8.28E+06	-1.08E+07	1.22	15
0	logSo_2-km0(t)	4.017	-8.28E+06	-1.08E+07	54.56	
1	logVm(t)	0.219	-8.28E+06	-1.08E+07	0.59	20
0	Co_2-km0(t)	11.982	-8.28E+06	-1.08E+07	244.2	
1	logVm(t)	0.388	-8.28E+06	-1.08E+07	1.6	15

According to table 4 and considering Hampel's criteria, the input variables affecting the estimation of SO_2 gas concentration at 5, 10, 15, and 20 km are SO_2 gas concentrations at the chimney site.

At first, the data were trained and tested on the support vector machine. Since neural networks work with data at a distance of (1, 0), in this process, data has been normalized from the beginning.

Determining the best structure of the SVM model in estimating pollutant concentrations at different distances from the power plant

After preparing the meteorological variables on a weekly basis to construct a potential set of input variables, the In order to calibrate (train) and validate (test) the SVM model and determine the optimal structure of the SVM model in estimating the concentration of pollutant gases: CO, CO_2, and NO_X at intervals of 5, 10, 15, and 20 km from the chimneys of Shazand thermal power plant, different kernel functions and different values of the kernel function parameters were considered. The values total number of training patterns were equal to 48 weeks. Of this total number of training models, 70% were used for training (calibration) for the SVM model and 30% were used for testing (validation) for the SVM model. In other words, (from December 2018 to the first week of September 2019) was determined as a time for training the model, and a period (from the second week of September 2019 to December 2019) was determined as a time for testing the model. The first set of data, so called training data, is used to determine network weights and biases. The second set of data, so called test data, is used to determine the accuracy of the model, or in other words, the model test.

of statistical indices and the optimal values of the parameters related to the best kernel function in the SVM model for estimating the SO_2gas concentration at different distances from the power plant chimney for the training and test periods for the SVM model are shown in table 5.

Table 5. Statistical indicators related to the best kernel function in L-SVM mode for estimating gas concentration of SO_2 at different distances from the power plant chimney for training and testing periods of the model

Distance (km)	Kernel function	Train			Test			Optimum values of parameters			
		RMSE	MAE	NSE	RMSE	MAE	NSE	ϒ	σ^2	d	c
5	Polynomial	3.72	3.28	0.953	2.88	2.33	0.994	50	-	3	1
10	Polynomial	3.11	2.73	0.953	2.39	1.93	0.994	150	-	3	1
15	Polynomial	2.90	2.55	0.953	2.32	1.89	0.994	10	-	2	1
20	Polynomial	2.49	2.19	0.953	1.917	1.55	0.994	50	-	3	1

According to table 5, the polynomial kernel function in the test period had a higher Nash-Sutcliffe coefficient and the root of the mean squared error and the mean of the absolute error less than the two basic radial and linear functions. In other words, the polynomial kernel function has better performance than the radial and linear basis function in estimating SO_2 gas concentration at different distances.

Figures 1 and 2 show a comparison of the changes in the gas concentration of SO_2 observed and estimated by the SVM model at different distances with the polynomial kernel function for the training and testing periods of the model. According to Figures 1 and 2, it can be seen that the SVM model is well trained in estimating gas concentration at different distances.

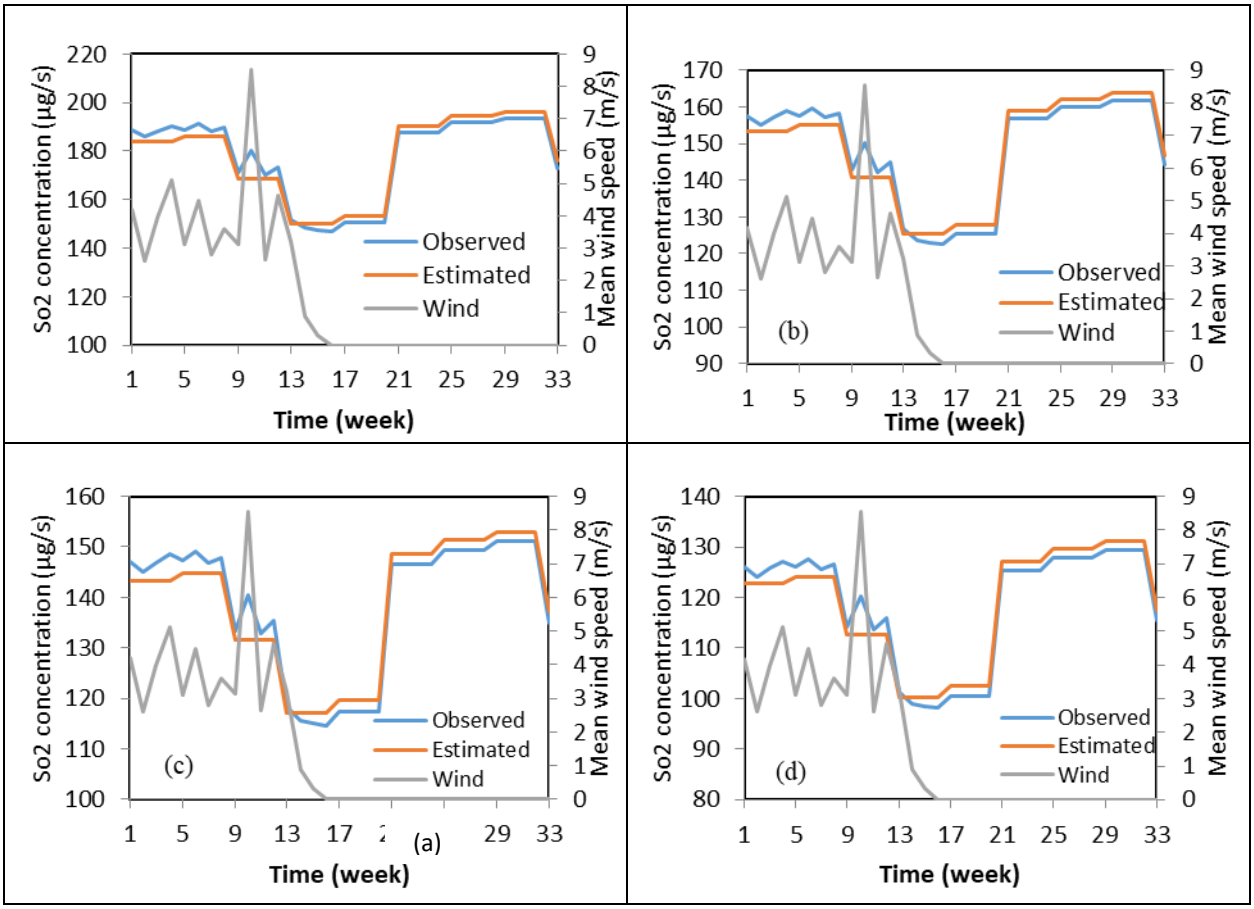


Fig.1- Comparison of changes in the gas concentration of SO_2 observed and estimated by the SVM model at distances of 5 km (a), 10 km (b), 15 km (c), and 20 km (d) for the training period

Due to the changes in different gas concentrations shown in Figures 1 to 8, it can be seen that at times when the wind speed has increased, the gas concentration has increased at different distances, but this increase in gas concentration compared to the effect Concentration of gas released from the chimney of the power plant, according to the results of PMI algorithm and comparison of time changes of different concentrations

of gas at different distances with the concentration of exhaust gas from the chimney of the power plant, it can be found that the concentration of exhaust gas from the chimney At different distances.

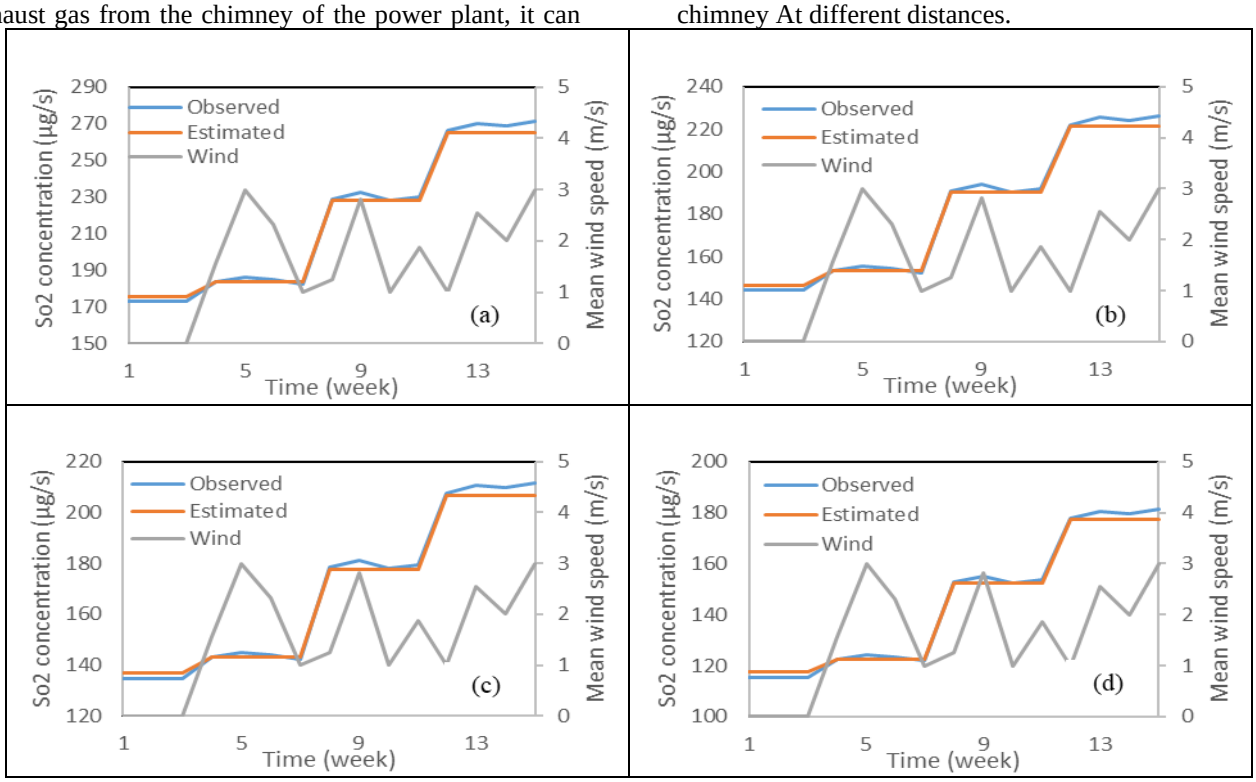


Fig.2- Comparison of SO_2 gas concentration changes observed and estimated by the SVM model at distances of 5 km (a), 10 km (b), 15 km (c), and 20 km (d) for the test period

Determining the best structure of the SVM model in estimating the concentration of SO_2 gas at different distances from the power plant chimney

In order to train and test the SVM model and determine the optimal structure of the SVM model in estimating the concentration of SO_2 gas at intervals of 5, 10, 15, and 20 km relative to the chimneys of Shazand thermal power plant, different kernel functions and different values of kernel function parameters were considered. The values of statistical indices and the optimal values of the parameters related to the best kernel function in the SVM model for estimating the gas concentration of SO_2 at different distances from the power plant chimney for the training and testing periods of the SVM model are shown in table 5. According to table 5, the linear kernel function in the test period had a higher Nash-Sutcliffe coefficient and a less root mean square error and mean absolute error were than the two radial and polynomial base functions. In other words, the linear kernel function is more accurate than the radial and polynomial base function in estimating gas concentration at different distances. Also, table 6 shows the results of the PMI algorithm for CO_2 concentration at km 5, 10, 15, and 20, respectively.

Table 6. Statistical indicators related to the best kernel function in the L-SVM model for estimating the gas concentration of CO_2 at different distances from the power plant chimney for training and testing periods of the model

Distance (km)	Kernel function	Train			Test			Optimum values of parameters			
		RMSE	MAE	NSE	RMSE	MAE	NSE	ϒ	σ^2	d	c
5	Linear	187.09	128.02	0.955	1134.50	833.13	0.985	5	-	-	-
10	Linear	173.59	122.37	0.95	1285.30	934.52	0.98	10	-	-	-
1	Linear	162.86	114.44	0.952	1222.10	888.76	0.974	24	-	-	-
20	Linear	155.06	109.89	0.952	1121.50	815.16	0.976	5	-	-	-

Figures 3 and 4 show a comparison of the gas concentration changes of SO_2 observed and estimated by the SVM model at different distances with the linear kernel function for the training and test time periods of

the model. According to Figures 3 and 4, it can be seen that the SVM model has well estimated the gas concentration of SO_2 different distances.

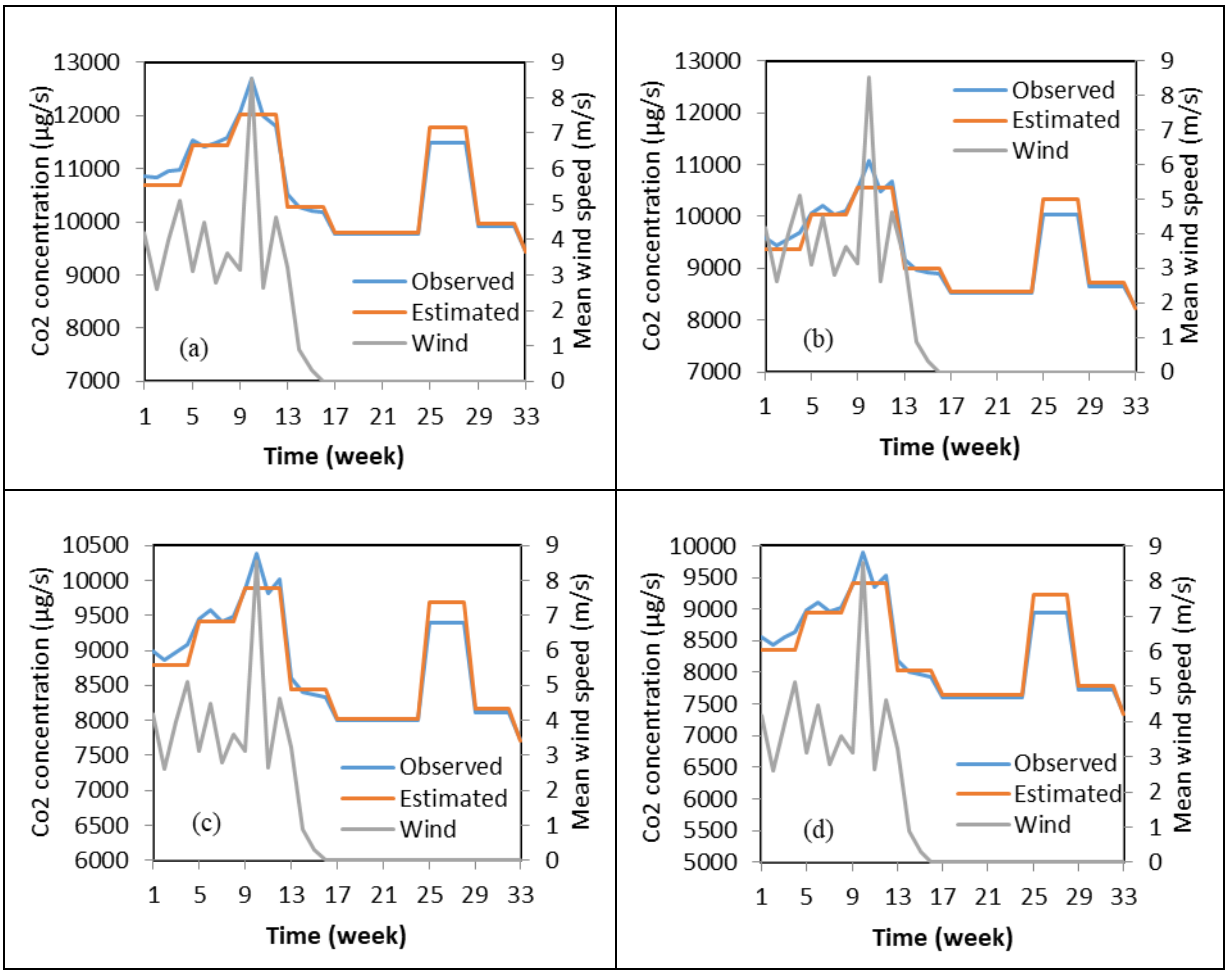
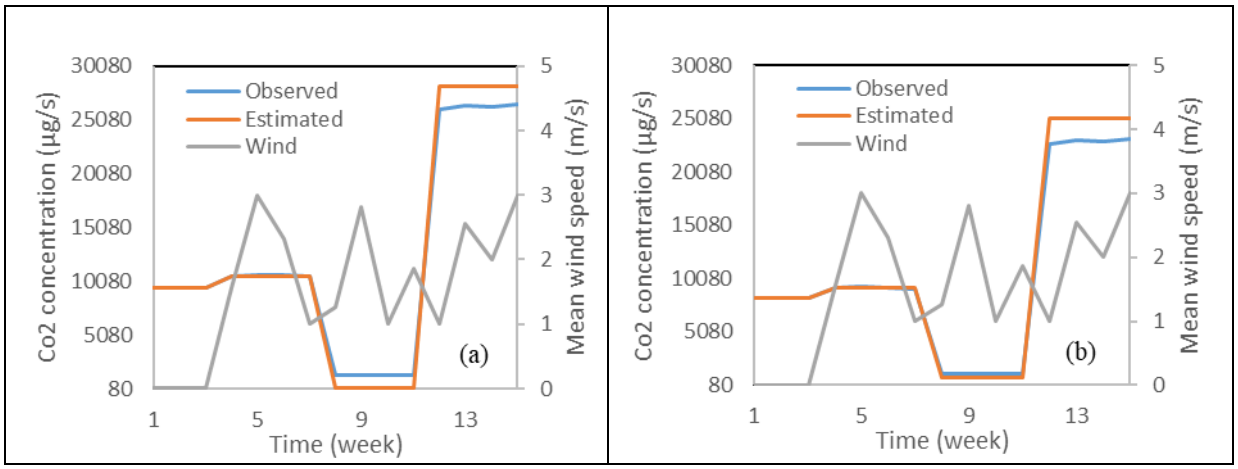


Fig.3- Comparison of changes in the concentration of CO_2 gas observed and estimated by the SVM model at distances of 5 km (a), 10 km (b), 15 km (c), and 20 km (d) for the training period



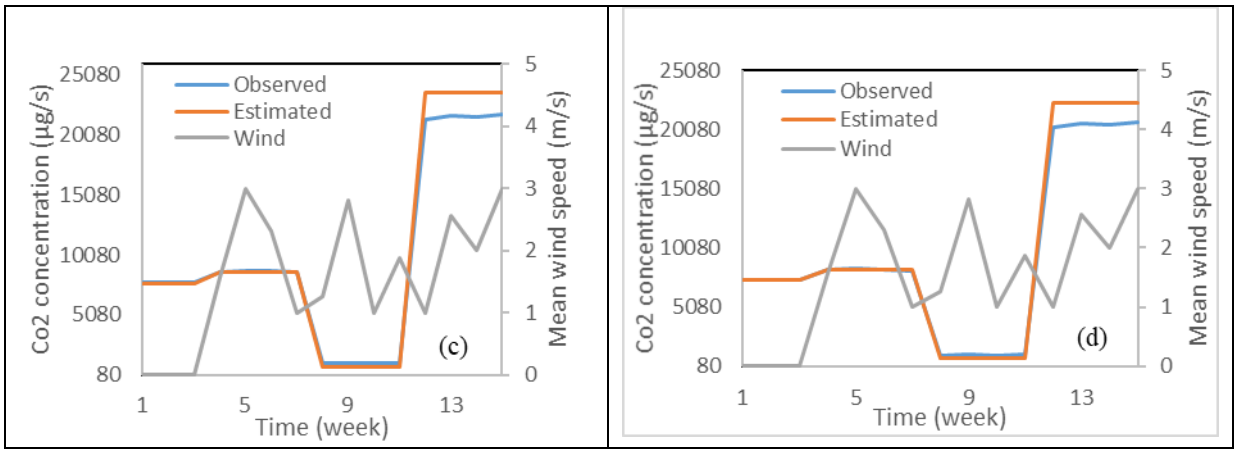


Fig.4- Comparison of changes in the concentration of CO_2 gas observed and estimated by the SVM model at distances of 5 km (a), 10 km (b), 15 km (c), and 20 km (d) for the test period

Determination of the best structure of the SVM model in estimating the concentration of CO pollutant different distances from the chimney of the power plant

In order to train and test the SVM model and to determine the optimal structure of the SVM model in the estimation of CO gas concentration, at distances of 5, 10, 15, and 20 km from the chimneys of Shazand thermal power plant, different kernel functions and different values of kernel function parameters were considered. The values of statistical indicators and optimum values of parameters related to the best kernel function in the SVM model to estimate CO gas concentration at different distances from the power plant chimney for training periods and SVM model test are shown in table 10. According to table 10, the linear kernel function in the time period of the test had a higher Nash-Sutcliffe coefficient and less mean square error root and less mean absolute value of error than the two radial and polynomial base functions. In other words, the linear kernel function is more accurate than the radial and polynomial base function in estimating the gas concentration of CO at different distances.

Table 7. Statistical indicators related to the best kernel function in the L-SVM model for estimating the gas concentration of CO different distances from the power plant chimney for training and testing periods of the model

Distance (km)	Kernel function	Train			Test			Optimum values of parameters			
		RMSE	MAE	NSE	RMSE	MAE	NSE	ϒ	σ^2	d	c
5	Linear	0.148	0.107	0.9998	0.095	0.082	0.9998	5	-	-	-
10	Linear	0.156	0.125	0.9998	0.104	0.085	0.9998	10	-	-	-
15	Linear	0.174	0.132	0.9998	0.116	0.081	0.9998	5	-	-	-
20	Linear	0.165	0.121	0.9998	0.102	0.071	0.9998	15	-	-	-

Figures 5 and 6 show a comparison of changes in the gas concentration of CO_2 observed and estimated by the SVM model at different distances with the linear kernel function for the training and test periods of the model. According to Figures 5 and 6, it can be seen that the SVM model has well estimated the gas concentration of CO at different distances.

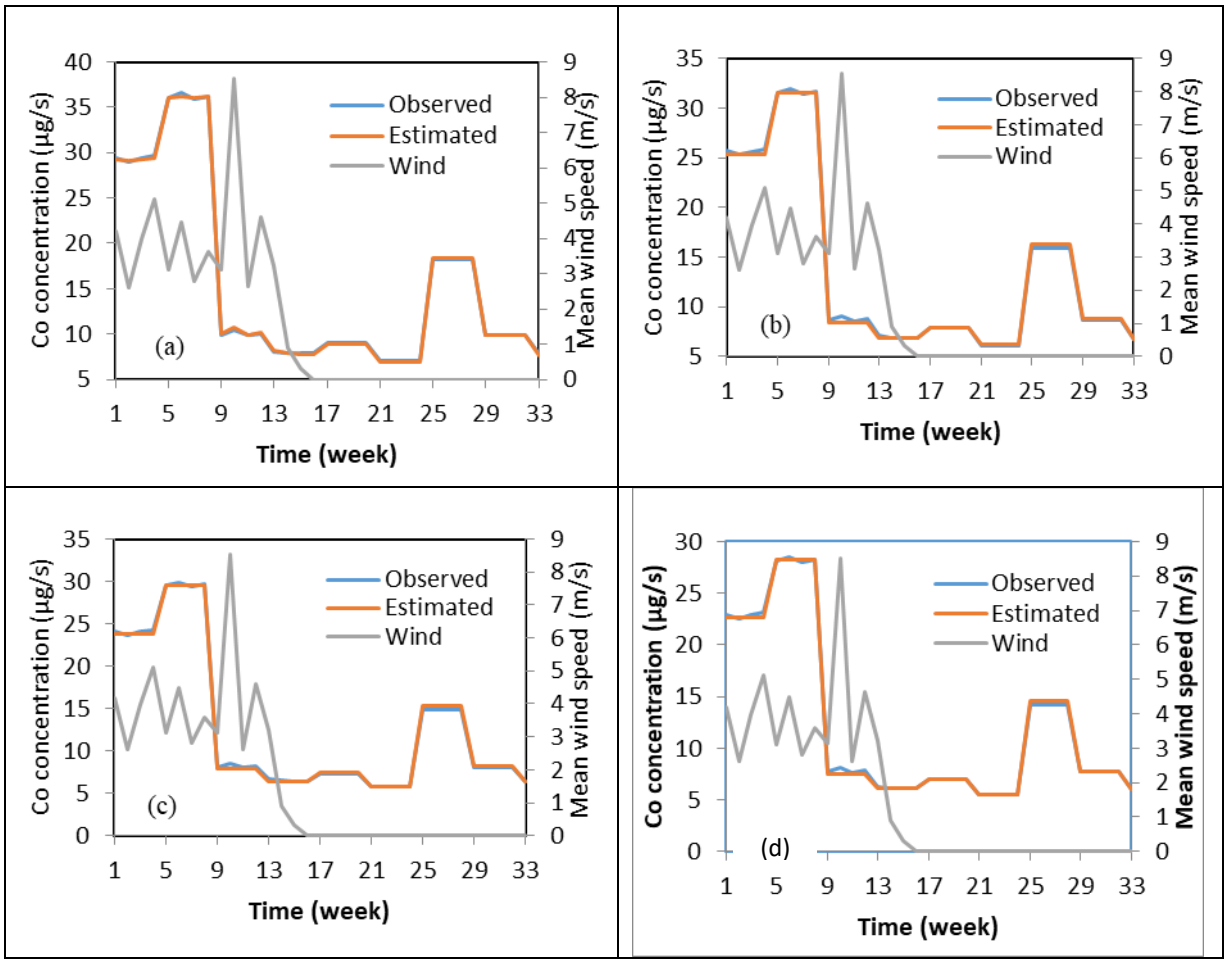
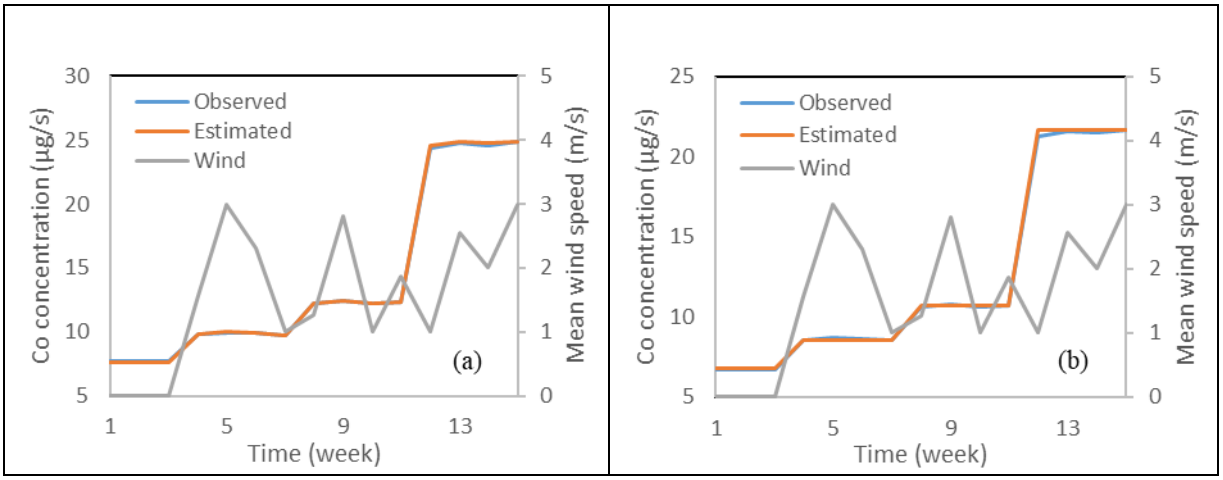


Fig.5- Comparison of changes in the concentration of CO gas observed and estimated by the SVM model at distances of 5 km (a), 10 km (b), 15 km (c), and 20 km (d) for the training period



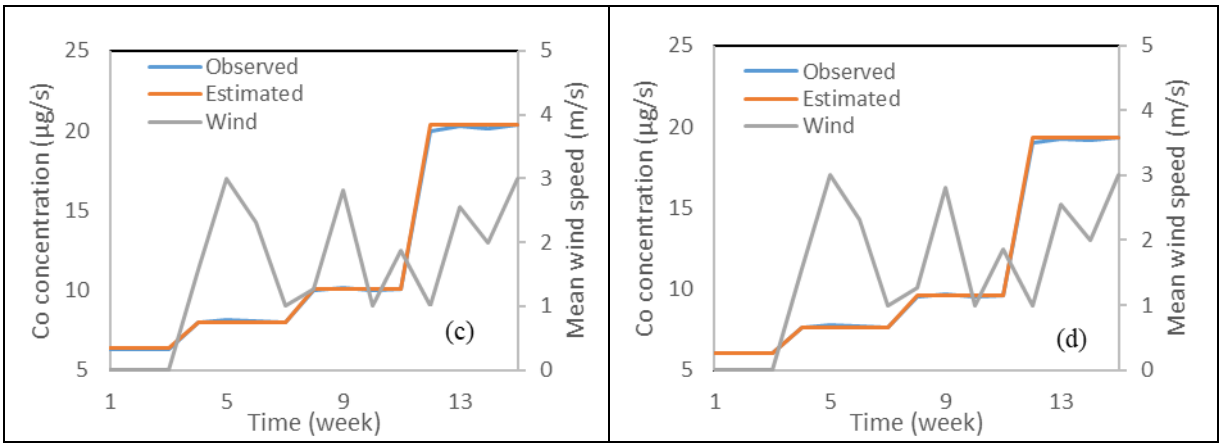


Fig.6- Comparison of changes in the concentration of CO gas observed and estimated by the SVM model at distances of 5 km (a), 10 km (b), 15 km (c), and 20 km (d) for the testing period

Determination of the best structure of the SVM model in estimating the concentration of NO_X pollutant at different distances

In order to train and test the SVM model and determine the optimal structure of the SVM model in estimating NO_X gas concentration at distances of 5, 10, 15, and 20 km relative to the chimneys of Shazand thermal power plant, different kernel functions and different values of kernel function parameters were considered. The values of statistical indices and the optimal values of the parameters related to the best kernel function in the SVM model for estimating the NO_X gas concentration at different distances from the power plant chimney for the training and test periods of the SVM model are shown in table 8. According to table 8, the polynomial kernel function in the test time period had a higher Nash-Sutcliffe coefficient and lower mean square error root, and lower mean absolute value of error than the two radial and linear base functions. In other words, the polynomial kernel function is more accurate than radial and linear base functions in estimating NO_X gas concentration at different distances. The results of the PMI algorithm for NO_X gas concentrations at 5, 10, 15, and 20 km, respectively, are shown in table 8.

Table 8. Statistical indicators related to the best kernel function in the SVM model for estimating NO_X gas concentration at different distances from power plant chimney for training and testing periods of the model

Distance (km)	Kernel function	train			test			Optimum values of parameters			
		RMSE	MAE	NSE	RMSE	MAE	NSE	ϒ	σ^2	d	c
5	Polynomial	0.408	0.286	0.974	3.118	2.305	0.963	20	-	3	1
10	Polynomial	0.346	0.242	0.974	2.628	1.943	0.963	100	-	4	1
15	Polynomial	0.315	0.221	0.974	2.486	1.836	0.961	5	-	3	1
20	Polynomial	0.295	0.206	0.9743	2.245	1.659	0.963	110	-	3	1

Also, according to table 8 and considering criteria, the input variables affecting the estimation of NO_X gas concentration at 5, 10, 15, and 20 km are the concentration of NO_X gas in the chimney site.

Figures 7 and 8 show a comparison of the changes in NO_X gas concentration observed and estimated by the SVM model at different distances with the linear kernel function for the training and test periods of the model. According to Figures 7 and 8, it can be seen that the NO_X gas concentration model has been well estimated at different distances.

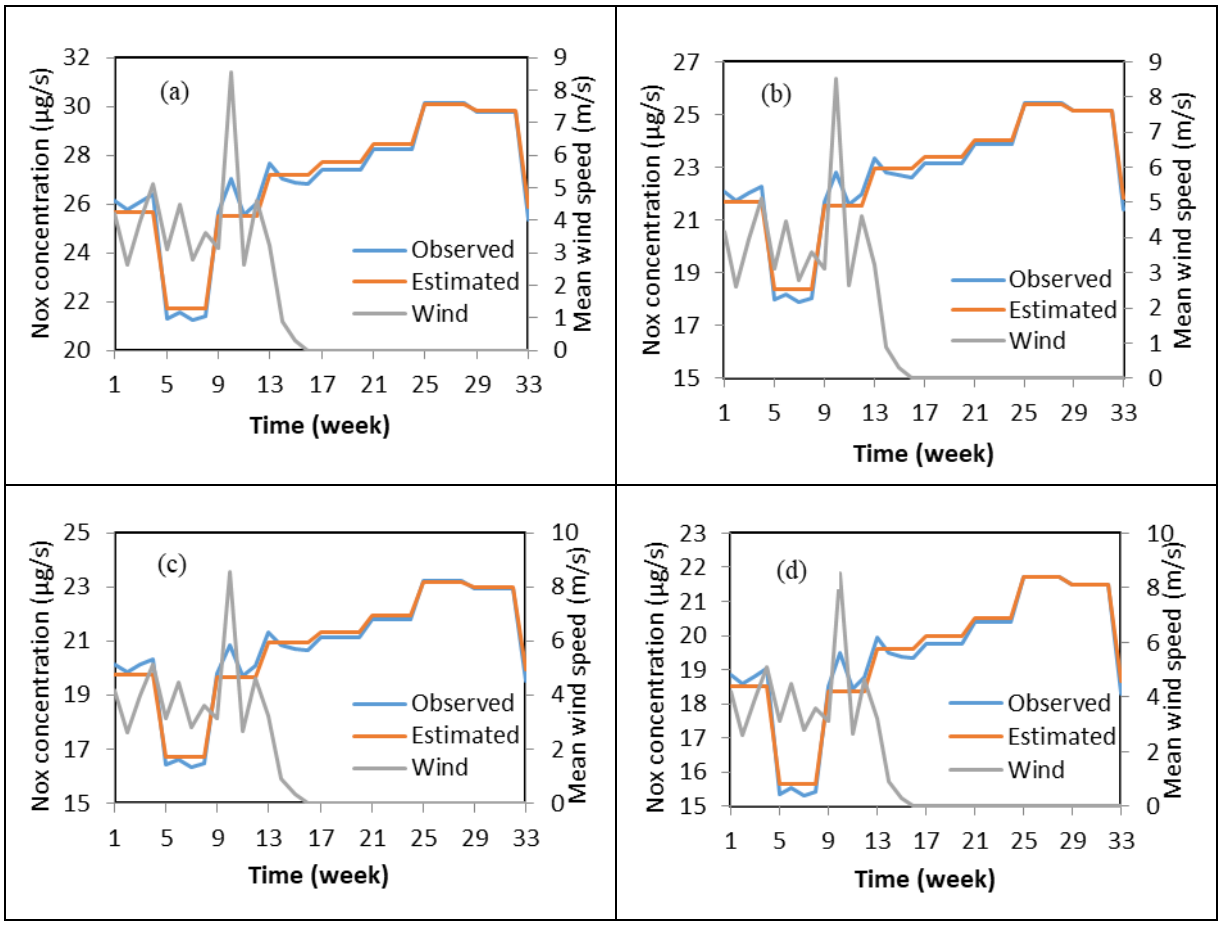
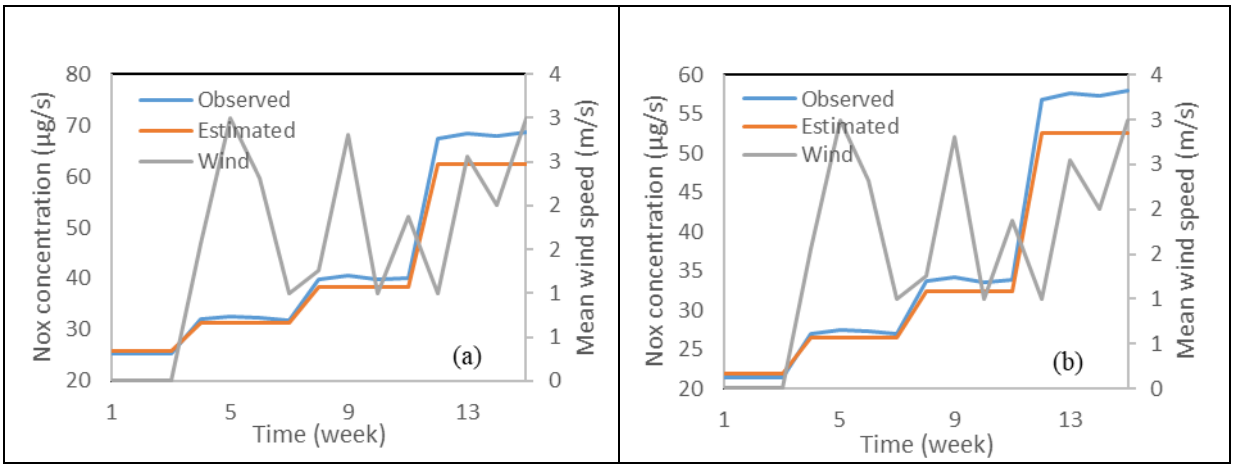


Fig.7- Comparison of changes in the concentration of NO_X gas observed and estimated by the SVM model at distances of 5 km (a), 10 km (b), 15 km (c), and 20 km (d) for the training period



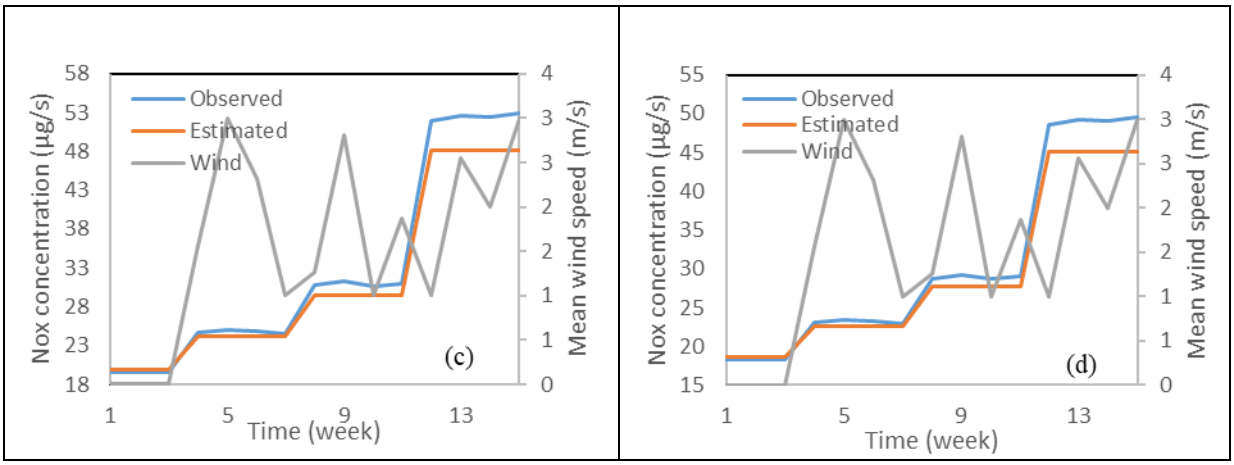


Fig.8- Comparison of changes in the concentration of NO_X gas observed and estimated by the SVM model at distances of 5 km (a), 10 km (b), 15 km (c), and 20 km (d) for the testing time period

According to the graph of changes in different gas concentrations in Figures 1 to 8, it can be seen that at times when the wind speed has increased, the gas concentration has increased at different distances, but this increase in gas concentration is small compared to the effect of the gas concentration coming out of the power plant chimney. According to the results of PMI algorithm and comparing the temporal changes of different gas concentrations at different distances with the concentration of exhaust gas from the power plant chimney, it can be seen that the concentration of exhaust gas from the power plant chimney has the greatest effect on changes in gas concentrations at different distances.

Evaluation of concentration emission pattern in estimating air pollutants

In order to assess the environmental concentrations of CO_2 and SO_2, NO_X, and CO pollutants estimated by the SVM model at different distances from the plant chimney, the concentrations of pollutants with EPA International Ambient Air Quality Standard for different groups, including workers, technical staff, office staff, and adjacent areas. The concentrations of CO_2, SO_2, NO_X, and CO pollutants are shown in table 9.

Table 9. Results of gas analysis and pollution parameters of Shazand power plant chimneys and its comparison with EPA International Ambient Air Quality Standard

Factors	Chemical Compounds	Concentration µg/m3	EPA International Ambient Air Quality Standard (µg/m3)
Workers	CO	1900	4000
	NO_2	450	200
	SO_2	80	196
Technical Area Staff	CO	1250	4000
	NO_2	375	200
	SO_2	206	196
Administrative Staff	CO	1250	4000
	NO_2	215	200
	SO_2	70	196
Adjacent Areas	CO	1500	4000
	NO_2	55	200
	SO_2	70	196

The EPA estimated for a variety of factors, as well as the International Ambient Air Quality Standard of the risk index assessed for workers on the site show that they are at serious risk for disease and complications from air pollutants. These people are exposed to direct pollutants from the chimneys of the power plant while working in

the open. Technical staff and workers are also prone to the non-cancerous effects of these contaminants. Workers are in a similar situation to technical staff while at the plant, but the frequency of contacts is different. Office workers, even though they are present in the buildings of the power plant during office hours, are indirectly exposed to pollutants from the power plants' chimneys in the closed environment. Also, the hazard index indicates that the risk of inhaling air pollutants around the plant is low. Carbon monoxide pollutants with a share of 62.5% and nitrogen dioxide with a share of 37.5% have the greatest role in creating hazards for power plant employees, both in the administrative and technical areas. Also, for workers, nitrogen dioxide pollutants with a hazard index of 64% and carbon monoxide with a hazard index of 36%, have the largest share in causing non-cancerous complications.

4. Discussion

Air quality is essential for human health and the environment. The spread of the consumption of electricity and fossil fuels in power plants has caused irreparable damage to the lives of humans and other living organisms. The extent of distribution of pollutants depends on the type and quality of fuels used. Consumption of fossil fuels and production of toxic gases such as nitrogen oxides NO_X, sulfur oxide (Sox) in Shazand thermal power plant has a destructive effect on the local, regional level. Also, the study of the distribution of concentrations of pollutants, especially NO_X simulated in the coldest period of the year shows the poor air quality of this pollutant to convert NO to NO_2 in the worst conditions, which is the reason for the areas close to the power plant. This can be attributed to the power plant's use of fuel oil in this season and the fact that none of its chimneys are equipped with systems to reduce this emission of pollutants. Therefore, the results of statistical analysis used for the SVM model in Shazand power plant also show a reasonable agreement between the data predicted by the model and the observational data at separate receptor distances around the power plant. Therefore, the PMI algorithm and support vector machine model with kernel functions can be used as a suitable scientific tool to analyze the control and policy strategies to reduce and prevent air pollution. In order to estimate the gas concentration, the values of meteorological variables of Shazand synoptic station were considered as inputs in the potential set of input variables. Also, the concentration of each gas in the chimney (km 0) was considered as input. The concentration of each gas at distances of 5, 10, 15, and 20 km was considered as output variables. After preparing the meteorological variables on a weekly basis to construct a potential set of input variables, the total number of training models was equal to 48 weeks. Of these total training models, 70% of the models were used for training (calibration) of the SVM model and 30% of the models were used for testing (validation) of the SVM model. In this study, the average concentration of pollutants in relation to the distance from the source of pollution (power plant) was measured, and changes in concentration relative to the distance from the power plant were identified to better show the contribution of power plants in limited pollution around the power plant to Arak. Among air pollutants, maximum gas concentration was related to CO2 with 2811.63 $\mu g/m3$ at a distance of 5 km from the power plant chimney in autumn and December and at weeks of 45-48 with an average wind speed of 2.14 m/s and the lowest gas concentration was related to Co with 5.5 $\mu g/m3$ at a distance of 20 km from the power plant in spring (June) in weeks 21 to 24 with an average wind speed of 1.61 m/s. Also, in a nearly similar study by K.P. Lin et al., 2011, in order to predict the concentration of air pollutants with the support of logarithm vector and safety algorithms in Taiwan (Lin, et al. 2011), two types of gaseous pollutants were analyzed. While in this article, we have tried to interpret the four pollutants along with comparing the changes in the concentrations of pollutants (CO_2 and SO_2, NO_X, and CO) observed and estimated at different distances from the source of the pollutant.

5. Conclusions

Accurate estimation of pollutant emission concentrations is important to effective monitoring. The purpose of this study was to present a plan to predict air pollution due to the concentration of pollutants in the Shazand thermal power plant dynamically. This forecast can be done on a weekly and monthly basis based on current data on the concentration of meteorological pollutants and spatial data. In this study, the SVM model was used to estimate the concentration of gaseous pollutants emitted from the Shazand thermal power plant at different distances from the power plant. The PMI algorithm was used to select the effective input variables in the SVM model. The results of using a PMI algorithm show that the input variables effective in estimating the concentration of pollutants (CO_2 and SO_2, NO_X, and CO) at different distances, taking into account the Hampel criterion, is the concentration of the same gas at the chimney of the power plant. The results show that the best kernel function of the SVM model for estimating the concentrations of SO_2, and NO_X pollutants at different distances is the polynomial kernel function. The linear kernel function is also the best kernel function for estimating the concentrations of CO_2 and CO pollutants at different distances. Comparison of changes in pollutant concentrations (CO_2, SO_2, NO_X, and CO) observed and estimated by the SVM model at different distances from the power plant, shows that the SVM model has good

accuracy and performance in estimating pollutant concentrations. Also, the selection of effective input variables in the SVM model with the PMI algorithm increases the accuracy and development of the model and reduces the time required for modeling. During the study of air pollutants, it was found that when fuel oil is burning, the power plant is one of the important sources of production in the region. At the same time, there is always the problem of production and propagation from the power plant to the environment. Proposing the use of the adsorption method of the porous substrate of activated carbon is the simplest method of removal. In Shazand power plant, the most logical proposal could be changing the type of fuel, from fuel oil to natural gas, which is being done. Also recommending the SCR method is a very good method of complete removal that has been used in many power plants and large industries in the world in recent years. Therefore, the reasons for the superiority in the proposed model can be summarized as follows: The use of the PMI partial cross-algorithm method increases the accuracy and improves the performance of the backup vector machine. Since the way data processing is significantly affecting the performance of the model, the use of other pre-processing methods to increase the model estimates (SVM) is recommended for future research. Integrating the proposed model with GIS and collecting more data onto real-time can make this model more useful in terms of application.

Finally, for the analysis of some uncertain factors, fuzzy set theory can be included in the support vector machine model. Experimental results showed that the results obtained in this study indicate the promising potential of the proposed model in estimating the concentration of pollutants, the advantage of which is to avoid making false errors in the actual estimation of emissions of these pollutants under maximum load power plant production conditions.

The results of this study show the promising potential for the proposed model in estimating the concentration of pollutants, the advantage of which is to avoid making false errors in the actual estimation of emissions of these pollutants under maximum load power plant production conditions.

6. Conflict of interest

The authors declare that they have no conflict of interest.

7. Additional Information And Declarations

Funding

There was no funding for this study.

Grant Disclosures

There was no grant funder for this study.

Competing Interests

The author declare there is no competing interests, regarding the publication of this manuscript

Author Contributions

Ghodratollah Siahpour, Seyed Ali Jozi, Neda Orak, Hossein Fathian, and Solmaz Dashti: Prposed the plan, conceived the experiments, analyzed the data, authored or revised drafts of the paper, approved the final draft.

Ethics Statement

Vice Chancellor for Research and Technology, Islamic Azad University, Ahvaz Branch

References

Adams, Derrick, et al. (2020) Prediction of SOx–NOx emission from a coal-fired CFB power plant with machine learning: Plant data learned by deep neural network and least square support vector machine. J Clean Prod 270:122310.

Akaike, H. (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716-723.

Alexiadis, MC, et al. (1998) Short-term forecasting of wind speed and related electrical power. Sol Energy 63(1):61-68.

Arain, MA, et al. (2007) The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. Atmos Environ 41(16):3453-3464.

Awasthi, Seema, Mukesh Khare, and Prashant Gargava 2006 General plume dispersion model (GPDM) for point source emission. Environ Model Assess 11(3):267-276.

Conti, John, et al. (2016) International energy outlook 2016 with projections to 2040. USDOE Energy Information Administration (EIA), Washington, DC (United States).

Cover, Thomas M (1991) J. A. Thomas, Elements oflnformation Theory: New York: Wiley.

David, Florence Nightingale (1938)Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples: Cambridge University Press.

Davies, Laurie, and Ursula Gather (1993) The identification of multiple outliers. J Am Stat Assoc 88(423):782-792.

Goebel, Bernhard, et al. (2005) An approximation to the distribution of finite sample size mutual information estimates. IEEE International Conference on Communications, 2005. ICC 2005. 2005, 2005. Vol. 2, pp. 1102-1106. IEEE.

Hamel, Lutz H (2011) Knowledge discovery with support vector machines. Volume 3: John Wiley & Sons.

Hasenfratz, David, et al. (2012) Participatory air pollution monitoring using smartphones. Mobile Sensing 1:1-5.

Hosseinnezhad, Vahid, and Ebrahim Babaei (2013) Economic load dispatch using θ-PSO. Int J Electr Power Energy Syst 49:160-169.

Kariniotakis, GN, GS Stavrakakis, and EF Nogaret (1996) Wind power forecasting using advanced neural networks models. IEEE Trans Energy Convers 11(4):762-767.

Lin, Kuo-Ping, Ping-Feng Pai, and Shun-Ling Yang (2011) Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms. APPL MATH COMPUT 217(12):5318-5327.

Lippmann, Morton, et al. (2003) The US Environmental Protection Agency Particulate Matter Health Effects Research Centers Program: a midcourse report of status, progress, and plans. Environ Health Perspect 111(8):1074-1092.

Liu, Xiyu, Hong Liu, and Huichuan Duan (2007) Particle swarm optimization based on dynamic niche technology with applications to conceptual design. Adv Eng Softw 38(10):668-676.

Lu, Wei-Zhen, et al. (2004) Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong. Environ Res 96(1):79-87.

May, Robert J, et al. (2006) Critical values of a kernel density-based mutual information estimator. The 2006 IEEE International Joint Conference on Neural Network Proceedings, 2006, pp. 4898-4903. IEEE.

May, Robert J, et al. (2008) Non-linear variable selection for artificial neural networks using partial mutual information. Environ Model Softw 23(10-11):1312-1326.

Pearson, Ronald K (2002) Outliers in process modeling and identification. IEEE Transactions on control systems technology 10(1):55-63.

Perera, Frederica (2018) Pollution from fossil-fuel combustion is the leading environmental threat to global pediatric health and equity: solutions exist. Int J Environ Res Public Health 15(1):16.

Pope III, C Arden (2007) Mortality effects of longer term exposures to fine particulate air pollution: review of recent epidemiological evidence. Inhal Toxicol 19(sup1):33-38.

Reikard, Gordon (2012) Forecasting volcanic air pollution in Hawaii: tests of time series models. Atmos Environ 60:593-600.

Shannon, Claude E (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379-423.

Wahid, H, et al. (2013) Neural network-based meta-modelling approach for estimating spatial distribution of air pollutant levels. APPL SOFT COMPUT 13(10):4087-4096.

Zheng, Yu, Furui Liu, and Hsun-Ping Hsieh (2013) U-air: When urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 1436-1444.

Zhou, Junyi, Jing Shi, and Gong Li (2011) Fine tuning support vector machines for short-term wind speed forecasting. ENERG CONVERS MANAGE 52(4):1990-1998.