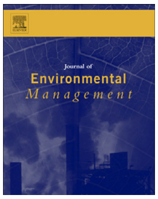
Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman



Research article

Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm

Wei Sun, Jingyi Sun*

Department of Business Administration, North China Electric Power University, Baoding 071000, China

ARTICLE INFO

Article history:
Received 2 March 2016
Received in revised form
1 December 2016
Accepted 5 December 2016
Available online 15 December 2016

Keywords:
$PM_{2.5}$
Concentration forecasting
PCA
LSSVM
CS

ABSTRACT

Increased attention has been paid to $PM_{2.5}$ pollution in China. Due to its detrimental effects on environment and health, it is important to establish a $PM_{2.5}$ concentration forecasting model with high precision for its monitoring and controlling. This paper presents a novel hybrid model based on principal component analysis (PCA) and least squares support vector machine (LSSVM) optimized by cuckoo search (CS). First PCA is adopted to extract original features and reduce dimension for input selection. Then LSSVM is applied to predict the daily $PM_{2.5}$ concentration. The parameters in LSSVM are fine-tuned by CS to improve its generalization. An experiment study reveals that the proposed approach outperforms a single LSSVM model with default parameters and a general regression neural network (GRNN) model in $PM_{2.5}$ concentration prediction. Therefore the established model presents the potential to be applied to air quality forecasting systems.

1. Introduction

In recent years, with the rapid development of industrialization in China, most major cities, especially in Beijing, Tianjin and Hebei province (Jing-Jin-Ji area) have suffered from serious air pollution (Wang and Zhang, 2009; Zheng et al., 2015; Zhang et al., 2016). $PM_{2.5}$, as the dominant pollutant, has attracted wide attention. It refers to the particulate matter whose aerodynamic diameter is 2.5 μm or less. $PM_{2.5}$ is made of toxic and hazardous substances with high activity and shows long residence time and far transportation distance in the atmosphere, thus it has been reported that exposure to high concentrations of $PM_{2.5}$ brings about an increase in cardiovascular and lung diseases (You et al., 2015; Sun et al., 2013).

To enhance air quality management, daily air quality index (AQI) has been published every day by the Chinese Ministry of Environment Protection since the year of 2012. It is noteworthy that this index adds $PM_{2.5}$ concentration which leads to haze pollution as one of the air quality monitoring indicators. China has set national ambient air quality standards for $PM_{2.5}$ based on annual average concentration (35 μg/m3) and 24-h average concentration (75 μg/m3) (State Bureau of Environment Protection, 2012). The publication and implementation of this standard has played an important role in monitoring air quality and improving living environment.

The adverse effects of $PM_{2.5}$ make it an urgent need for researchers to simulate and forecast its concentration. A series of real-time air quality forecasting (RT-AQF) models have been proposed (Zhang et al., 2012a). The accuracy and skills are gradually improved with the scientific advances and major numerical, statistical and computational techniques (Zhang et al., 2012b). Wang et al. (2013) employed a sequential factorial analysis approach for supporting regional air quality forecasting. Vlachogianni et al. (2011) outlined a stepwise multiple linear regression model which used hourly concentrations of pollutants together with meteorological parameters as predictors to forecast PM_{10} and NO_x concentrations in Athens and Heisinki. Donnelly et al. (2015) put up with a multiple linear regression based model for real time air quality prediction. The results showed that the proposed model could meet the satisfaction of air quality forecasting with high accuracy and computational efficiency. Jian et al. (2012) aimed at improving the prediction accuracy of submicron particle concentration under busy traffic conditions by using an auto-regression integrated moving average (ARIMA) model with the consideration of meteorological factors. These regression analysis and time series may be weak in predicting the extreme points and difficult to deal with influential factors involved in the forecasting. Some mathematical optimization models have also been introduced into air

* Corresponding author. Department of Business Administration, North China Electric Power University, 689 Huadian Road, Baoding 071000, China.
E-mail addresses: bdsunwei@126.com (W. Sun), sunjingyi0224@126.com (J. Sun).

http://dx.doi.org/10.1016/j.jenvman.2016.12.011
0301-4797/© 2016 Elsevier Ltd. All rights reserved.

quality management. Wang and Huang (2013a) proposed a coupled factorial-analysis-based interval programming approach to tackle uncertainties that exist in the objective function and constraints. The results can help decision makers identify desired pollution mitigation strategies with minimized total cost and maximized environmental efficiency. To further address dual uncertainties, an interactive fuzzy boundary interval programming was applied to make the tradeoff between the operating cost and the constraint violation risk in air quality management (Wang and Huang, 2013b).

With the development of artificial intelligence, intelligent algorithms have gradually been popular for air quality prediction. Gennaro et al. (2013) presented an artificial neural network (ANN) for PM_{10} daily concentration prediction in regional and urban background sites respectively. The results indicated that ANN could be a useful tool to obtain information on air quality status. Cheng et al. (2014) formulated a neural network based ensemble methodology for regional air quality modeling. Díaz-Robles et al. (2008) proposed a hybrid model that combined ANN and ARIMA to predict PM_{10} concentration. The results proved that this approach could improve the forecasting accuracy obtained by either of the models used separately. Feng et al. (2015) introduced air mass trajectory analysis and wavelet transformation to improve ANN forecast precision of daily average $PM_{2.5}$ concentrations. The air mass trajectory was used to recognize distinct corridors and wavelet transformation was applied to deal with the $PM_{2.5}$ concentration fluctuation efficiently. Singh et al. (2012) made a comparison in urban air quality prediction performance among regression approaches and three different types of ANN models. Their findings revealed that general regression neural network (GRNN) showed better forecasting performance than multilayer perceptron network and radial-basis function network. Though ANNs can yield good results, the drawbacks of this technique are the need of lots of training samples and the instability of training results. Support vector machines (SVMs) (Vapnik, 1995, 1998) are effective alternatives on the basis of statistical learning theory methods to overcome the shortcomings of ANNs. Instead of using empirical risk minimization principle to minimize the training error, an upper bound on the generalization error can be minimized through applying structural risk minimization principle in SVM so that the global optimal solutions can be obtained theoretically. Wang et al. (2015) investigated SVM to forecast PM_{10} and SO_2 concentrations in Taiyuan, China. The good results elaborated on the applicability of SVM in air quality forecasting. Yeganeh et al. (2012) focused on an innovative model that combined partial least square as data selection and SVM as a predictor. The findings indicated that this hybrid model presented more accurate and faster prediction ability. Least squares support vector machine (LSSVM) is a modified form for SVM with improved operation speed and convergence accuracy. The LSSVM model has been successfully applied to forecasting problems in many fields, such as annual electric load (Li et al., 2012), wind power (Yuan et al., 2015), gasoline price (Mustaffa et al., 2014), traffic flow (Cong et al., 2016) and so on. However, there are few papers for $PM_{2.5}$ concentration prediction using LSSVM.

The fitting accuracy and generalization ability of LSSVM mainly depend on its two parameters' selection (γ andσ^2). Therefore it's important to employ an appropriate heuristic algorithm to determine the values. Several optimization algorithms have been taken to select parameters for LSSVM, such as genetic algorithm (Wu, 2011; Yuan and Lee, 2015) and particle swarm optimization (Yu et al., 2016; Gorjaei et al., 2015). However, these methods show low calculation efficiency and poor accuracy. Cuckoo search algorithm, as a new metaheuristic method, was introduced by Yang and Suash (2009). Preliminary studies indicate that the mechanism of Levy flight in CS makes the global optimal solution obtained faster and the results insensitive to the change of parameters. Therefore, considering the excellent performance of CS in parameter optimization, this paper employs CS to automatically determine the appropriate values in LSSVM model.

In addition, few $PM_{2.5}$ concentration forecasting methodologies take the input selection into consideration. Thus, principal component analysis (PCA), which can merge the original features and reduce the dimension to simplify computation, is exploited in this paper to select proper input for the forecasting model.

This paper is organized as follows: Section 2 presents a brief review of PCA, LSSVM and CS. Then the novel hybrid prediction technique (PCA-CS-LSSVM) is discussed in detail. Section 3 presents the evaluation criteria of the results. Section 4 provides an experiment study to validate the proposed model and Section 5 obtains the conclusion.

2. Methods

2.1. Principal component analysis

PCA was initially introduced in the discussion of non-random variables by Pearson (1901) and extended to random one by Hotelling (1933). This method can effectively reduce the dimensionality of a data set on the premise of retaining main variance. It is achieved by applying orthogonal transformation to convert the data into a new set of indexes, also called PCs, which meet: (i) Each PC is a linear combination of original variables. (ii) PCs are uncorrelated to each other. The first PC covers the most information of original index and accounts for the largest proportion of variability, while the subsequent PCs interpret the remaining information and variability which their predecessors have not explained one by one. In this paper, the PCA calculation was performed on SPSS v.19.0 and the accumulative explained variation of the selected PCs should be more than 0.85.

2.2. Least squares support vector machine

LSSVM, proposed by Suykens and Vandewalle (1999), is an improvement of SVM. There are two main differences between these two models: (i) LSSVM transforms the inequality constraints into equality ones. (ii) The quadratic programming problems are converted into linear equation problems by utilizing sum squares error loss function as the loss experience in LSSVM model. Therefore, LSSVM simplifies the computational complexity and increases the operation speed. The structure of SVM is shown in Fig. 1.

In LSSVM model, the training sample is set as $\{(x_k, y_k)|k = 1, 2, ..., n\}$, where$x_k \in R^n$ is the input variable and $y_k \in R^n$ is the corresponding output. The specific form of LSSVM model is described as follows:

$$y(x) = \omega^T \cdot \phi(x) + b \tag{1}$$

where $\varphi(x)$ is the nonlinear mapping function that maps the training data into a highly dimensional linear feature space; ω represents weight and b is the bias.

Considering both the complexity of function and error of fitting, the regression problem can be expressed as a constrained optimization problem as follows:

$$\min_{\omega,b,e} (\omega, e) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{k=1}^{n} e_k^2 \tag{2}$$

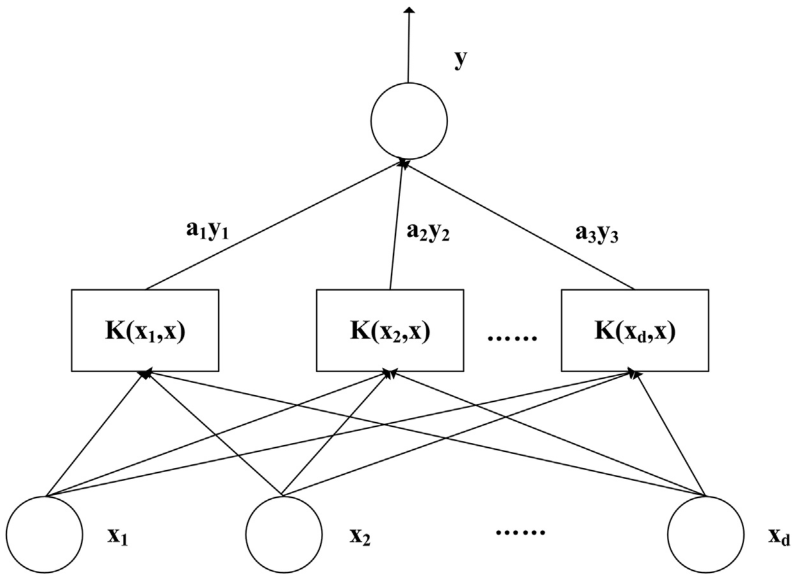


Fig. 1. Structure of SVM.

$$s.t.\ \ y_k = \omega^T \phi(x_k) + b + e_k, \quad k = 1, 2, ..., n \tag{3}$$

where γ equals regularization parameter and e_k is the error.

In order to solve the optimization problem above, the Lagrange function can be defined as follows:

$$L(\omega, b, e, \alpha) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{k=1}^{n} e_k^2 - \sum_{k=1}^{n} \left\{ \alpha_k \left[\omega^T \phi(x_k) + b + e_k - y_k \right] \right\} \tag{4}$$

where $\alpha_k \in R$ equals Lagrange multipliers.

According to the Karush-Kuhn-Tucker (KKT) conditions, Eq. (5) is obtained as follows:

$$\begin{cases} \omega = \sum\limits_{k=1}^{n} \alpha_k \varphi(x_k) \\ \sum\limits_{k=1}^{n} \alpha_k = 0 \\ \alpha_k = e_k \gamma \\ \omega^T \varphi(x_k) + b + e_k - y_k = 0 \end{cases} \tag{5}$$

Thus, after eliminating the variables of ω and e_k, the optimization problem can be transformed into linear equation problems as displayed in Eq. (6):

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(x_1, x_1) + 1/\gamma & \cdots & K(x_1, x_l) \\ \cdots & \cdots & \cdots & \cdots \\ 1 & K(x_l, x_1) & \cdots & K(x_l, x_l) + 1/\gamma \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \cdots \\ \alpha_l \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ \cdots \\ y_l \end{bmatrix} \tag{6}$$

The final form of LSSVM model can be described as follows:

$$f(x) = \sum_{k=1}^{l} \alpha_k K(x, x_i) + b \tag{7}$$

where$K(x, x_i) = \varphi(x)^T \bullet \varphi(x_l)$ is the kernel function which satisfies Mercer's condition. The radial basis function (RBF) is used as the kernel function in this paper and its expression is presented as follows:

$$K(x, x_i) = \exp\left(-\frac{|x - x_i|^2}{2\sigma^2} \right) \tag{8}$$

where σ^2 represents the width of the kernel parameter.

The performance of LSSVM model seriously depends on the input and parameters. The determination of these two parameters, regularization parameter γ and kernel parameter σ^2 is generally based on experience, which easily leads to randomness and inaccuracy in the application of LSSVM. Therefore, cuckoo search algorithm is utilized to optimize these two parameters so that the prediction precision of LSSVM model can be improved.

2.3. *Cuckoo search algorithm*

The technique CS is based on cuckoos' obligate brood parasitic behavior and their way in egg laying and breeding. There exist two main search methods for cuckoos to compete for survival: (i) a random search on the basis of probability of being discovered by a host bird; (ii) a direct search based on Levy flights. Accordingly, cuckoo search algorithm needs to set fewer parameters and is simple for application. Thus CS algorithm is superior to other optimization techniques especially for non-convex and complex optimization problems.

Three ideal conditions must be satisfied: (i) Each cuckoo only lays one egg at a time and randomly selects nest to hatch. (ii) The nest that lays the egg of the highest quality will be held to the next generation. (iii) The number of available host nests is fixed, and a host bird can discover an alien egg with a probability$p_a \in [0, 1]$.

Two search capabilities have been combined in cuckoo search. Local search can be described as follows:

$$x_i^{(t+1)} = x_i^t + \alpha s \oplus H(p_a - \varepsilon) \otimes \left(x_j^t - x_k^t \right) \tag{9}$$

where x_j^t and x_k^t are two different series; $H(u)$ is Heaviside function; ε represents a random number; s means step lengths.

Global search based on Levy flight is given as follows:

$$x_i^{(t+1)} = x_i^t + \alpha \oplus L(s,\lambda) \tag{10}$$

where $L(s,\lambda) = \frac{\lambda \Gamma(\lambda) \sin(\pi\lambda/2)}{\pi} \frac{1}{s^{1+\lambda}}, s > > s_0, 1 < \lambda \leq 3 \lim\limits_{x \to \infty}$; α is the step proportion factor related to the extent of issues at stake. The product$\oplus$ means entry-wise multiplications, which is similar to those used in PSO, but the random walk process via Levy flight here is more efficient in exploring the search space, for its step length is much longer in the run.

The false code of CS is displayed in Table S1 (see Appendix A part).

2.4. LSSVM optimized by CS algorithm

$PM_{2.5}$ concentration forecasting model incorporating PCA, CS and LSSVM is constructed as presented in Fig. 2. In this proposed forecasting approach, there exist two assumptions: (i) The monitoring of each index is independent of each other. (ii) The research area remains stable for a period of time with no natural disasters such as earthquakes and acid rain.

On the basis of CS-LSSVM model, the optimal parameters of LSSVM can be derived as follows:

(1) Input selection. Historical $PM_{2.5}$ concentrations are treated as endogenous variables and other air quality data and temperatures are selected as exogenous variables. After PCA pretreatment, the original data are divided into the training set and the test set.
(2) Parameters initialization. In this paper, suppose the number of host nests is 25, the maximum iteration number is 200, the search ranges of γ and σ^2 is [0.01, 100] and [0.1, 10], respectively.
(3) Population initialization. Suppose that the probability of a cuckoo egg to be discovered by a host bird in its nest is $p_a = 0.25$. Initialize n nest locations$p_i^0 = [x_1^0, x_2^0, ..., x_n^0]^T$, each nest position is a component by γ and σ^2. Calculate the fitting degree of each nest position and obtain the current best nest location x_b^0 and the minimum fitting degree $F_{\min}$.
(4) Update positions. Reserve the best nest position x_b^0 and update other nest locations via Levy flight. Then a new group of nest positions is produced and the fitness degree F is calculated. Compare the new nest positions with p_{i-1} of the preceding generation according to the fitting degree and update the nest position with a better one, thus a new set nest position is derived as follows: $p_t = [x_1^t, x_2^t, ..., x_n^t]^T$.

A random number r is made comparison with p_a. Reserve the nests with lower probability to be discovered in p_t and randomly update the higher one. Calculate the fitting degree of the new nests and replace worse location with better one in p_t in contrast with the precedent fitness degree. Therefore, a current nest location with better test values p_t is found.

(5) Output the optimal solution. Repeat Step (4) up to the maximum number of iterations. Output the global optimal solution x_b^t, corresponding to the optimal parameters γ and σ^2 in LSSVM model. Thus, a $PM_{2.5}$ concentration forecasting model is established.

3. Error measures

To determine which forecasting model outperforms other models, three criteria including mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE) are used to measure the average prediction ability of the model on each data point. These three error indexes are defined as follows:

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |y_t - y_t^*| \tag{11}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N} \left| \frac{y_t - {y_t}^*}{y_t} \right| \times 100\% \tag{12}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N} (y_t - {y_t}^*)^2} \tag{13}$$

where y_t and y_t^* are the actual and forecast $PM_{2.5}$ concentrations at time period t, respectively; N represents the number of time points in test dataset.

4. Experiment study

4.1. Initial selection of input

Baoding City, located in Hebei Province in China, has experienced serious air pollution in recent years with the burgeoning increase of urbanization. The statistic of primary pollutants of Baoding city in the whole year of 2015, as presented in Fig. 3, has shown that $PM_{2.5}$ accounts for 65.80% as the major source of pollution followed by PM_{10} and O_3. Therefore, $PM_{2.5}$ has become the most important factor that causes severe air pollution in this city.

The source of $PM_{2.5}$ pollution is related to the city's industrial and energy structure. According to the investigation of environment protection department in Baoding, the sources can be divided into four groups: (a) Coal burning. It's noted that coal burning, which brings a large amount of SO_2, accounts for more than 70% in total energy consumption, while clean energy only occupies less than 2% in this city. (b) Fugitive dust. The bad road hardening and greening condition, unsuitable management measures in construction process and arid climate all contribute to the increase of PM_{10} concentrations. (c) Exhaust of automobile. By the end of 2014, the total number of motor vehicles has reached 1.7 million in Baoding, which is still on an upward trend. NO_x, CO and PM_{10} caused by cars have gradually become important components of pollution sources. (d) Industrial pollution. The enterprises in high pollution and energy consuming, such as electricity, building material and chemical industry all add the total discharge amount of SO_2, NO_x and dust.

Thus, the initial selection of input is decided according to the factors mentioned above. The air quality data including daily average $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3 range from January 1, 2015 to December 10, 2015. Considering temperature affects the diffusion rate of pollutants, the meteorological forecast data consist of maximum and minimum day temperatures (MaxT and MinT) recorded for the same period. The data from January to October are used as a training set, and the remaining data from November 1 to December 10 are used as a test set. The descriptive statistics of the measured variables are given in Table 1. Fig. 4 displays the original $PM_{2.5}$ concentrations of the 344 data points, which shows obvious seasonal variation. From the figure it can be seen that the concentrations of $PM_{2.5}$ were relatively high in winter, especially in January and December, mainly due to coal fired heating and the climate which was unfavorable to air diffusion. The increased temperature since March accelerated $PM_{2.5}$ to spread and its

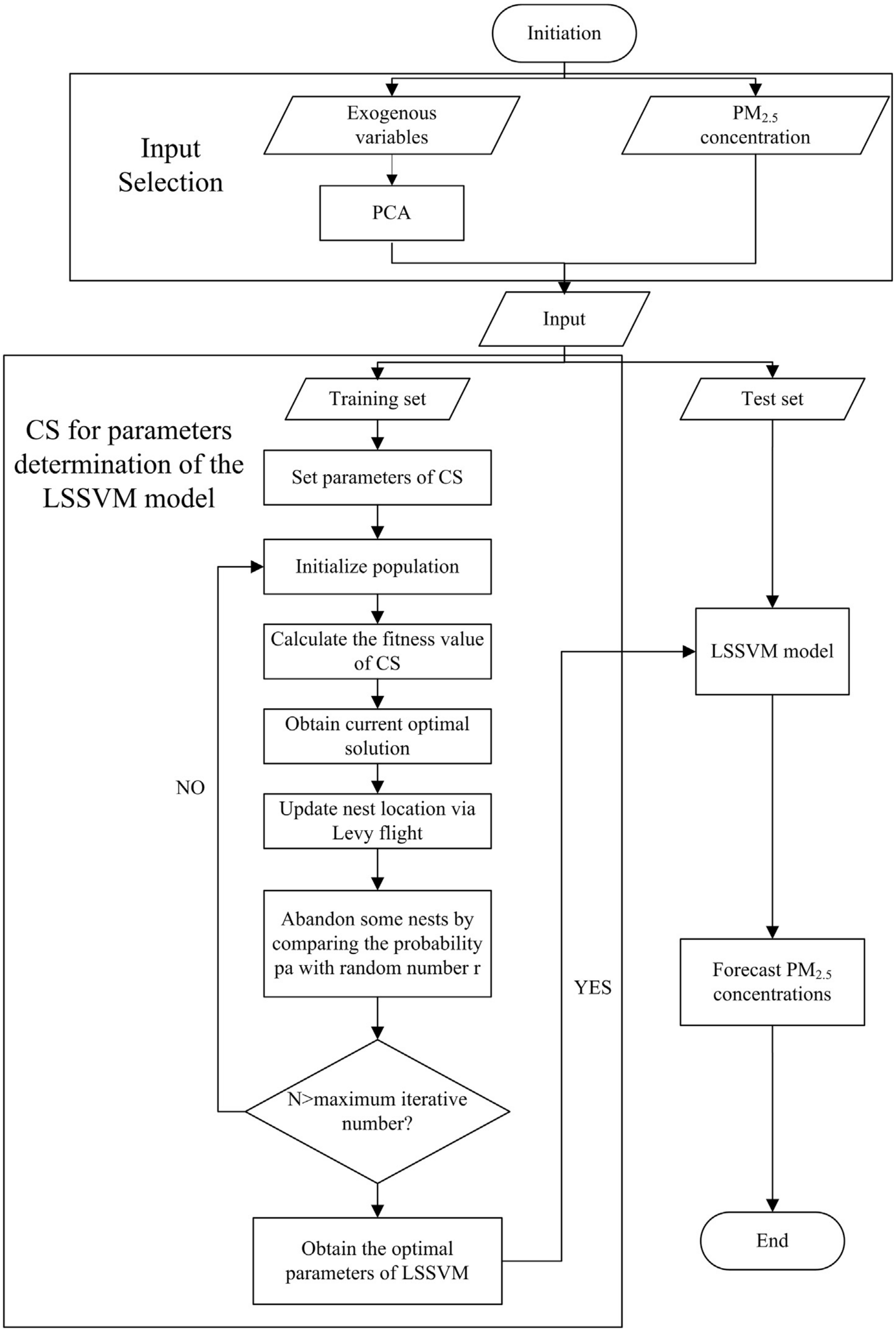

Initiation
Input Selection
Exogenous variables
PCA
PM2.5 concentration
Input
CS for parameters determination of the LSSVM model
Training set
Set parameters of CS
Initialize population
Calculate the fitness value of CS
Obtain current optimal solution
Update nest location via Levy flight
Abandon some nests by comparing the probability pa with random number r
N>maximum iterative number?
NO
YES
Obtain the optimal parameters of LSSVM
Test set
LSSVM model
Forecast PM2.5 concentrations
End

Fig. 2. The flowchart of PCA-CS-LSSVM model.

concentration was lower in that period.

4.2. Correlation analysis and PCA

Mining the relationships between $PM_{2.5}$ and other variables are essential for the establishment of a good prediction model. Pearson coefficient and bilateral significance test are selected for correlation analysis in this paper. Table 2 presents the values of correlation coefficients. It can be found that there is significant correlation between $PM_{2.5}$ and other air quality data except O_3. Temperatures, which represent the coldest and warmest condition of a day, also have great impact on $PM_{2.5}$ formation and transportation. In addition, most variables are correlated with each other.

PCA is utilized to remove the multi collinearity presented in the predictors. We mine the major information containing in the data for the day except $PM_{2.5p,}$ for it stands for the concentration level on the previous day. The PCA process result is shown in Table 3 and Fig. 5. It can be seen that the first two principal components explain more than 85% of the factors, so these two principal components are utilized to replace the predictors as a part of the input in addition to $PM_{2.5p}$.

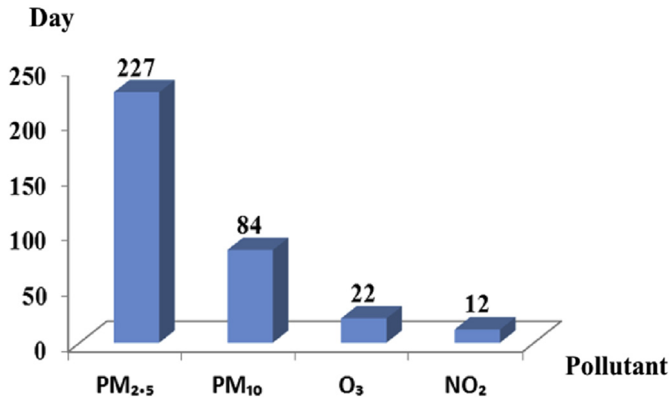


Fig. 3. The sources of primary air pollutants in Baoding in 2015.

Table 1
Statistics of measured variables from January 1, 2015 to December 10, 2015.

Variable	Unit	Range	Mean	Standard deviation
$PM_{2.5}$	ug/m^3	[13, 532.9]	99.69	84.23
PM_{10}	ug/m^3	[28.6, 665.3]	168.07	114.06
SO_2	ug/m^3	[7.3, 243.3]	52.27	46.96
CO	mg/m^3	[0.25, 7.96]	1.70	1.69
NO_2,	ug/m^3	[15.8, 146.5]	50.79	28.31
O_3	ug/m^3	[11, 282]	117.90	66.43
MaxT	°C	[-2, 38]	19.69	10.85
MinT	°C	[-12, 25]	8.55	10.51

4.3. Forecast evaluation and model comparison

The actual and forecast $PM_{2.5}$ values are displayed in Table S2 (see Appendix A part). We can find that the predictions of CS-LSSVM ($\gamma = 50.24$, $\sigma^2 = 8.04$) are much better than LSSVM ($\gamma = 25$, $\sigma^2 = 1$) and GRNN. The prediction error range of CS-LSSVM is [0.35% 29.48%], while the error ranges of LSSVM and GRNN are [1.54% 63.25%] and [4.60% 43.15%] respectively.

The number of errors under 20% generated by CS-LSSVM is 33, which accounts for 82.5% of the total forecasting points as shown in Table 4. In addition, no point of the results of CS-LSSVM is over 30%, but the number of LSSVM and GRNN is 6 and 9, respectively. Thus CS-LSSVM model displays a good performance in $PM_{2.5}$ concentration forecasting.

Fig. 6 shows the deviation between the actual value and the prediction results for the three forecasting models. The error of CS-LSSVM is the minimum one and is more stable than others, which proves CS optimization part can effectively improve the forecasting accuracy. November 15th is the beginning data of heating in Baoding. Keeping warm with coal is the main form in this area. It can be seen that several peaks of $PM_{2.5}$ concentration appear after accumulation of pollutants in a certain period of time.

The performance comparison results of the forecasting models are measured by MAE, MAPE and RMSE as presented in Table 5. It indicates CS-LSSVM outperforms other models in terms of MAE, MAPE and RMSE. This is mainly due to the fact that the PCA process can extract significant information so as to detect duplication and the CS optimization part avoids the randomness of the parameters' setting in LSSVM model.

The daily average $PM_{2.5}$ concentration can be categorized into six levels (Zhou et al., 2014) as shown in Table 6. In order to display the prediction results more directly and clearly, the forecasting classes of daily average $PM_{2.5}$ concentrations are listed in Table 7. We find that 80% of the forecasting $PM_{2.5}$ concentrations of CS-LSSVM model are in the same class with actual ones, which validates CS-LSSVM presents better prediction accuracy than LSSVM and GRNN. Therefore the proposed model provides a basis for traffic restrictions based on even-numbered and odd-numbered license plates in accordance with $PM_{2.5}$ concentration classes so as to reduce haze in Baoding.

Good fit shows that the factors selected in this paper have significant effects on $PM_{2.5}$ concentration, which forms the basis to control it from the sources. The following measures can be taken: (a) Clean energy such as electricity, biomass energy and geothermal power should be utilized instead of coal, and the construction of cogeneration and central heating must be accelerated to achieve

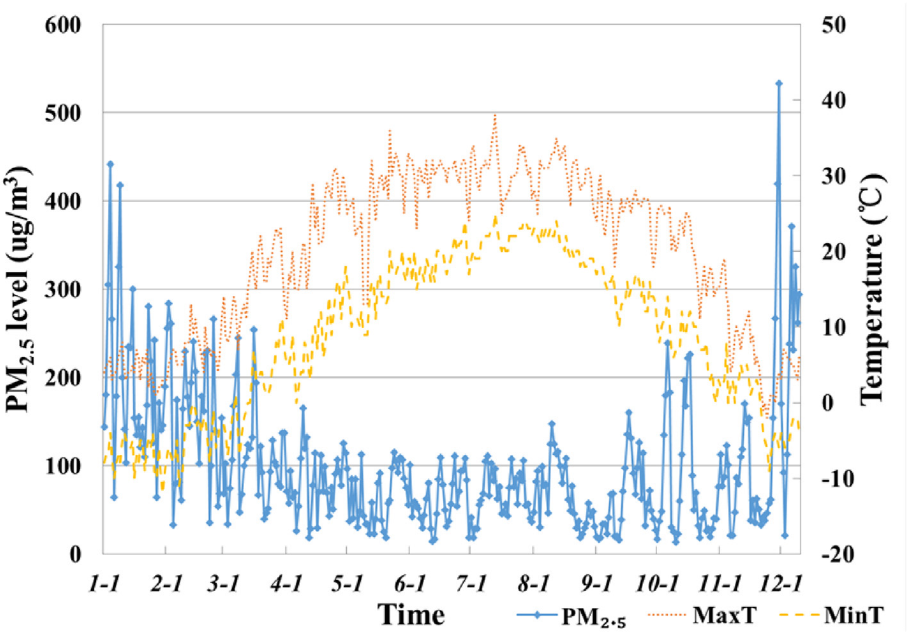


Fig. 4. The original $PM_{2.5}$ data.

Table 2
The correlation coefficients between different air pollutant predictors.

	$PM_{2.5}$	$PM_{2.5,p}$	PM_{10}	SO_2	CO	NO_2	O_3	MaxT	MinT
$PM_{2.5}$	1	0.701	0.96	0.754	0.901	0.779	−0.259	−0.455	−0.450
$PM_{2.5p}$	0.701	1	0.688	0.551	0.663	0.521	−0.302	−0.459	−0.467
PM_{10}	0.96	0.688	1	0.751	0.861	0.755	−0.241	−0.429	−0.464
SO_2	0.754	0.551	0.751	1	0.880	0.654	−0.473	−0.689	−0.716
CO	0.901	0.663	0.861	0.880	1	0.757	−0.490	−0.664	−0.643
NO_2	0.779	0.521	0.755	0.654	0.757	1	−0.377	−0.492	−0.506
O_3	−0.259	−0.302	−0.241	−0.473	−0.490	−0.377	1	−0.832	−0.766
MaxT	−0.455	−0.459	−0.429	−0.689	−0.664	−0.492	0.832	1	0.938
MinT	−0.450	−0.467	−0.464	−0.716	−0.643	−0.506	0.766	0.938	1

Notes: $PM_{2.5p}$ represents $PM_{2.5}$ concentration level a day before the forecast day.

Table 3
Component matrix.

Component	PC1	PC2
PM_{10}	0.776	0.553
SO_2	0.895	0.177
CO	0.915	0.290
NO_2	0.778	0.386
O_3	−0.710	0.598
MaxT	−0.867	0.394
MinT	−0.868	0.446

clean heating in winter. (b) Traffic restrictions based on the last digit of license plate numbers, more strictly based on even-numbered and odd-numbered license plates, can be implemented to reduce pollutant emissions and traffic pressure when $PM_{2.5}$ concentration reaches Class IV or above. (c) Road dust should be controlled and enterprises that cause serious environmental pollution must be closed.

5. Conclusions

A five-year plan for air pollution prevention was carried out in China in 2013 to speed up the steps of reducing haze and improving air quality. High-precision $PM_{2.5}$ forecasting is critical to the accuracy of weather monitoring and warning system and implementation of air pollution control measures. Therefore, the established model in this paper which is based on PCA and improved LSSVM with CS algorithm appears to be very attractive and shows a great extent of improvement. Based on the $PM_{2.5}$ concentration forecasting results, several conclusions can be obtained as follows: (a) the PCA process extracts significant information and reduces the dimension of input which is conducive to improve the prediction accuracy; (b) compared with GRNN, LSSVM shows a better predictive capability in $PM_{2.5}$ concentration forecasting especially in a small-sample situation; (c) the MAE, MAPE and RMSE values of LSSVM algorithm optimized by CS are the lowest, indicating that the proposed method is a promising methodology in $PM_{2.5}$ concentration prediction; (d) the hybrid

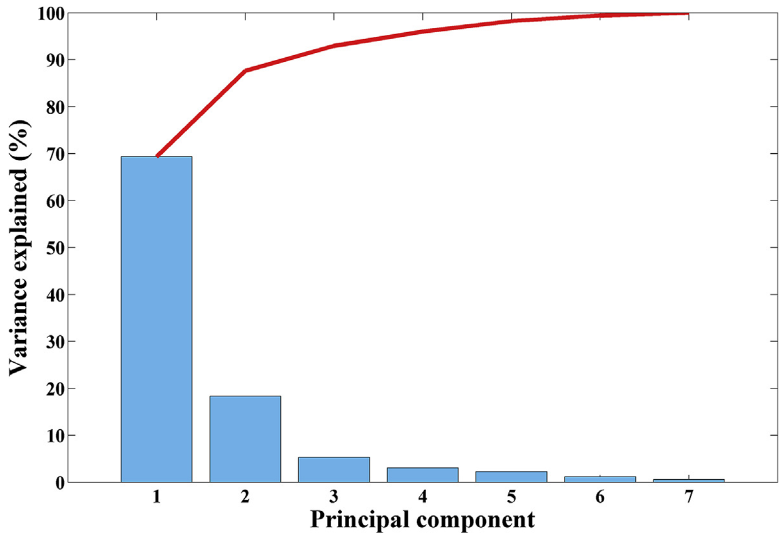


Fig. 5. Scree plot.

Table 4
Accuracy estimation of forecasting models for test samples.

Forecasting models	<20%		20%–30%		>30%	
	Number	Percentage	Number	Percentage	Number	Percentage
CS-LSSVM	33	82.5%	7	17.5%	0	0%
LSSVM	29	72.5%	5	12.5%	6	15%
GRNN	23	57.5%	8	20%	9	22.5%

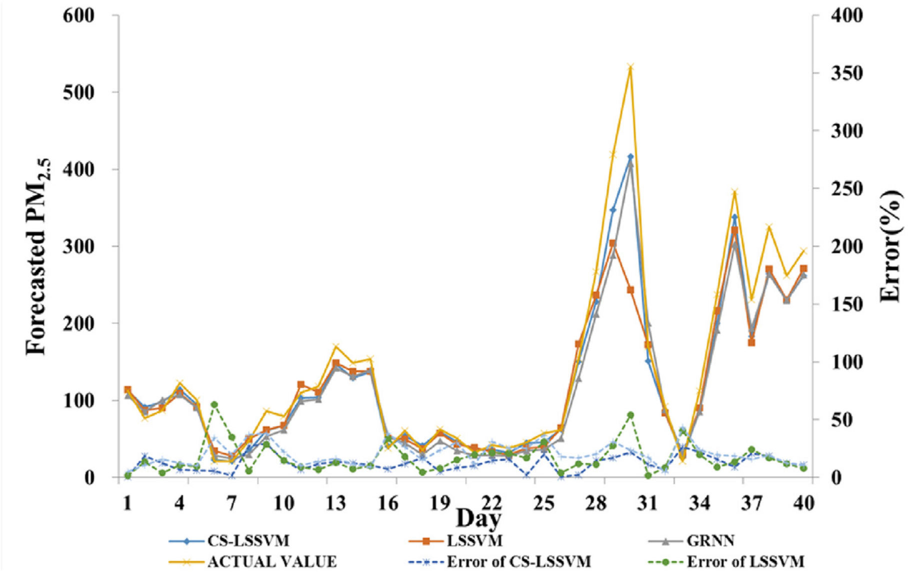


Fig. 6. Forecasting results of different models.

Table 5
Statistical error measures of prediction methods.

Forecasting models	Indexes		
	MAE (ug/m3)	MAPE (%)	RMSE (ug/m3)
CS-LSSVM	18.84	12.56	14.47
LSSVM	24.64	17.15	21.75
GRNN	26.09	20.71	22.89

Table 6
Classes of $PM_{2.5}$ concentration.

$PM_{2.5}$ concentration	Class	Degree of pollution
$PM2.5 \leq 50$	I	Excellent
$50 < PM2.5 \leq 100$	II	Good
$100 < PM2.5 \leq 150$	III	slightly polluted
$150 < PM2.5 \leq 200$	IV	moderately polluted
$200 < PM2.5 \leq 300$	V	heavily polluted
$PM2.5 > 300$	VI	severely polluted

Table 7
Forecasting classes of daily average $PM_{2.5}$ concentration.

Forecasting models	Proportion	
	In the same class	In different classes by one
CS-LSSVM	80%	20%
LSSVM	70%	30%
GRNN	57.5%	42.5%

model in this paper provides a basis for constructing an effective air quality warning system. Due to the complexity of $PM_{2.5}$ formation, this research is mainly applicable for short-term $PM_{2.5}$ forecasting. In our further study, we will attempt to shorten the running time of this forecasting model to make it more practical and explore to predict the concentration of other air pollutants based on this approach.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jenvman.2016.12.011.

References

Cheng, S.Y., Li, L., Chen, D.S., Li, J.B., 2014. A neural network based on ensemble approach for improving the accuracy of meteorological fields used for regional air quality modeling. J. Environ. Manag. 112, 404–414.

Cong, Y.L., Wang, J.W., Li, X.L., 2016. Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm. Proced. Eng. 137, 59–68.

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

Donnelly, A., Misstear, B., Broderick, B., 2015. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmos. Environ. 103, 53–65.

Feng, X., Li, Q., Zhu, Y.J., Hou, J.X., Jin, L.Y., Wang, J.J., 2015. Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118–128.

Gennaro, G.D., Trizio, L., Gilio, A.D., Pey, J., Pérez, N., Cusack, M., Alastuey, A., Querol, X., 2013. Neural network model for the prediction of PM_{10} daily concentrations in two sites in the Western Mediterranean. Sci. Total. Environ. 463–464, 875–883.

Gorjaei, R.G., Songolzadeh, R., Torkaman, M., Safari, M., Zargar, G., 2015. A novel PSO-LSSVM model for predicting liquid rate of two phase flow through wellhead chokes. J. Nat. Gas. Sci. Eng. 24, 228–237.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.

Jian, L., Zhao, Y., Zhu, Y.P., Zhang, M.B., Bertolatti, D., 2012. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. Sci. Total. Environ. 426, 336–345.

Li, Z.H., Guo, S., Zhao, R.H., Su, B.C., Wang, B., 2012. Annual electric load forecasting by a least squares support vector machine with a fruit fly optimization algorithm. Energies 5, 4430–4445.

Mustaffa, Z., Yusof, Y., Kamaruddin, S.S., 2014. Gasoline price forecasting: an application of LSSVM with improved ABC. Proced. Soc. Behav. Sci. 129, 601–609.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos. Mag. 2, 559–572.

Singh, K.P., Gupta, S., Kumar, A., Shukla, S.P., 2012. Linear and nonlinear modeling approaches for urban air quality prediction. Sci. Total. Environ. 426, 244–255.

State Bureau of Environment Protection, 2012. Ambient Air Quality Standard (GB3095-2012). http://www.cnemc.cn/publish/106/news/news_25941.html (Accessed 12 March 2005).

Sun, W., Zhang, H., Palazoglu, A., Singh, A., Zhang, W.D., Liu, S.W., 2013. Prediction of

24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in Northern California. Sci. Total. Environ. 443, 93–103.

Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9, 293–300.

Vapnik, V., 1995. The Nature of Statistic Learning Theory. New York.

Vapnik, V., 1998. Statistical Learning Theory. New York.

Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., Kukkonen, J., 2011. Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM_{10} in Athens and Helsinki. Sci. Total. Environ. 409, 1559–1571.

Wang, P., Liu, Y., Qin, Z.D., Zhang, G.S., 2015. A novel hybrid forecasting model for PM_{10} and SO_2 daily concentrations. Sci. Total. Environ. 505, 1202–1212.

Wang, S., Huang, G.H., Veawab, A., 2013. A sequential factorial analysis approach to characterize the effects of uncertainties for supporting air quality management. Atmos. Environ. 67, 304–312.

Wang, S., Huang, G.H., 2013a. A coupled factorial-analysis-based interval programming approach and its application to air quality management. J. Air Waste Manag. 63, 179–189.

Wang, S., Huang, G.H., 2013b. Interactive fuzzy boundary interval programming for air quality management under uncertainty. Water Air Soil Poll. 224, 1–16.

Wang, Y., Zhang, Y.S., 2009. Air quality assessment by contingent valuation in Ji'nan, China. J. Environ. Manag. 90, 1022–1029.

Wu, Q., 2011. Hybrid model based on wavelet support vector machine and modified genetic algorithm penalizing Gaussian noises for power load forecasts. Expert Syst. Appl. 38, 379–385.

Yang, X.S., Suash, D., 2009. Cuckoo search via Lévy flights. NaBIC2009. In: World Congress on IEEE, pp. 210–214.

Yeganeh, B., Motlagh, M.S.P., Rashidi, Y., Kamalan, H., 2012. Prediction of CO concentrations based on a hybrid partial least square and support vector machine model. Atmos. Environ. 55, 357–365.

You, W., Zang, Z.L., Pan, X.B., Zhang, L.F., Chen, D., 2015. Estimating PM2.5 in Xi'an, China using aerosol optical depth: a comparison between the MODIS and MISR retrieval models. Sci. Total. Environ. 505, 1156–1165.

Yu, H.H., Chen, Y.Y., Hassan, S.G., Li, D.L., 2016. Prediction of the temperature in a Chinese solar greenhouse based on LSSVM optimized by improved PSO. Comput. Electron. Agric. 122, 94–102.

Yuan, F.C., Lee, C.H., 2015. Using least square support vector regression with genetic algorithm to forecast beta systematic risk. J. Comput. Sci. Neth 11, 26–33.

Yuan, X.H., Chen, C., Yuan, Y.B., Huang, Y.H., Tan, Q.X., 2015. Short-term wind power prediction based on LSSVM–GSA model. Energy Convers. Manag. 101, 393–401.

Zhang, H.F., Wang, S.X., Hao, J.M., Wang, X.M., Wang, S.L., Chai, F.H., Li, M., 2016. Air pollution and control action in Beijing. J. Clean. Prod. 112, 1519–1527.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012a. Real-time air quality forecasting, part I: history, techniques, and current status. Atmos. Environ. 60, 632–655.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012b. Real-time air quality forecasting, part II: state of the science, current research needs, and future prospects. Atmos. Environ. 60, 656–676.

Zheng, S.M., Yi, H.T., Li, H., 2015. The impacts of provincial energy and environmental policies on air pollution control in China. Renew. Sust. Energy Rev. 49, 386–394.

Zhou, Q.P., Jiang, H.Y., Wang, J.Z., Zhou, J.L., 2014. A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci. Total. Environ. 496, 264–274.