Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 21 | No. 1 | pp. 393-403 | 2022

Original Research Paper

https://doi.org/10.46488/NEPT.2022.v21i01.047

Open Access Journal

PM2.5 Concentration Prediction Based on Pollutant Pattern Recognition Using PCA-clustering Method and CS Algorithm Optimized SVR

Wei Liu*, Fuji Chen*† and Yihui Chen*
*School of Economics & Management, Fuzhou University, Fuzhou 350116, PR China
†Corresponding author: Fuji Chen; chenfuji@fzu.edu.cn

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 17-04-2021
Revised: 27-06-2021
Accepted: 14-07-2021

Key Words:
PM2.5 concentrations
PCA-clustering
Cuckoo search algorithm
SVR

ABSTRACT

Environmental issues, particularly air pollution, are a matter of concern for people all around the world. PM2.5 levels that are too high harm people's physical and mental health. For government air pollution control, more accurate PM2.5 concentration predictions are critical. In this paper, we explored the relationship between pollutants (PM10, SO_2, NO_2, O_3, CO) and meteorological factors (atmospheric pressure, relative humidity, air temperature, wind speed, wind direction, cumulative precipitation) that affect the generation and transmission of PM2.5. To better predict the concentration of PM2.5, we innovatively combined principal component analysis (PCA) and clustering methods to extract pollutant variables and patterns as important PM2.5 concentration predictors of different models such as support vector regression (SVR), multivariate nonlinear regression (MNR), and artificial neural network (ANN). Compared to MNR and ANN models, SVR presented better prediction accuracy. Moreover, cuckoo search (CS), cross-validation (CV), and particle swarm optimization (PSO) algorithms were used to further optimize the parameters in the process of SVR. And to evaluate the above PM2.5 concentration prediction results, we introduced several evaluating indicators including root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and person correlation coefficient (R) between predicted and measured values. The obtained results confirmed that when the pollutant data was divided into three patterns, the best prediction accuracy was achieved by the CS-SVR model.

INTRODUCTION

Particulate matter (PM) in the atmosphere has gotten a lot of attention in recent decades because it has a big impact on human health. PM2.5 is made up of harmful chemicals such as heavy metals and carcinogenic organic compounds with aerodynamic diameters less than 2.5m. It can easily and deeply penetrate the human lungs to cause serious health issues (Thomaidis et al. 2003, Yuan et al. 2019, Badaloni et al. 2017). High PM2.5 level exposure is correlated to the increase in respiratory and cardiovascular diseases (Ostro et al. 1999, Biancofiore et al. 2017) and population mortality (Di et al. 2017, Liang et al. 2018). It has also been proven that prenatal exposure to PM2.5 can decrease corpus callosum volume and affect children's neuropsychological development (Mortamais et al. 2019, Suades-González et al. 2015).

International environmental organizations and countries all over the world pay great attention to the negative effects of PM2.5. According to the WHO guideline and China's current situation, in 2012 the Ministry of Ecology and Environment (MEE) published Chinese ambient air quality standards, in which the daily and annual average PM2.5 limits were set as 75 and 35 $\mu g.m^{-3}$ (State Bureau of Environment Protection 2012). In Beijing, for example, days with air quality surpassing the MEE limit accounted for 43.5% of the total, which is higher than days with other pollutants such as O_3, PM10, and NO_2 (Beijing Municipal Ecology and Environment Bureau 2019, 2018). PM2.5 concentration predictions that are more accurate help not only people in planning their daily activities but also government regulation.

PM2.5 concentration is influenced by a number of factors, the most important of which are pollutant emission factors and meteorological conditions. The former takes part in the chemical process of PM2.5 formations, and the latter influences the dissipation of PM2.5 (Liang et al. 2015, Wang et al. 2015). With the development of statistical methods, data mining, and artificial intelligence technology, researchers hope to use more simple and effective methods to predict the PM2.5 concentrations. Many efforts have been committed to the data algorithm and optimization. Marsha and Larkin (2019) used a multiple linear regression scheme to forecast daily PM2.5 concentrations using the previous day's PM2.5 measurements as well as fire and smoke-related variables from satellite observations. Sun et al. (2013) used hidden Markov models to forecast daily average PM2.5 concentrations for the next 24 hours. Liu and Sun (2019) used the

supplementary ensemble empirical modal decomposition algorithm, in which the random forest was applied to the decomposition sequence, to effectively reflect the trend of PM2.5 concentration.

In recent years, more effective data mining methods like artificial neural networks and support vector machines have also been successfully implemented in air pollution forecasting. Artificial neural networks, principal component analysis, and k-means clustering technology were combined by Franceschi et al. (2018) to forecast the PM10 and PM2.5 concentrations in Bogotá, Colombia. A hybrid model based on principal component analysis (PCA) and cuckoo search algorithm (CS) optimized least square support vector machine (LSSVM) method was developed by Sun and Sun (2016) to predict PM2.5 concentrations. Gan et al. (2018) proposed a new method based on the secondary-decomposition-ensemble learning paradigm to forecast hourly PM2.5 concentration, in which the least square support vector was used to model all reconstructed components independently. These findings show that the support vector machine method is very effective at predicting PM2.5 concentrations.

In this work, we introduced the PCA-clustering method to CS algorithm optimized SVR for the prediction of PM2.5 concentrations in Beijing. In the beginning, we investigated the correlation between pollutant factors, meteorological factors, and PM2.5 concentrations, and extracted the pollutant variables and patterns using the PCA-clustering method to assist prediction. Then, contrastive studies on parameters optimization algorithms for SVR have been carried out, including cross-validation (CV), particle swarm optimization (PSO), and cuckoo searching (CS) algorithms, to achieve better prediction efficiency. Evaluation metrics such as RMSE, MAE, MAPE, and R were introduced as part of the process. Finally, to further verify the effectiveness of our method, other predictive models like multivariate nonlinear regression (MNR) and artificial neural network (ANN) were also tested. The obtained results indicated that The PCA-clustering approach with SVR optimized by CS algorithm produced the best prediction accuracy.

MATERIALS AND METHODS

Sites and Data

Yizhuang station in Beijing has the most advanced meteorological observation equipment in China, enabling it to provide the most accurate data. Therefore, we model and simulate the PM2.5 predictions with the pollutant data and meteorological data from the Yizhuang observation station as shown in Fig. 1.

The pollutant factors include PM10, SO_2, NO_2, O_3, CO which were collected from Beijing Municipal Environmental Monitoring Center, and the meteorological factors include atmospheric pressure (P), relative humidity (RH), air temperature (T), wind speed (WS), wind direction (WD), 20-20 hours' cumulative precipitation (CP) which were collected from the National Meteorological Information Center. The details of the original data are shown in Table 1. The 24 h average of the pollutant factors was calculated for the purpose of PM2.5 concentration prediction. Fig. 2 shows the PM2.5 concentration and temperature of Yizhuang station from October 14, 2014, to December 31, 2017. When some variables' data was missing for several days in a row, the associated dates data was removed, and the sporadic missing data was imputed using the EM imputation method.

As shown in Fig. 2, the trends of PM2.5 concentration and temperature are opposite. Low PM2.5 values were observed in the warm period from April to September, while high PM2.5 values were observed in the cold days from October

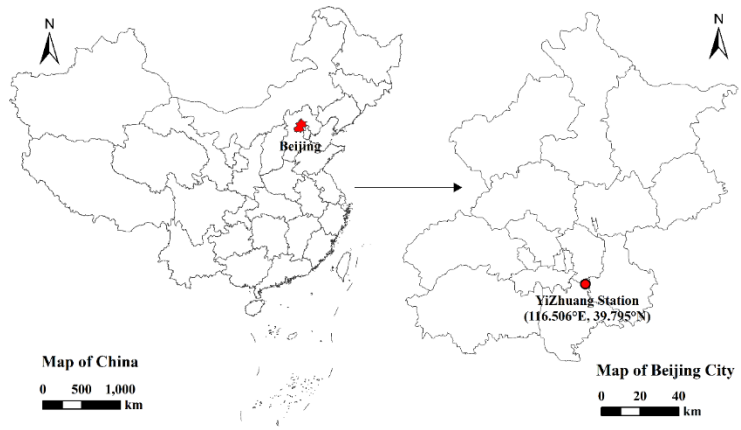

Fig. 1: The geography of YiZhuang pollutant and meteorological monitoring Station (39.795N, 116.506 E).

Table 1: Original pollutant and meteorological variables.

Pollutant Variables	Frequency	Pollutant Variables	Frequency
PM2.5	hourly	atmospheric pressure (P)	24 h average
PM10	hourly	relative humidity (RH)	24 h average
SO_2	hourly	air temperature (T)	24 h average
NO_2	hourly	wind speed (WS)	Max wind speed
O_3	hourly	wind speed (WD)	Max wind direction
CO	hourly	cumulative precipitation (CP)	20-20 h

to March next year, some of which were even more than 500μg/m^3. Considering that high-level PM2.5 concentration has a great impact on people's lives, we use the atmospheric environment data of cold days to establish a prediction model for PM2.5 concentration.

Methods

PCA-clustering method to extract pollutant variables and pollutant patterns: The principal component analysis (PCA) algorithm has been widely applied for reducing the dimension of the data set on the premise of retaining the main variance. A new set of variables can be achieved by PCA transforming which are called principal components (PCs). To simplify the structure of the dataset, only the first few PCs with large variance are usually chosen to reflect the information of the original variables in the real research process. In most cases, a cumulative variance contribution rate of more than 85% for the first several major components is appropriate. For the purpose of this study, PCA was combined with the correlation coefficient between PM2.5 concentration and related covariates to find the primary influencing factors. Moreover, k-meaning clustering was further introduced to extract pollutant patterns.

Support vector machine regression model: Support vector machine (SVM), originally developed by Vapnik in the 1990s (Vapnik 1995, 1998), is one of the most robust and accurate data mining algorithms, mainly including support vector machine classification (SVC) and support vector machine regression (SVR). It is very flexible to solve all kinds of nonlinear classification regression problems (Wu & Kumar 2013). In this paper, SVR has been used to build the PM2.5 concentration prediction model for satisfying results.

In the SVR model, the training data is set as $\{(x_i, y_i)|i=1,2,\cdots,n\}$, where $x_i \in R^n$ is the input variable and y_i is the corresponding dependent variable. To learn a $g(x)$ close to y, the SVR linear regression model can be as follows:

$$g(x) = w^T \cdot x_i + b \quad \ldots(1)$$

where w, b are the pending parameters. To obtain larger intervals and smaller amounts of noise data, relaxation variables ξ_i and $\hat{\xi}_i$ are further introduced, and the SVR regression problem can be expressed as follows:

$$\min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) \quad \ldots(2)$$

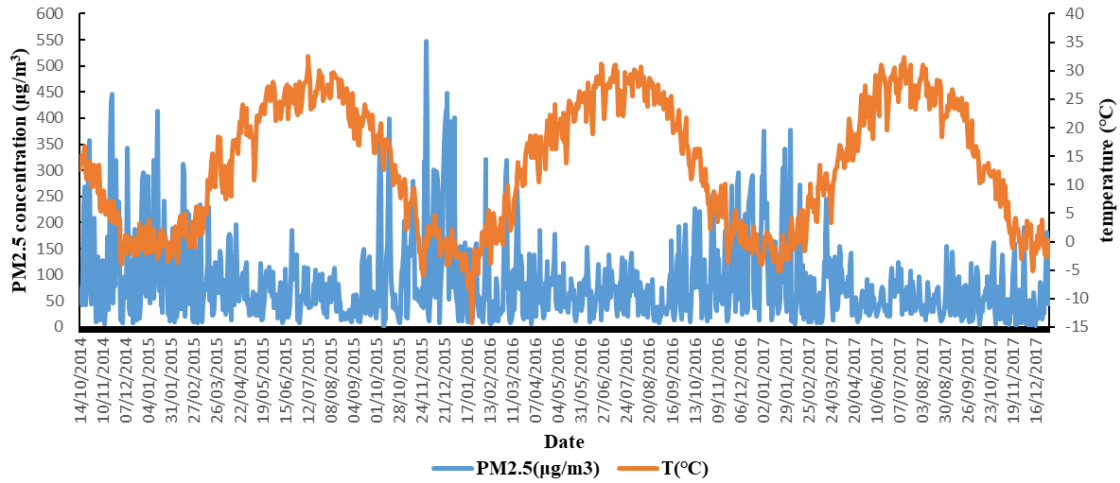

Fig. 2: The PM2.5 concentration and temperature of Yizhuang from October 14, 2014, to December 31, 2017.

$$\text{s.t } g(x_i) - y_i \le \epsilon + \xi_i,$$
$$y_i - g(x_i) \le \epsilon + \hat{\xi}_i,$$
$$\xi_i \ge 0,\ \hat{\xi}_i \ge 0,\ i = 1,2,\cdots,n.$$

where parameter C is the penalty factor to solve the optimization problem, where $\alpha_i, \mu_i \in R$ are Lagrange multiplier. Meanwhile, $\alpha, \hat{\alpha}, \mu, \hat{\mu} \in R$ are introduced to build the Lagrange function as follows:

$$L(\omega, b, \alpha, \hat{\alpha}, \xi, \hat{\xi}, \mu, \hat{\mu}) = \frac{1}{2}\|\omega\|^2 + C$$
$$\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) - \sum_{i=1}^{n}\mu_i\xi_i - \sum_{i=1}^{n}\hat{\mu}_i\hat{\xi}_i +$$
$$\sum_{i=1}^{n}\alpha_i\,(g(x_i) - y_i - \varepsilon - \xi_i) + \sum_{i=1}^{n}$$
$$\hat{\alpha}_i\,(y_i - g(x_i) - \varepsilon - \hat{\xi}_i) \quad \ldots(3)$$

In Eq. (3), by fixing $\alpha, \hat{\alpha}, \mu, \hat{\mu},$, calculating derivation of $\omega, b, \xi, \hat{\xi}$ and setting the results as 0, the following Eqs. are obtained:

$$w = \sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i)\,x_i,, \quad \ldots(4)$$
$$0 = \sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i),,$$
$$C = \alpha_i + \mu_i,$$
$$C = \hat{\alpha}_i + \hat{\mu}_i.$$

Putting the four Eqs. (4) into Eq. (3), and adding Karush-Kuhn-Tucker conditions to the obtained duality problem, Eq. (5) is achieved as follows:

$$max \sum_{i=1}^{n} y_i(\hat{\alpha}_i - \alpha_i) - \varepsilon(\hat{\alpha}_i + \alpha_i) - \frac{1}{2}\sum_{i=1}$$
$$\sum_{j=1}^{n}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)\langle x_i, x_j\rangle \quad \ldots(5)$$
$$\text{s.t } \sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i) = 0,$$
$$0 \le \alpha_i, \hat{\alpha}_i \le C.$$
$$\text{KKT: } \alpha_i(g(x_i) - y_i - \varepsilon - \xi_i) = 0,$$
$$\hat{\alpha}_i(y_i - g(x_i) - \varepsilon - \hat{\xi}_i) = 0,$$
$$\alpha_i\hat{\alpha}_i = 0,\ \xi_i\hat{\xi}_i = 0$$
$$(C - \alpha_i)\xi_i = 0,\ (C - \hat{\alpha}_i)\hat{\xi}_i = 0.$$

To efficiently solve the above optimization problem, the SMO algorithm is used. After determining the optimal Lagrange multiplier, the values of ω and b can be obtained. Accordingly, the final SVM regression modal can be defined as follows:

$$g(x) = \sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i)\,\langle\emptyset(x_i), \emptyset(x)\rangle + b$$
$$= \sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i)\,k(x_i, x) + b \quad \ldots(6)$$

where $\emptyset(x)$ is the nonlinear mapping function that maps the data into a linear feature space with a higher dimension. The kernel function $k(x_i, x)$ satisfying Mercer's condition can be used instead of the mapping function to solve the complex dimensions and computing problems. In this paper, the radial basis function is used as the kernel function (Eq. (7)):

$$k(x_i, x) = exp\left(-\frac{|x - x_i|^2}{2\sigma^2}\right) \quad \ldots(7)$$

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

SVR Optimized by the Cuckoo Search Algorithm

In the SVR nonlinear prediction model with radial basis function as kernel function, the penalty C and the width σ^2 are the parameters. In this paper, the cuckoo search (CS) algorithm is introduced to optimize these two parameters to improve the efficiency and accuracy of prediction.

Yang & Deb (2009) presented the Cuckoo search natural heuristic method, which mimics cuckoo brood parasitism behavior. The algorithm can be enhanced by Levy flight rather than a simple isotropic random walk. The CS algorithm combines global search and local search which are controlled by discovery probability (P_a). This makes it possible to explore the search space more efficiently in the global scope, achieving global optimum with a higher probability. Although the PSO algorithm may converge to local optimization prematurely, it is not necessarily the global optimal solution. While CS can usually converge to global optimization.

In the D dimensional space, the population of the nest is n, $X_{Nestpop} = [X_1, X_2, ..., X_n]^T$, and each nest is the solution to the problem. In each nest, there is a D dimension vector$\{X_i = [X_{i1}, X_{i2}, ..., X_{iD}]^T | i = 1,2,..,n\}$.

After the nest population is formed randomly, CS updates the individual through two paths:

i) the cuckoo uses Levy flight-based Eq (8) to find the nest and lay an egg.

$$X_{t+1} = X_t + \alpha S = X_t + \alpha \otimes Levy(\beta) \quad \ldots(8)$$

$$Levy(\beta) \sim \mu = t^{-\beta}, 1 \le \beta \le 3 \quad \ldots(9)$$

In combined Eq. (8) and Eq. (9), S is the random step size obeying Levy distribution.

$$S = \frac{U}{|V|^{\frac{1}{\beta}}}, (U \sim N(0, \sigma^2), V \sim N(0,1)) \quad \ldots(10)$$

$$\sigma = \left\{ \frac{\Gamma(1+\beta) * \sin(\frac{\pi\beta}{2})}{\beta * \Gamma(\frac{1+\beta}{2}) * 2^{\frac{\beta-1}{2}}} \right\}^{\frac{1}{\beta}} \quad \ldots(11)$$

where α is the scaling factor of step size, which is to 0.01, and β is set to 1.5.

ii) the host uses random walk to rebuild its nest after finding the alien egg with the probability of P_a (Eq. (12)).

$$X_{t+1} = X_t + \gamma \otimes Heaviside(P_a - \epsilon) \otimes (X_i - X_j) \quad \ldots(12)$$

where $P_a = 0.25$ is recommended. γ, ∈ are random numbers subject to a uniform distribution. $Heaviside(P_a - \epsilon)$ is the Heaviside step function. When $P_a > \epsilon$, $Heaviside(P_a - \epsilon) = 1$, when $P_a < \epsilon$ $Heaviside(P_a - \epsilon) = 0$, when $P_a = \epsilon$ $Heaviside(P_a - \epsilon) = 0.5$ X_i, X_j are any other nests.

The flow chart of SVR prediction optimized by the CS algorithm (CS-SVR) is shown in Fig. 3.

Evaluation index for prediction results: To investigate the accuracy of different PM2.5 concentration prediction models, four evaluation indexes are applied, including person correlation coefficient (R), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). R can release the relevance between the observed value and the predicted value (Eq. (12)). Mean squared error (MSE) is the expected value of the square of the difference between the predicted value and the observed value. Correspondingly, RMSE is the square root of MSE which is more intuitive in order of magnitude (Eq. (13)). And the smaller the RMSE value, the better the accuracy of the prediction model. MAE represents the mean of the absolute error between the predicted value and the observed value, which can better reflect the actual predicted error (Eq. (14)). MAPE is used to better evaluate different models with the same set of data (Eq. (15)).

$$R = \frac{cov(observed, predicted)}{\sigma_{observed}\sigma_{predicted}} \quad \ldots(12)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2} \quad \ldots(13)$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|observed_t - predicted_t| \quad \ldots(14)$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{observed_t - predicted_t}{observed_t}\right| \times 100\% \quad \ldots(15)$$

RESULTS AND DISCUSSION

The Results of Extracting Pollutant Variables for Pattern Calculation

The atmospheric environment affecting PM2.5 concentration consists of pollutant factors and meteorological factors. We

Table 2: Correlation coefficient (R) between the PM2.5 concentration and 6 pollutant variables, 5 meteorological variables.

	PM2.5	PM10	SO_2	NO_2	O_3	CO	P	RH	T	WS	WD	CP
PM2.5	1	-	-	-	-	-	-	-	-	-	-	-
PM10	.913**	1	-	-	-	-	-	-	-	-	-	-
SO_2	.575**	.541**	1	-	-	-	-	-	-	-	-	-
NO_2	.785**	.745**	.574**	1	-	-	-	-	-	-	-	-
O_3	-.296**	-.245**	-.373**	-.523**	1	-	-	-	-	-	-	-
CO	.832**	.755**	.574**	.782**	-.434**	1	-	-	-	-	-	-
P	.086**	.040	.324**	.226**	-.671**	.233**	1	-	-	-	-	-
RH	.406**	.279**	-.126**	.262**	-.049	.374**	-.271**	1	-	-	-	-
T	-.245**	-.202**	-.486**	-.351**	.758**	-.381**	-.880**	.269**	1	-	-	-
WS	-.362**	-.300**	-.251**	-.515**	.286**	-.376**	-.049	-.379**	.057	1	-	-
WD	.275**	.241**	.159**	.275**	-.038	.276**	-.081**	.353**	.077**	-.264**	1	-
CP	.009	.004	-.072*	-.004	.070*	.021	-.125**	.128**	.136**	.006	.067*	1

*Note: **p-value≤0.01, *p-value≤0.05*

Table 3: The total variance explained of 6 pollutant variables and 6 meteorological variables.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.847	40.388	40.388	4.847	40.388	40.388
2	2.517	20.975	61.362	2.517	20.975	61.362
3	1.105	9.212	70.574	1.105	9.212	70.574

Table 4: The rotated component matrix of pollutant variables of the 12 variables calculation, indicating their interpretive competence to each principal component.

pollutant variables and meteorological variables	Component		
	1	2	3
PM2.5	0.929	-0.050	0.223
PM10	0.928	0.015	0.116
SO_2	0.728	-0.321	-0.189
NO_2	0.799	-0.304	0.291
O_3	-0.232	0.863	-0.159
CO	0.831	-0.249	0.265

examined their relationship with PM2.5 concentrations. The correlation coefficient (R) between the 11 variables (from 14th October 2014 to 31st December 2017) and PM2.5 concentrations is shown in Table 2. The R values between pollutant factors and PM2.5 concentrations are mostly higher than 0.5, but the R values in the meteorological parts are between 0.1 and 0.4. These results indicate that pollutant factors have a greater impact on the PM2.5 concentrations. Therefore, we decided to extract pollutant patterns to improve the prediction accuracy. The PCA-clustering method was employed for extracting needed variables.

In the beginning, all the 12 atmospheric environment variables including 6 pollutant components (PM2.5, PM10, SO_2, NO_2, O_3, CO) and 6 meteorology factors (P, RH, T, WS, WD, CP) were examined by PCA methods, and the results are shown in Tables 3 and 4. In our previous work, we have confirmed that relative humidity (RH), temperature (T), and wind speed (WS) have a more significant impact on the concentration level of PM2.5 (Liu et al. 2019). As a result, the PCA method was used to construct an examination using nine variables (PM2.5, PM10, SO_2, NO_2, O_3, CO, RH, T, and WS) (Tables 5 and 6). As shown in Table 3 and Fig. 4, the first three principal components can explain the degree of data variation as 70.57%, with the first one accounting for 40.39%. However, for the 9 variables calculation, the first three principal components can explain 82.017%, and the first one accounts for 51.321% (Table 5 and Fig. 4). Therefore, 3 meteorological variables (RH, T, WS) will be introduced for predicting PM2.5 concentrations in the next part.

The rotated component matrix in Table 4 shows the interpretive competency of pollutant variables to each primary component in the 12 variables (6 pollutant and 6 meteorology variables) computation. PM2.5, PM10, SO_2, NO_2, and CO all strongly explain the first principal component, however, O_3 is shifted to the second principal component. The same pattern may be seen in the findings of the 9 variable calculation (6 pollutant and 3 meteorology variables) and the results as shown in Table 6. PM10, SO_2, NO_2, and CO are used in the

Table 5: The total variance explained of 6 pollutant variables and 3 meteorological variables.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	total	% of Variance	Cumulative %
1	4.619	51.321	51.321	4.619	51.321	51.321
2	1.749	19.429	70.750	1.749	19.429	70.750
3	1.014	11.267	82.017	1.014	11.267	82.017

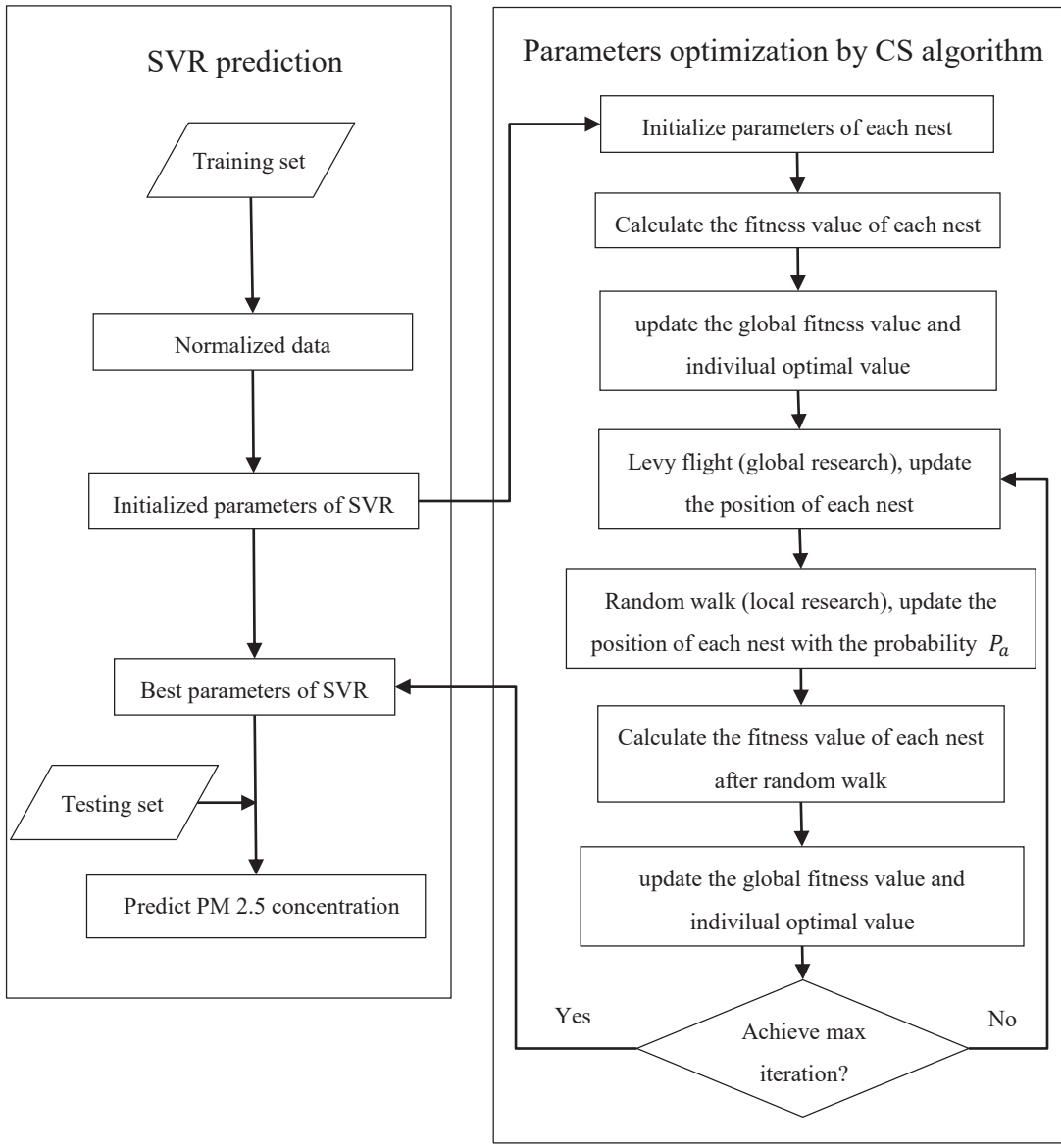

Fig. 3: The flow chart of CS algorithm optimized SVR prediction (CS-SVR).

extraction of pollutant patterns, which will be generated using the clustering approach and used as a key input factor for the subsequent prediction model, based on the study above.

The Results of PM2.5 Concentration Prediction by CS-SVR Model

All cold day data from October to the following March 2014-2017 was used as a training set, while the data from November and December 2017 was used as a testing set for PM2.5 concentration prediction. The population of the nest is set to 20 in the CS optimization procedure. The discovery probability P_a = 0.25 is recommended. Penalty C and the width σ^2 as parameters to be optimized are set to [0.01, 100]. In the optimization process, 100 iterations have been carried out.

First, three prediction models including CS-SVR, multivariate nonlinear regression (MNR), and artificial neural network (ANN) have been studied using 11 variables. As shown in Fig. 5, the absolute value of relative errors of CS-SVR is much better than those of MNR and ANN. The

prediction accuracy of different models was further compared using four evaluation indexes such as R, MAE, RMSE, and MAPE (Table 7). The R values between the predicted and observed PM2.5 concentration by CS-SVR (0.9430) are higher than ANN (0.9342) and MNR (0.9326). Meanwhile, the MAE, RMSE, and MAPE indicators of CS-SVR decreased by 30.10%, 10.22%, 70.87% than MNR, and by 13.88%, 3.4451%, 51.48% than ANN, respectively. All of these findings indicate that the CS-SVR model outperforms the MNR and ANN models in predicting PM2.5 concentrations. In addition, different optimization methods for SVR were also investigated (Table 7). All the four indexes (R, MAE, RMSE, and MAPE) of CS-SVR are better than those of CV-SVR and POS-SVR.

In addition, eight factors were used in the CS-SVR prediction model. CS-SVR still outperforms ANN and MNR in terms of prediction performance, as demonstrated in Fig. 6 and Table 8, despite the greater R-value and smaller MAE,

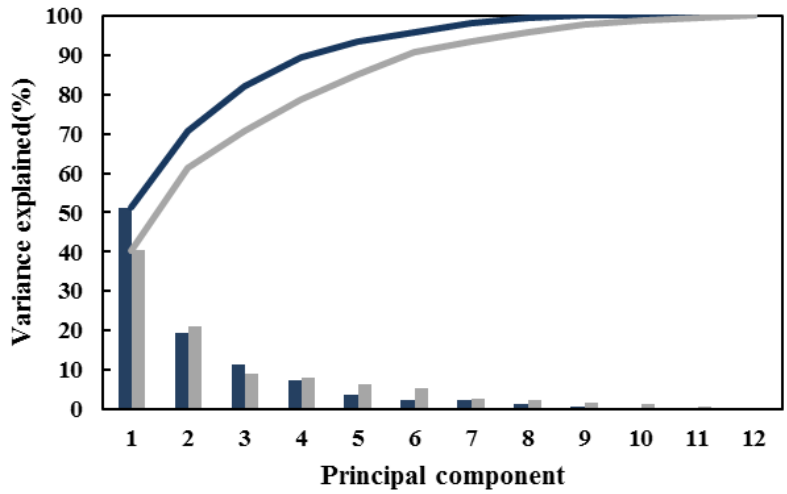

Fig. 4: The scree plots of 11 variables (gray) vs. 9 variables (black).

Table 6: The rotated component matrix of pollutant variables of the 9 variables calculation, indicating their interpretive competence to each principal component.

pollutant variables and meteorological variables	Component		
	1	2	3
PM2.5	0.924	-0.067	0.271
PM10	0.927	-0.016	0.157
SO_2	0.710	-0.407	-0.195
NO_2	0.758	-0.363	0.342
O_3	-0.166	0.896	-0.228
CO	0.817	-0.262	0.303

Table 7: The evaluation index results of CS-SVR, CV-SVR, POS-SVR, ANN, and MNR models using 11 pollutant and meteorological variables.

	Index			
Models	R	MAE	RMSE	MAPE
CS-SVR	0.9430	11.5785	18.2506	0.3141
CV-SVR	0.9341	13.2025	19.6715	0.5225
POS-SVR	0.9423	11.6407	18.7494	0.3158
ANN	0.9342	13.4451	18.9453	0.6480
MNR	0.9326	16.5625	20.3276	1.0794

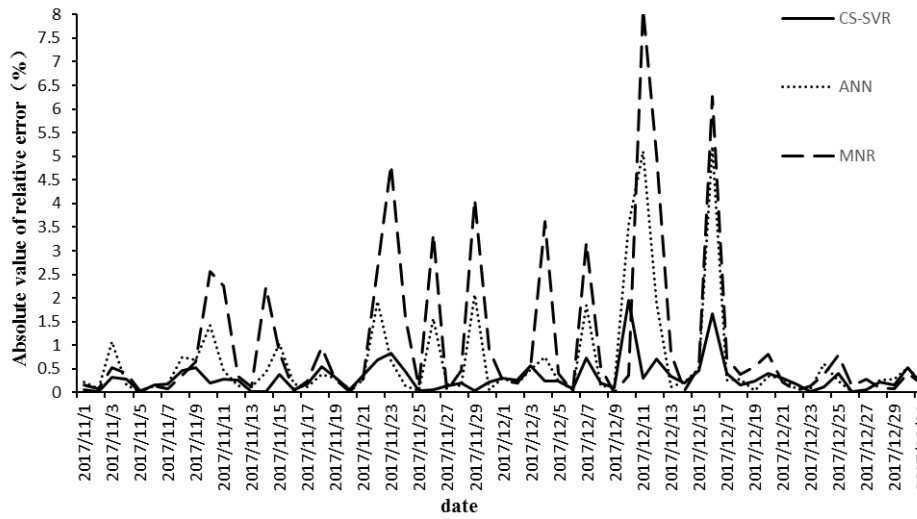

Fig. 5: The absolute value of the relative error of different prediction models using 11 pollutant and meteorological variables (PM10, SO_2, NO_2, O_3, CO and P, RH, T, WS, WD, CP).

RMSE, and MAPE indicators. Therefore, from the above two groups of comparative experiments, the prediction accuracy of the CS-SVR model is better than the other two models in terms of each index.

Interestingly, although using fewer variables, the 8 variables CS-SVR prediction shows acceptable prediction accuracy. However, the R-value of the 8 variables CS-SVR prediction is a little bit lower than the 11 variables one (0.9388 *vs* 0.9430). To further improve the prediction accuracy of the 8 variables CS-SVR model, we used the PCA-clustering method to extract the pollutant pattern as an additional variable for PM2.5 concentration prediction. The obtained prediction results taking into account the pollutant pattern variable are shown in Table 9. It is quite clear that when the pollutant pattern variable is involved in the calculation, the prediction accuracy of all models improved. Especially, for the CS-SVR model, when the pollutant data was divided into three patterns (k = 3) by k-meaning clustering, the best prediction accuracy was achieved. The R-value increased to 0.9455, while the MAE, RMSE, and MAPE values decreased to 11.2523, 16.7114, and 0.3006, respectively, the lowest values of all models. As a result, the PCA-clustering extracted pollutant pattern-based CS-SVR model predicts PM2.5 concentrations the best.

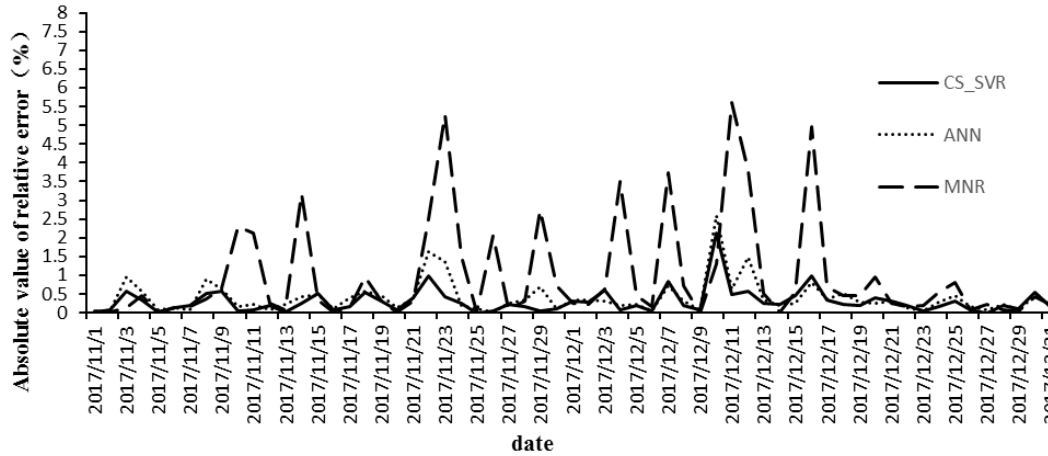

Fig. 6: The absolute value of the relative error of different prediction models using 8 pollutant and meteorological variables (PM10, SO_2, NO_2, O_3, CO and RH, T, WS).

Table 8: The evaluation index results of CS-SVR, ANN, and MNR models using 8 pollutant and meteorological variables.

	Index			
Models	R	MAE	RMSE	MAPE
CS-SVR	0.9388	11.4269	18.6665	0.3008
ANN	0.9316	12.7832	19.2167	0.4021
MNR	0.9367	15.7408	19.8749	0.9721

Table 9: The evaluation index results of CS-SVR, CV-SVR, PSO-SVR, ANN, and MNR models using 5 pollutant variables, 3 meteorological variables, and 1 pollutant pattern variable.

		Index			
Models	k	R	MAE	RMSE	MAPE
CS-SVR	1	0.9388	11.4269	18.6665	0.3008
	2	0.9378	12.0629	18.9555	0.3207
	3	0.9455	11.2523	16.7114	0.3006
	4	0.9453	11.6977	16.9446	0.422
	5	0.9392	11.6424	18.6715	0.316
CV-SVR	1	0.91158	13.6238	20.8764	0.6888
	2	0.9253	12.0441	19.0600	0.4241
	3	0.9292	11.8964	18.3937	0.5254
	4	0.9366	11.9500	17.6161	0.5364
	5	0.9295	12.5774	19.1370	0.5066
PSO-SVR	1	0.9338	11.7656	18.9985	0.3722
	2	0.9345	11.7015	18.8554	0.3350
	3	0.9424	11.7059	18.8421	0.3119
	4	0.9464	11.7066	18.4610	0.3328
	5	0.9377	11.9483	19.0073	0.3990
ANN	1	0.9316	12.7832	19.2167	0.4021
	2	0.9298	13.3388	20.0169	0.3789
	3	0.9455	11.6921	17.1260	0.3241
	4	0.9480	12.5176	17.7708	0.4623
	5	0.9397	11.97	18.0200	0.5673
MNR	1	0.9367	15.7408	19.8749	0.9721
	2	0.9388	15.2120	19.3763	0.8988
	3	0.9478	14.9337	18.8504	0.8527
	4	0.9471	15.0780	18.6994	0.8325
	5	0.9383	15.7837	19.8648	0.9405

k is the number of pollutant patterns

CONCLUSION

In this work, we applied principal component analysis (PCA)-clustering-based pollutant pattern recognition and CS algorithm optimized SVR for the PM2.5 concentration prediction. In comparison to a prediction based on all meteorological factors, relative humidity, air temperature, and wind speed could be chosen to provide an acceptable prediction accuracy. To further improve the prediction results, a new variable (k) of pollutant pattern was extracted by the PCA-clustering method and added to the calculation. Support vector regression outperforms multivariate nonlinear

regression and artificial neural network models, as seen by indices like RMSE, MAE, MAPE, and R. Furthermore, the cuckoo search was used to further optimize the parameters in the SVR process, which resulted in a better prediction of PM2.5 concentration than cross-validation and particle swarm optimization algorithms. According to these comparative studies, the best PM2.5 concentration prediction accuracy could be obtained by the CS-SVR model with three pollutant patterns (k = 3). Pollutant and meteorological data from more observation stations will be introduced in the future to further prove the reliability of our prediction model and acquire higher prediction accuracy.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Natural Science Foundation of China (71271056).

REFERENCES

Badaloni, C., Cesaroni, G., Cerza, F., Davoli, M., Brunekreef, B. and Forastiere, F. 2017. Effects of long-term exposure to particulate matter and metal components on mortality in the Rome longitudinal study. Environ. Int., 109: 146-154.

Beijing Municipal Ecology and Environment Bureau. 2019. Beijing Ecology and Environment Statement 2018.

Beijing Municipal Ecology and Environment Bureau. 2018. Beijing Ecology and Environment Statement 2017.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Tommaso, S.D., Colangeli, C., Rosatelli, G. and Carlo, P.D. 2017. Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmos. Pollut. Res., 8: 652-659.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F. and Schwartz, J.D. 2017. Air pollution and mortality in the medicare population. N. Engl. J. Med., 376: 2513-2522.

Franceschi, F., Cobo, M. and Figueredo, M. 2018. Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. Atmos. Pollut. Res., 9: 912-622.

Gan, K., Sun, S., Wang, S. and Wei, Y. 2018. A secondary-decomposition-ensemble learning paradigm for forecasting $PM_{2.5}$ concentrations. Atmos. Pollut. Res., 9: 989-999.

Liang, F., Xiao, Q., Gu, D., Xu, M., Tian, L., Guo, Q., Wu, Z., Pan, X. and Liu, Y. 2018. Satellite-based short- and long-term exposure to PM 2.5, and adult mortality in urban Beijing, China. Environ. Pollut., 242: 492-499.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Huang, H. and Chen, S.X. 2015. Assessing Beijing's PM 2.5 pollution: severity, weather impact, APEC and winter heating. Proc. R. Soc. A., 471: 20150257.

Liu, D. and Sun, K. 2019. Short-term PM2.5 forecasting based on CEEMD-RF in five cities of China. Environ. Sci. Pollut. Res., 26: 32790-32803.

Liu, W., Guo, G., Chen, F. and Chen, Y. 2019. Meteorological pattern analysis assisted daily PM2.5 grades prediction using SVM optimized by PSO algorithm. Atmos. Pollut. Res., 10: 1482-1491.

Marsha, A. and Larkin, N.K. 2019. A statistical model for predicting $PM_{2.5}$ for the western United States. J. Air Waste Manag. Assoc., 69: 1215-1229.

Mortamais, M., Pujol, J., Martínez-Vilavella, G., Fenoll, R. Reynes, C., Sabatier, R., Rivas, I., Forns, J., Vilor-Tejedor, N., Alemany, S., Cirach, M., Alvarez-Pedrerol, M., Nieuwenhuijsen and M., Sunyer, J. 2019. Effects of prenatal exposure to particulate matter air pollution on corpus callosum and behavioral problems in children. Environ. Res., 178: 108734.

Ostro, B., Chestnut, L., Vichit-Vadakan, N. and Laixuthai, A. 1999. The impact of particulate matter on daily mortality in Bangkok, Thailand. J. Air & Waste Manag. Assoc., 49: 100-107.

State Bureau of Environment Protection of China. 2012. Ambient air quality standards. (GB3095-2-12).

Suades-González, E., Gascon, M., Guxens, M. and Sunyer, J. 2015. Air pollution and neuropsychological development: A review of the latest evidence. Endocrinology, 156: 3473-3482.

Sun, W. and Sun, J. 2016. Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by the cuckoo search algorithm. J. Environ. Manag., 188: 144-152.

Sun, W., Zhang, H., Palazoglu, A., Singh, A., Zhang, W. and Liu, S. 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.

Thomaidis, N.S., Bakeas, E.B. and Siskos, P.A. 2003. Characterization of lead, cadmium, arsenic, and nickel in PM2.5 particles in the Athens atmosphere, Greece. Chemosphere. 52: 959-66.

Vapnik, V. (ed.). 1995. The Nature of Statistical Learning Theory. Springer, New York.

Vapnik., V. (ed.). 1998. Statistical Learning Theory. Springer, New York.

Wang, L., Zhang, N., Liu, Z., Sun, Y., Ji D. and Wang, Y. 2015. The influence of climate factors, meteorological conditions, and boundary-layer structure on severe haze pollution in the Beijing-Tianjin-Hebei region during January 2013. Adv. Meteorol., 2014: 1-14.

Wu, X. and Kumar, V. (ed.) 2013. The top ten algorithms in data mining. CRC Press, Boca Raton, Florida, USA.

Yang, X.S. and Deb, S., 2009. Cuckoo Search via Lévy Flights. World Congress on Nature & Biologically Inspired Computing IEEE, 9-11 December 2009, Coimbatore, India, IEEE, Piscataway, NJ, pp. 1-7.

Yuan, Y., Wu, Y., Ge, X., Nie, D., Wang, M., Zhou, H. and Chen, M. 2019. In vitro toxicity evaluation of heavy metals in urban air particulate matter on human lung epithelial cells. Sci. Total. Environ., 678: 301-308.