Technological Forecasting & Social Change 146 (2019) 41–54

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

A hybrid PSO-SVM model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting

Shuixia Chen, Jian-qiang Wang*, Hong-yu Zhang

School of Business, Central South University, Changsha 410083, PR China

ARTICLE INFO

Keywords:
Short-term atmospheric pollutant concentration forecasting
Influential factors analysis
Clustering algorithm
Particle swarm optimisation
Support vector machine

ABSTRACT

Air pollution can lead to a wide range of hazards and can affect most organisms on Earth. Therefore, managing and controlling air pollution has become a top priority for many countries. An effective short-term atmospheric pollutant concentration forecasting (SAPCF) can mitigate the negative effects of atmospheric pollution. In this paper, we propose a new hybrid forecasting model for SAPCF. Firstly, we analyse the influential factors of pollutants to obtain the optimal combination of input variables. Secondly, we use a clustering algorithm to enhance the regularity of our modelling data. Thirdly, we build a particle swarm optimisation (PSO)–support vector machine (SVM) hybrid model called PSO–SVM and perform a case study in Temple of Heaven, Beijing to test its forecasting accuracy and validate its performance against three contrastive models. The first model inputs all possible variables in equal weight without influence factor analysis. The second model integrates the same input variables used in the proposed model without clustering. The third model inputs these same variables with genetic-algorithm optimised SVM parameters. The comparison amongst these models demonstrates the superior performance of our proposed hybrid model. We further verify the forecasting results of our hybrid model by conducting statistical tests.

1. Introduction

As one of the four elements of life, air plays a major role in maintaining the ecosystem. However, human activities have seriously aggravated the degree of air pollution, thereby prompting researchers to conduct pollution analysis (Bollen, 2015) and predict pollutant concentrations (Wu et al., 2018). Given its important role in formulating effective precautionary measures, atmospheric pollutant concentration forecasting (APCF) has received much research attention (Bai et al., 2018).

Previous studies have classified APCF into short-term APCF (SAPCF) (Niu et al., 2016), medium and long term APCF. Medium and long term APCF forecasts the concentration of pollutants over a relatively long period, usually ranging from months to years (Nebenzal and Fishbain, 2018), and is mainly used for planning the distribution of industrial sites or residential areas. Meanwhile, SAPCF reveals the vital status in many basic operations and is often used in planning abatement actions and transportation networks in advance (Li and Tao, 2018; Xie et al., 2018; Zhai and Chen, 2018). SAPCF can also help governments save time in responding to pollution-related problems (Y.F. Wang et al., 2018) and help individuals prevent exposure to pollutants (Soh et al., 2018). Therefore, an accurate SAPCF is of great significance at the social and individual levels. Researchers have proposed numerous methods for SAPCF in recent years as will be discussed in Section 1.1.

1.1. SAPCF models

SAPCF can be achieved by using statistical and artificial intelligence (AI) methods. The most commonly applied statistical methods include regression methods (Kumar and Goyal, 2011), autoregressive integrated moving average (Zhang et al., 2018), projection pursuit model (Huber, 1985), principal component analysis (PCA) (Sun and Sun, 2017) and fuzzy time series analysis (Rahman et al., 2015). However, these methods have their limitations. For instance, the results of regression forecasting mainly rely on the proposed hypotheses. Therefore, the desired forecasting accuracy cannot be easily achieved given the differences amongst these hypotheses. PCA methods also greatly depend on the employed dataset and the forecasting accuracy is discounted when extreme and nonlinear values are involved. To address these shortcomings, scholars have proposed some AI methods, including artificial neural network (ANN) (Russo et al., 2015), back propagation (BP) neural network (Bai et al., 2016), wavelet neural network (WNN) (Wang et al., 2015b) and support vector machine (SVM) (Wang et al., 2015a). ANN shows a promising performance in

* Corresponding author.
E-mail address: jqwang@csu.edu.cn (J.-q. Wang).

https://doi.org/10.1016/j.techfore.2019.05.015
Received 17 September 2018; Received in revised form 10 April 2019; Accepted 8 May 2019
0040-1625/ © 2019 Elsevier Inc. All rights reserved.

fitting nonlinear variables, but the complex relationships amongst some variables can affect its performance (Bai et al., 2018). Meanwhile, traditional neural networks (NN) face some problems related to convergence and local optimisation. To address these problems, SVM is used for SAPCF given its excellent performance in dealing with small samples and in addressing global optimisation and high-dimensional feature space problems. Based on the structural risk minimisation (SRM) principle, SVM can avoid the local minima and improve its classification accuracy (Zendehboudi et al., 2018). These features have motivated the extensive implementation of SVM in pollution forecasting (Chen et al., 2010). However, SVM also has several shortcomings. For instance, this method does not work well with large sample data and is susceptible to missing data. The forecasting accuracy of SVM is also highly correlated with the choice of parameters (Hong, 2010). To address these limitations, researchers have proposed some hybrid SVM models (Wang et al., 2017) that integrate different models to improve its forecasting accuracy as will be discussed further in Section 1.2.

1.2. Hybrid SVM models for SAPCF

A hybrid model is a combination of two or more methods (Inman et al., 2013) and an appropriate combination of models can lead to an improved forecasting accuracy. Shah et al. (2018) proposed an error minimisation algorithm based on artificial bee colony by using combining NN and AI for stock market prices prediction. Their experiment results reveal that their proposed hybrid model has a higher forecasting performance compared with conventional methods. Two hybrid SVM methods are commonly used for SAPCF. Firstly, SVM can be combined with other intelligent methods to address its most common problems. For example, Ghaemi et al. (2015) proposed a distributed computing based on the Hadoop platform to overcome the inability of SVM in working well with massive data. However, this algorithm combination faces some limitations. When dealing with large-scale data, the commonly used method for data splitting node selection is subjective. Moreover, some hybrid models need to undergo two stages before obtaining the final result, thereby leading to information loss and time waste. Secondly, other algorithms are combined with SVM to optimise its parameters. Some of the most widely used algorithms include the genetic algorithm (GA) (Garg, 2015), ant swarm optimisation (Zheng et al., 2008), cuckoo search algorithm (Sun and Sun, 2017) and gravitational search algorithm (Garg, 2019). However, these algorithms are unable to store the best particle information (Barman et al., 2018) whilst other algorithms take too long to optimise, thereby limiting their potential to be combined with SVM.

To address these problems, a new hybrid SVM model for SAPCF needs to be proposed. Given the influence of certain parameters on SVM, this study introduces the particle swarm optimisation (PSO) algorithm (Patwal et al., 2018) into SVM to optimise its parameters and improve its forecasting fitness. PSO can be easily implemented and only uses two parameters, namely, the position and velocity of particles (Garg, 2016). PSO also shows excellent performance in achieving algorithm convergence and global optimisation (Xiao et al., 2017), thereby motivating us to use this algorithm to optimise the parameters of SVM. We build a hybrid PSO-optimised SVM parameters (PSO–SVM) model as will be discussed in Section 1.3.

1.3. Hybrid PSO–SVM model based on clustering algorithm for SAPCF

Previous studies have reported a high correlation between the concentration of atmospheric pollutants and some meteorological variables (Whiteman et al., 2014), that is, considering these meteorological variables would significantly improve the SAPCF accuracy. However, various meteorological factors may generate different effects on the concentration of atmospheric pollutants. For example, high-intensity wind can send atmospheric pollutants over long distances. Therefore, wind speed and direction are relatively more important compared with other meteorological factors. The same meteorological factor may also demonstrate various effects across different regions. Therefore, the impact of meteorological variables on the concentration of atmospheric pollutants must be investigated. Although previous studies have considered the influence of meteorological variables (Cortina-Januchs et al., 2015), only few have specifically examined the effect of each meteorological variable. To address this gap, we design an influential factor analysis in our proposed PSO–SVM model to investigate the dependency and direction of dependency between the influential factors and the forecasting target as well as to obtain the optimal combination of influential variables (i.e. the combination of variables and the weight of each variable). To ensure the high accuracy and reliability of a forecasting model, the regularity of the modelling data must be improved. Therefore, we perform unsupervised clustering to classify our dataset into several categories, with all the data clustered into a single category sharing the same characteristics. As mentioned above, SVM parameter optimisation plays a key role in SAPCF. Therefore, we use PSO to obtain the parameters of SVM. In sum, our proposed hybrid PSO–SVM model includes three processes, namely, influential factor analysis, data clustering and forecasting.

Given the advantages of the aforementioned methods and the importance of each process, the main objective of this paper is to propose a new hybrid PSO–SVM model to improve the accuracy of SAPCF. This hybrid model performs an influential factor analysis to determine the optimal combination of influential factors for specific regions. We also introduce the clustering method into our hybrid model to strengthen the regularity of data. The small datasets obtained by clustering meet the demand of SVM for a small sample data volume. Given the ability of PSO to optimise the SVM parameters, the hybrid PSO–SVM model addresses the inherent limitations of SVM. We test the accuracy and stability of this model by performing a comparative analysis, which highlight the advantages of our hybrid model in terms of forecasting accuracy and runtime. We also perform additional statistical tests to confirm this conclusion.

The rest of this study is organised as follows. Section 2 briefly discusses the data collection and pre-processing. Section 3 discusses the influential factor analysis. Section 4 presents the architecture of the proposed hybrid model. Section 5 performs a case study, comparison analysis and statistical tests. Section 6 presents some discussions of the results. Section 7 concludes the paper.

2. Data collecting and pre-processing

This section introduces the sources and pre-processing of the data used in this study.

2.1. Data sources

Each country adopts a unique method and set of criteria for monitoring and evaluating air quality. For instance, China adopts the National Ambient Air Quality Standards (GB3095-2012) (Duan et al., 2014) to convert its pollutant monitoring values into simple conceptual numerical forms, such as PM2.5, PM10, SO_2, NO_2, O_3 and CO, and uses the air quality index (AQI) (Xie et al., 2018) as a quantitative indicator for describing the status of air quality. Therefore, we use AQI and these six numerical forms as influential variables for SAPCF. Meteorological variables also play a vital role in SAPCF (Cortina-Januchs et al., 2015). Therefore, we choose five meteorological variables, namely, temperature, relative humidity, windspeed, wind direction and air pressure, as influential factors that may affect air pollution.

The data used in this paper are collected from the website of the Beijing Municipal Environmental Monitoring Center (http://www.bjmemc.com.cn), which collects its data through its multiple monitors. The air quality automatic monitoring system of Beijing covers 35 monitoring points, including the Dongsi Subdistrict, Temple of Heaven,

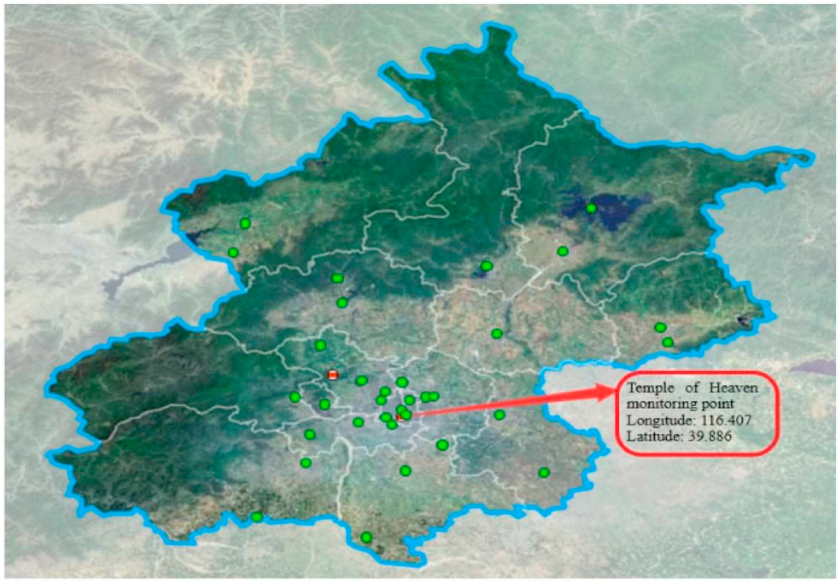

Fig. 1. Satellite image of monitoring points in Beijing.

West Park officials and so on. Fig. 1 presents the location of each of these monitoring points on a satellite map. We collected data on the six pollutant concentrations, AQI and the five meteorological variables of the Temple of Heaven monitoring point from January 1, 2017 to December 31, 2017. The Temple of Heaven monitoring point is located in the Dongcheng district of Beijing (116.407° longitude and 39.886° latitude). To simplify our data collection procedure without losing any important information, we collected our data at three-hour intervals starting from 2:00 (i.e. 2:00, 5:00, 8:00, 11:00, 17:00, 20:00 and 23:00) every day. We eventually collected 2921 pieces of data in 365 days, with each piece of data having 14 attributes, including date, time, pollutant concentration and so on.

2.2. Data pre-processing

We prepare the data after their collection. Data pre-processing is a crucial step in data processing. Previous studies show that preparing the raw data can help improve their processing accuracy (Coussement et al., 2017; Y. Wang et al., 2018a). Our main data pre-processing methods include data cleaning and data transformation, which will be explained in the following subsections.

2.2.1. Data cleaning

We initially perform data cleaning to maintain the consistency of our dataset. The raw data should be cleaned before the forecasting given that the data obtained from monitoring instrument may be incomplete due to power or instrument failures. These data may also show some noise, redundancy and inconsistencies. Data cleaning involves two procedures, namely, consistency checks and missing value processing.

(1) Consistency checks

Consistency checks examine the relationship amongst the data based on the value range and correlation of each variable. Through this procedure, we can identify those values that exceed the normal range or contradict one another. Consistency checks can be classified into range, time and variable consistency checks. Range check finds those values that exceed a specified measurement range based on the scope of monitoring value (Schlechtingen et al., 2013). Following the standards of HJ 653-2013 and HJ 654-2013 (Zeng et al., 2015), we obtain the monitoring range of each variable as shown in Table 1.

The statistics of the variables included in the collected dataset are presented in Table 2. A comparison between Tables 1 and 2 reveal that the maximum monitoring value of PM10 is outside the specified monitoring range.

The monitoring values of PM10 are presented in detail in Fig. 2. We deleted those values that lie outside the specified scope.

The change of the variable has a certain regularity with time (Rohde

Table 1
Monitoring range of each variable.

Variable	Unit	Range
PM2.5	$\mu g/m^3$	0~1000$\mu g/m^3$
PM10	$\mu g/m^3$	0~1000$\mu g/m^3$
AQI	NA	≥0
SO_2	$\mu g/m^3$	0~1428$\mu g/m^3$
NO_2	$\mu g/m^3$	0~1026$\mu g/m^3$
O_3	$\mu g/m^3$	0~1071$\mu g/m^3$
CO	mg/m^3	0~62.5mg/m^3
Temperature	$^{\circ}C$	$-50\sim+80\,^{\circ}C$
Relative humidity	NA	0~100%
Windspeed	m/s	0~75m/s
Wind direction	°	0~360°
Air pressure	$mmHg$	412.5339~795.065$mmhg$

Table 2
Statistics of the variables.

Variables	Minimum	Maximum	Mean
PM2.5	2.0	734.0	59.061
PM10	0.0	1010.0	88.180
AQI	2.0	507.0	88.394
SO_2	1.0	111.0	6.920
NO_2	0.0	336.0	45.389
O_3	1.0	325.0	55.705
CO	0.0	12.7	0.977
Temperature	−9.9	38.0	14.240
Relative humidity	5.0	97.0	49.878
Windspeed	0.0	8.0	2.097
Wind direction	0.0	3.0	1.862
Air pressure	739.9	776.8	758.264

The red box in the table highlights that the maximum monitoring value of PM10 is outside the specified monitoring range.

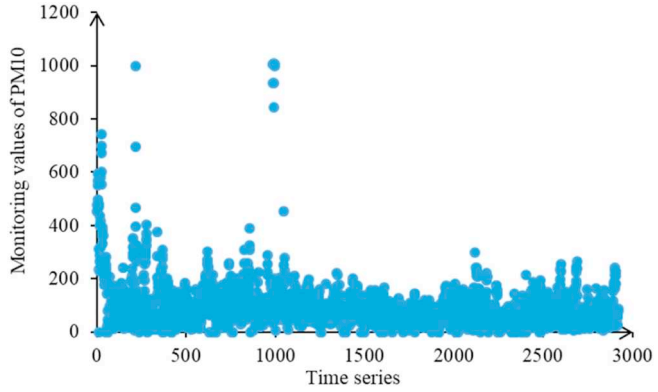

Fig. 2. Monitoring values of PM10.

and Muller, 2015). Therefore, each variable must be examined at the time scale. The time consistency checks aim to discover outliers in time series. To determine how these variables change over time, we calculate the absolute deviation between the adjacent variables. Fig. 3 presents the results. The general ratios of the adjacent monitoring value deviations less than $40\mu g/m^3$, less than $60\mu g/m^3$ and greater than $100\mu g/m^3$ are greater than 80%, greater than 90% and less than 1%, respectively. Given that the magnitude of the monitoring value for CO at adjacent times differs from that of the other monitoring values, the results for CO are reported in a separate figure (Fig. 4). The huge gap in the deviation of meteorological data cannot be presented in the same graph and is therefore not reported in this paper.

We use the data for the adjacent time points as a reference in determining the maximum variation range at a specified moment. We utilise the following mathematical model:

$$\begin{Bmatrix} |y(d,t) - y(d,t-1)| > \theta_1 \\ |y(d,t) - y(d,t+1)| > \theta_2 \end{Bmatrix}, \tag{1}$$

where $y(d,t) = \frac{y(d,t-1)+y(d,t+1)}{2}$ represents the value of variable d at time t whilst θ_1 and θ_2 denote the threshold values. We determine the threshold of different variables according to the rule of 3σ (Yousefzadeh et al., 2017), that is, any data located outside the specified range will be treated as poor data. Then we adopt the average value to ensure a smooth calculation.

The variable consistency checks examine whether the relations amongst elements conform to objective laws (Wang, 2011). Certain correlations can be observed amongst the variables used in this study. For example, PM2.5 and PM10 are positively related to CO, NO_2 and SO_2 whilst O_3 is negatively correlated with the other gaseous pollutants (Abdul-Wahab et al., 2005). The correlations amongst these variables can be ascribed to the sources of pollutants and the weather conditions and often demonstrate strong local and seasonal characteristics. We adopt the robust regression method to perform the variable consistency checks based on the positive correlation between NO_2 and CO. In the regression analysis, the residual error of the fitting value between the dependent variable and the actual value can be used as a criterion in the outlier test. We perform M estimation (Zou et al., 2000) to calculate the regression parameters. However, given that these procedures are not the focus of our study, we do not explain their principles in this paper.

(2) Missing value processing

We deal with the missing values that may be included in the collected dataset in two ways. Firstly, we fill a large number of missing values by crawling data from other monitoring sites or by collecting data from adjacent areas at the same time. Secondly, if the time interval of missing data is not large, then we use the following linear interpolation method to fill the gap:

$$S_{t+j} = S_t + \frac{S_{t+i} - S_t}{i} \cdot j,\ 0 < j < i, \tag{2}$$

where S_t and S_{t+i} denote the monitoring values of time t and $t+i$, respectively.

2.2.2. Data transformation

To avoid the impacts of different dimensions, the raw dataset needs to be normalised. We adopt the following data transformation methods to this end:

(1) Normalise data to [0,1].

$$\hat{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{3}$$

where $x_{\max}$ and $x_{\min}$ represent the maximum and minimum values of the raw dataset. x_i denotes the actual data and $\hat{x}_i$ denotes the normalised data. All notations presented below follow the same definition.

(2) Normalise data to [−1,1].

$$\hat{x}_i = \frac{x_i - 0.5(x_{\max} + x_{\min})}{0.5(x_{\max} - x_{\min})}. \tag{4}$$

(3) Convert the data into zero mean value and single azimuth.

$$\hat{x}_i = \frac{x_i - \mu}{\sigma^2}, \tag{5}$$

where μ and σ^2 represent the mean and variance of the original data, respectively.

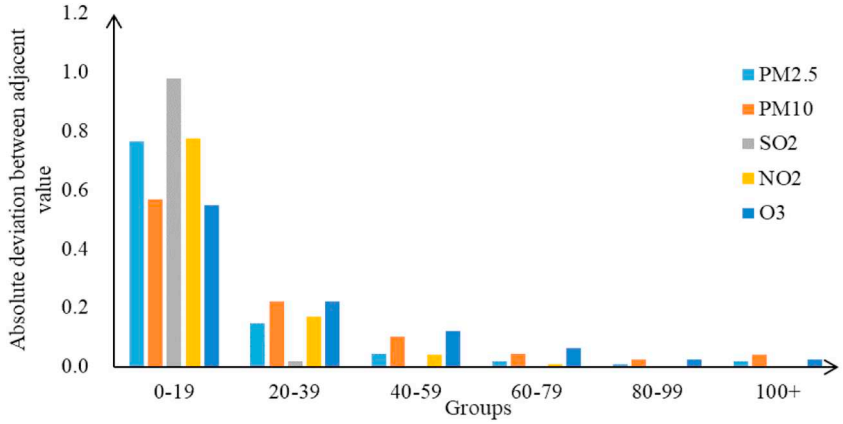

Fig. 3. Absolute deviation between the adjacent values of PM2.5, PM10, SO_2, NO_2 and O_3.

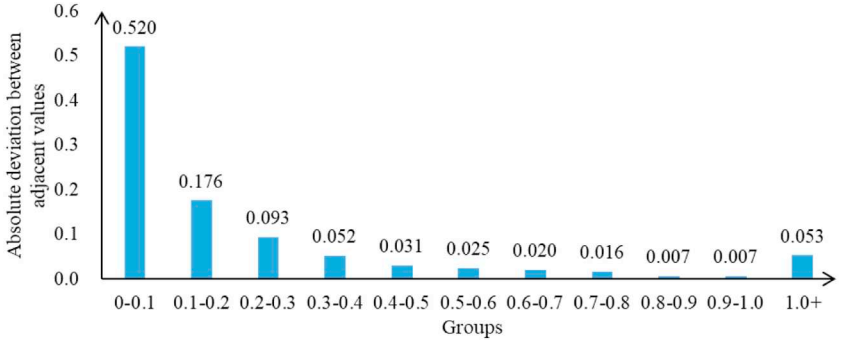

Fig. 4. Absolute deviation between the adjacent values of CO.

3. Influential factors analysis for SAPCF

Given the impacts of meteorological factors and historical pollutant data, we choose the six pollutants, AQI and five meteorological factors as influential factors. As mentioned above, the roles of various factors in the study area must be examined in detail. We use the following simple linear correlation coefficient for the preliminary analysis:

$$r = \frac{Cov(x,y)}{\sqrt{D(x)D(y)}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \quad (6)$$

where $\bar{x}$, $\bar{y}$ denote the mean values of x and y, respectively, and r denotes the correlation coefficient (Zhou et al., 2016). The correlation coefficients between the variables are presented in Table 3.

We used AQI as an example to demonstrate the correlation analysis process. Here, AQI is treated as a dependent variable whilst all the other variables are treated as independent variables (i.e. input variables). Table 3 shows a zero correlation between AQI and air pressure, thereby indicating the absence of any correlation between these factors. If we set the confidence coefficient to 95%, then AQI shows positive correlations with PM2.5, PM10, SO_2, NO_2, CO and relative humidity but shows negative correlations with O_3, temperature, windspeed and wind direction. After obtaining the correlations between each independent and dependent variable, we attempt to establish a concrete relationship between these variables for further analysis. Accordingly, we perform a stepwise regression analysis on the AQI and other related variables by using IBM SPSS Statistic. The model summary is presented in Table 4. R denotes the goodness of fit. An R value closer to 1 indicates a better model. In addition, the R square represents the decision coefficient and the last column of Table 4 presents the estimated random error.

Table 4 shows that the R values of cases 5, 6 and 7 are equal and better than those of the other cases whilst the error of case 7 is less than those of cases 5 and 6. Therefore, the coefficient combination in case 7 serves as an excellent representation of the relationship between the independent and dependent variables. The coefficients of case 7 are presented in Table 5, where the second and third columns present the partial regression coefficients whilst the fourth column presents the standard regression coefficient. $Y = 0.685x_1 + 0.370x_2 + 0.045x_3 + 0.016x_4 + (-0.078)$ $x_5 + 0.010x_6 + 0.023x_7$ can be obtained, with $x_i(i = 1, \ldots, 7)$ representing the variables in the first column of the table (with numbers 1 to 7 denoting PM2.5, PM10, SO_2, NO_2, CO, relative humidity and windspeed, respectively) and Y denoting the value of AQI. Column 6

Table 4
Model summary of the stepwise regression analysis on AQI and the other variables.

Case	R	R square	Adjusted R square	Error of the estimate
1	0.945[a]	0.894	0.894	23.635
2	0.969[b]	0.939	0.939	12.890
3	0.970[c]	0.942	0.942	17.531
4	0.971[d]	0.943	0.943	17.269
5	0.972[e]	0.944	0.944	17.187
6	0.972[f]	0.944	0.944	17.160
7	0.972[g]	0.944	0.944	17.154

Letters in the table represent the different combinations of independent variables.

Table 5
Coefficients of case 7.

Variable	B	Std. error	Beta	t	Sig.
Constant	10.187	1.311		7.770	0.000
PM2.5	0.789	0.013	0.685	59.917	0.000
PM10	0.332	0.008	0.370	43.776	0.000
CO	−5.608	0.574	−0.078	−9.773	0.000
SO_2	0.387	0.043	0.045	8.935	0.000
Windspeed	1.225	0.282	0.023	4.337	0.000
O_3	0.021	0.006	0.016	3.330	0.001
Relative humidity	0.028	0.017	0.010	1.688	0.092

Table 3
Correlation coefficients between the variables.

	PM2.5	PM10	AQI	SO_2	NO_2	O_3	CO	Temperature	Relative humidity	Windspeed	Wind direction	Air pressure
PM2.5	1	0.842	0.945	0.375	0.637	−0.146	0.820	−0.158	0.356	−0.225	−0.110	0.039
PM10	0.842	1	0.911	0.359	0.515	−0.053	0.652	−0.084	0.171	−0.090	−0.086	−0.023
AQI	0.945	0.911	1	0.409	0.579	−0.076	0.732	−0.112	0.265	−0.142	−0.096	0.000
SO_2	0.375	0.359	0.409	1	0.382	0.003	0.307	−0.169	−0.165	0.020	0.018	0.115
NO_2	0.637	0.515	0.579	0.382	1	−0.530	0.674	−0.370	0.205	−0.383	−0.040	0.218
O_3	−0.146	−0.053	−0.076	0.003	−0.530	1	−0.268	0.613	−0.275	0.388	−0.113	−0.486
CO	0.820	0.652	0.732	0.307	0.674	−0.268	1	−0.294	0.363	−0.258	−0.040	0.168
Temperature	−0.158	−0.084	−0.112	−0.169	−0.370	0.613	−0.294	1	0.122	0.112	−0.094	−0.847
Relative humidity	0.356	0.171	0.265	−0.165	0.205	−0.275	0.363	0.122	1	−0.492	−0.046	−0.227
Windspeed	−0.225	−0.090	−0.142	0.020	−0.383	0.388	−0.258	0.112	−0.492	1	0.186	−0.012
Wind direction	−0.110	−0.086	−0.096	0.018	−0.040	−0.113	−0.040	−0.094	−0.046	0.186	1	0.115
Air pressure	0.039	−0.023	0.000	0.115	0.218	−0.486	0.168	−0.847	−0.227	−0.012	0.115	1

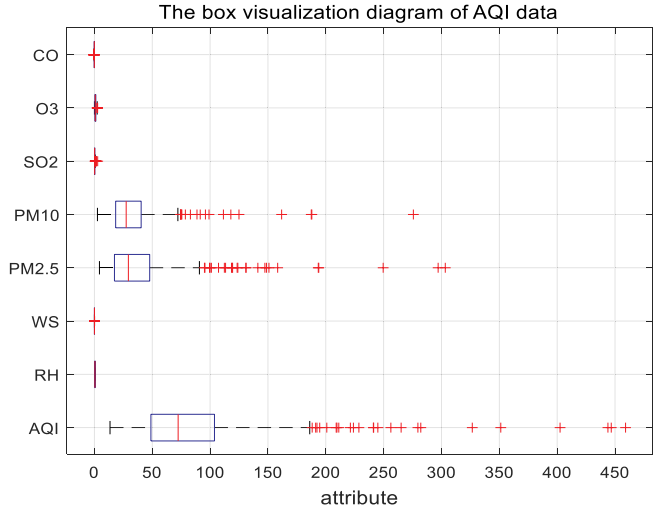

Fig. 5. Box diagram of AQI and the other relevant variables.

and 7 present the results for the other parameters used in the stepwise regression analysis. The relationships between AQI and the other relevant variables are illustrated in Fig. 5. The detailed relationships between each independent and dependent variable are illustrated in Fig. 6, where attributes 1 to 8 represent the eight related variables, including AQI.

4. Architecture of the proposed hybrid PSO–SVM model based on K-means

After obtaining the optimal combination of input variables, we build the architecture of the hybrid PSO–SVM model based on K-means. This section introduces SVM, PSO, and the hybrid PSO–SVM model based on K-means.

4.1. Support vector machine

Support vector machine (SVM) is a common discriminant method (Cortes and Vapnik, 1995) that follows the SRM principle and presents unique advantages in dealing with small sample and high-dimensional feature space problems. Initially applied to address pattern recognition problems, SVM is now being used to deal with nonlinear regression estimation problems by introducing the insensitive loss function ε. The SVM employed for addressing regression problems is called support vector regression (SVR), which main idea is to map the dataset $x_i(i = 1, \ldots, n)$ to a higher-dimensional feature space by employing a nonlinear function. The relationship between the values can be by expressed as

$$f(x) = \omega^T \phi(x) + b, \tag{7}$$

where $f(x)$ is the output value, ω and b are the coefficients and $\phi(x)$ is the nonlinear mapping function that can transform the input values into a high-dimensional feature space. The regulated values of ω and b can be obtained as

$$\begin{aligned} &\underset{\omega,b,\xi*,\xi}{Min}\, R_\varepsilon(\omega, \xi*, \xi) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}(\xi_i^* + \xi_i) \\ &\begin{cases} y_i - \omega^T\phi(x_i) - b \le \varepsilon + \xi_i^*, \ i = 1, 2, ..., n \\ -y_i + \omega^T\phi(x_i) + b \le \varepsilon + \xi_i, \ i = 1, 2, ..., n \\ \xi_i, \xi_i^* \ge 0, \ i = 1, 2, ..., n \end{cases} \end{aligned} \tag{8}$$

where $R_\varepsilon()$ is the empirical risk (Barman et al., 2018), C is the regularisation parameter, ξ_i^* represents the errors above ε and ξ_i represents the errors below $-\varepsilon$. The above function represents a quadratic optimisation problem that can be transformed into a dual problem. The final equation for SVM is

$$f(x) = \sum_{i=1}^{n}(\beta_i^* - \beta_i)K(x_i, x_j) + b, \tag{9}$$

where β_i^*, β_i are the *Lagrangian* coefficients and $K(x_i, x_j)$ is the kernel function of SVM that represents the inner product of two vectors. The kernel function of vectors x_i and x_j is defined as

$$K(x_i, x_j) = \phi(x_i)\cdot\phi(x_j). \tag{10}$$

Several kernel functions are available, including the linear kernel function and Gaussian kernel function. Amongst these functions, the Gaussian kernel function is one of the most popular and is also called the radial basis function (RBF). This function can map data to an infinite dimension with less computational complexity. Therefore, we use

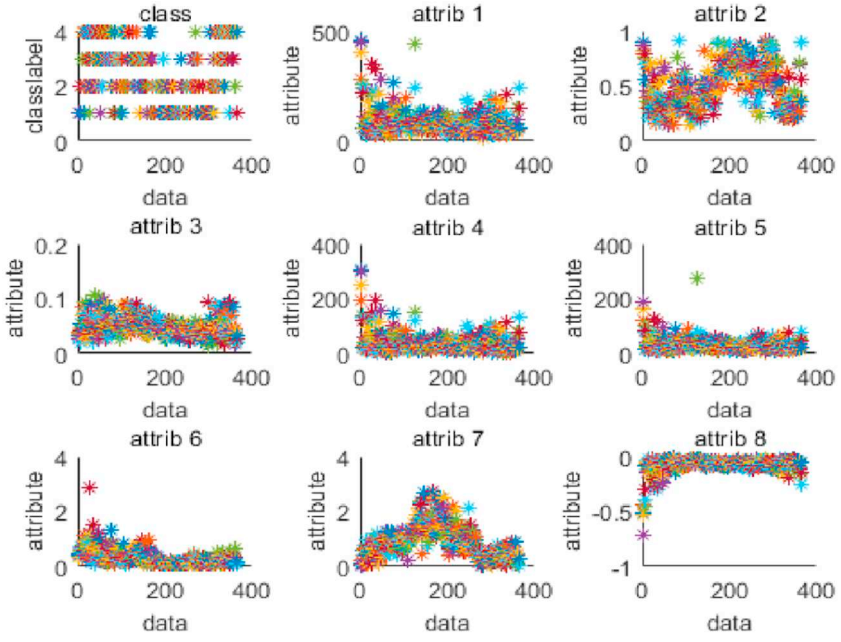

Fig. 6. Relationships between AQI and each variable.

RBF for SVM and define this function as

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\gamma^2}\right), \tag{11}$$

where γ is the Gaussian parameter. The appropriate combination of SVM parameters (C, $\varepsilon$and γ) plays a critical role in achieving a high forecasting accuracy. Therefore, we use the PSO algorithm to determine the appropriate parameters. This algorithm is discussed further in the following subsection.

4.2. Particle swarm optimisation

Particle swarm optimisation (PSO) is a swarm computing technology developed on the basis of iterative optimisation. This algorithm initialises a set of particles and then updates the velocity and position of these particles in the next iteration by tracking two extremum values, namely, the individual extreme value P_{ibest} and the global extreme value P_{gbest}. After discovering these two extremities, PSO identifies the speed and distance of each particle.

Suppose that there is a population of m particles in a $d-dimensional$ search space. The $i-th$ particle is represented as $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$, $i = 1, 2, \dots, m$. In other words, the position of the $i-th$ particle is x_i. The velocity of the $i-th$ particle is also a vector denoted by $v_i = (v_{i1}, v_{i2}, \dots, v_{id})$. The optimal position of this particle is $p_i = (p_{i1}, p_{i2}, \dots, p_{id})$ whilst that of the whole population is $p_g = (p_{g1}, p_{g2}, \dots, p_{gd})$. The standard PSO algorithm updates v_i and x_i as

$$\begin{aligned} v_{i,k+1}^d &= \bar{\omega} v_{i,k}^d + c_1 r_1 (P_{i,k}^d - x_{i,k}^d) + c_2 r_2 (P_{g,k}^d - x_{i,k}^d) \\ x_{i,k+1}^d &= x_{i,k}^d + v_{i,k+1}^d \end{aligned}, \tag{12}$$

where $\bar{\omega}$ is the weight coefficient of inertia, c_1, c_2 are two non-negative constants called acceleration constants and r_1, r_2 are random numbers that are uniformly distributed within $[0, 1]$.

The main problem of PSO lies in the premature convergence in optimisation (Jordehi, 2015). We use an improved PSO algorithm to control the characteristics of the population. To prevent falling into the local optimum, we introduce the following average grain spacing in choosing the initial population:

$$D(t) = \frac{1}{mL} \sum_{i=1}^{m} \sqrt{\sum_{d=1}^{n} (p_i^d - \overline{p^d})^2}, \tag{13}$$

where L is the maximum length of the search space diagonal, n denotes the dimension of the solution space, P_i^d represents the $d-dimensional$ coordinate value and $\overline{P^d}$ represents the mean value of P_i^d. The average particle spacing represents the distribution dispersion degree of each particle. A smaller $D(t)$ corresponds to a higher population concentration, and vice versa.

Judging premature convergence is critical in dealing with premature convergence problems. Given that the fitness of a particle is largely determined by its position, the current status of the population can be determined according to the overall change in the fitness of all particles. We denote the current fitness by f_i and the current average fitness by $\overline{f}$ and then define the fitness variance of the population as

$$\lambda^2 = \sum_{i=1}^{m} \left(\frac{f_i - \overline{f}}{f}\right)^2, \tag{14}$$

where f represents the normalisation scaling factor that is employed to limit the size of λ^2. We compute f as

$$f = \begin{cases} \max |f_i - \overline{f}|, \max |f_i - \overline{f}| > 1 \\ 1, \ else \end{cases}, \tag{15}$$

where λ^2 represents the aggregation degree of particles. A smaller λ^2 corresponds to a greater aggregation degree, and vice versa. As the number of iterations increases, the fitness becomes closer and smaller, that is, the value of λ^2 gradually decreases. Meanwhile, when $\lambda^2 \prec \beta$ (where β is a given threshold value), the algorithm enters the later search stage.

4.3. Hybrid PSO–SVM model based on K-means

To enhance the regularity of the data and shorten the forecasting time, we employ the K-means algorithm (Hartigan and Wong, 1979) to incorporate the similar meteorological variables. As a classic distance-based clustering algorithm, K-means has been widely used in various fields of forecasting (Benmouiza and Cheknane, 2013). The process of this algorithm is described as follows. Firstly, k points are randomly selected as the centre of the initial cluster. Secondly, other points are allocated to their nearest centre to form the initial cluster. Thirdly, the mean value of all points in each cluster is calculated. This mean value is taken as the new centre point and the other points are redistributed to the nearest centre point. This procedure is repeated until the centre of each cluster no longer changes. The key step in K-means is determining the cluster number k. We identify k when the Euclidean distance between each point and its clustering centre stops showing significant changes. Take the AQI as an example. This index is related to two meteorological variables, namely, relative humidity and windspeed. The Euclidean distance between each point and its centre obtained by K-means clustering is presented in Fig. 7. This distance is significantly

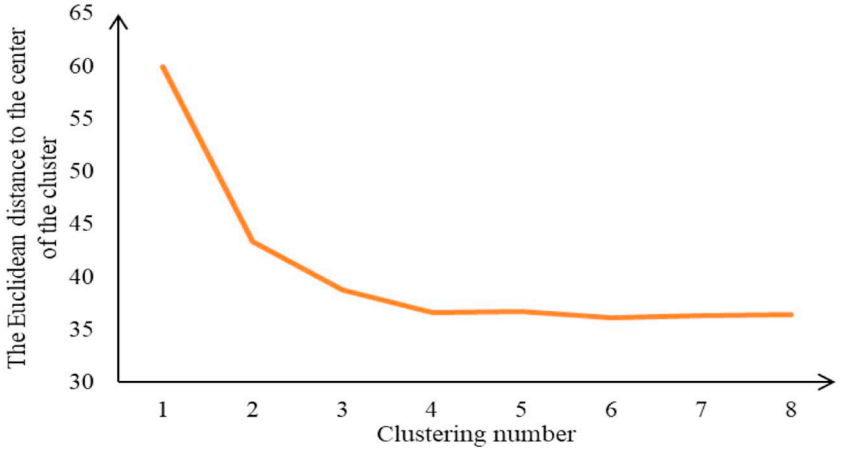

Fig. 7. Euclidean distance between points and their centres.

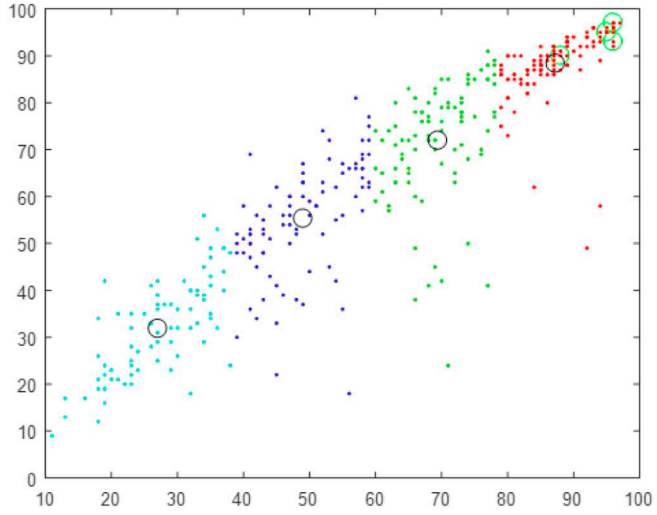

Fig. 8. Clustering results of $k = 4$.

reduced until reaching $k = 4$. After reaching this point, the distance no longer shows significant changes as k increases. Therefore, we choose $k = 4$ as our cluster number. The clustering results of $k = 4$ are illustrated in Fig. 8, where each colour denotes different clusters and the bold circles in each class denotes the centre of the cluster. To highlight the differences amongst various clusters, we choose the first two columns of data as the x and y axes in this figure. The forecasting values for the meteorological data used in this paper are obtained from the weather forecast provider.

In the above analysis, we obtain the optimal combination of input variables and new classes of training data. We then develop a new hybrid PSO–SVM model to obtain the forecasting results. We use PSO to obtain the optimal parameters of SVM. Fig. 9 illustrates the process of the proposed PSO–SVM model in detail. The particle population is initialised before setting the population size m, the initial and final inertial weight values $\overline{\omega}_{\max}$, $\overline{\omega}_{\min}$, the acceleration constants c_1 and c_2, the maximum evolutionary algebra $T_{\max}$ or iterative termination threshold. We set $m = 20$, use [0.4,0.9] as the value range for $\overline{\omega}$ and set the maximum evolutionary algebra $T_{\max} = 200$. To balance the impacts of random factors, we set the initial value of c_1 and c_2 to 2. The values of SVM parameters C, ε and γ can be obtained through an automatic optimisation.

5. Case study and forecasting result analysis

Many performance indexes can be used to evaluate the SAPCF performance of the proposed hybrid model. We select mean absolute percentage error (MAPE) to test the forecasting accuracy of our proposed hybrid model. MAPE represents the average of N absolute percentage error and can be computed as

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{A_i - F_i}{A_i}\right| \times 100\%, \tag{16}$$

where N represents the number of time instances, and A_i, F_i represent the actual and forecasting data, respectively.

We take AQI as an example to demonstrate the forecasting process. The values of AQI from January 1, 2017 to December 31, 2017 are presented in Fig. 10. We input the following variables into the hybrid model: (I) AQI at the same hour in a similar cluster, (II) relative humidity at the same hour in a similar cluster, (III) windspeed at the same hour in a similar cluster, (IV) PM2.5 at the same hour in a similar cluster, (V) PM10 at the same hour in a similar cluster, (VI) SO_2 at the same hour in a similar cluster, (VII) O_3 at the same hour in a similar cluster and (VIII) CO at the same hour in a similar cluster.

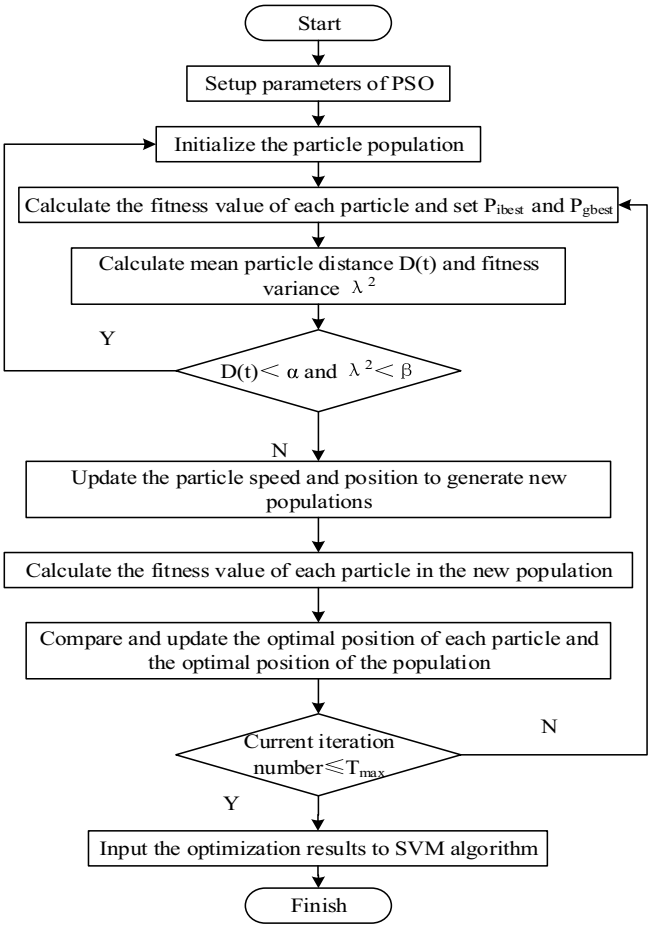

Fig. 9. Optimisation process of the PSO–SVM model.

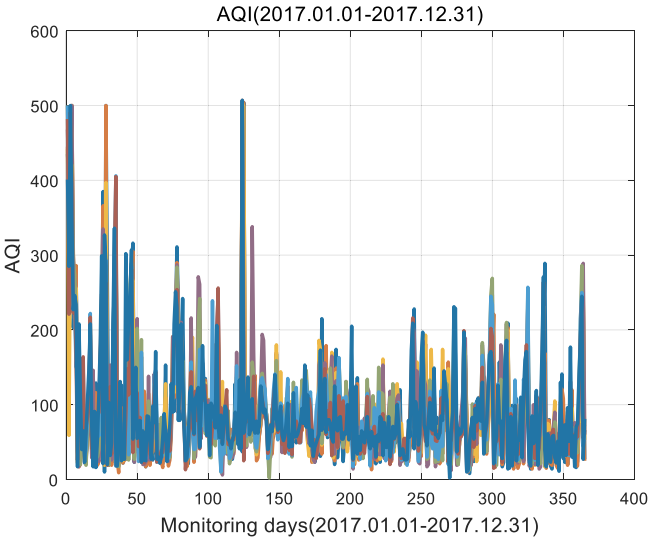

Fig. 10. AQI values from January 1, 2017 to December 31, 2017.

We perform three tests to highlight the performance of our proposed hybrid model. Firstly, we input all possible variables in equal weight into the proposed model without analysing the influential factors. Secondly, we input the data analysed by the influential factors into the PSO–SVM model without performing the clustering process. Thirdly, we input the same data used in the proposed model into the GA-optimised SVM parameters (GA–SVM) model. The proposed model and the three

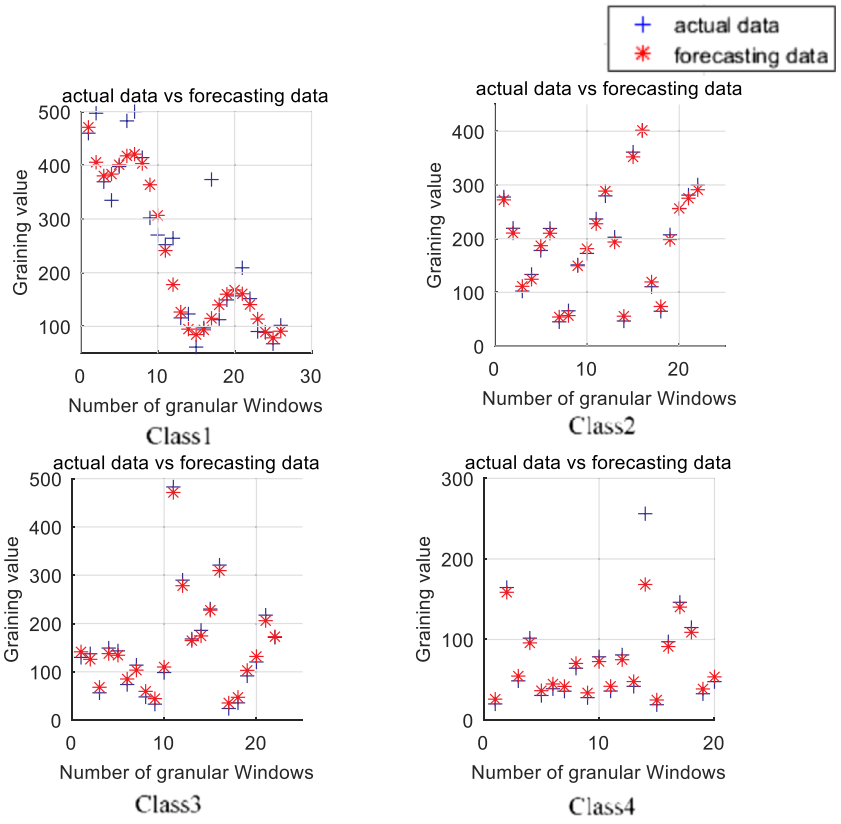

Fig. 11. Forecasting results of Model1.

other models used in these tests are referred to as Model1, Model2, Model3 and Model4 in the following subsections, respectively.

The forecasting results of Model1 are presented in Fig. 11. All data including the following contrastive models are granulated before forecasting, which results in the scale of abscissa. This figure roughly illustrates the fitness of the forecasted and actual values in each class. A detailed analysis and performance comparison with the other models are presented in the following subsections.

5.1. Forecasting results of contrastive models

The forecasting results of the contrastive models are presented in this subsection. To highlight the effects of each step in the proposed model, we control one part of the model in each comparative analysis. Without analysing the influential factors, Model2 shows the influence of the related variables on forecasting performance. We input all monitored variables into this model without analysing the correlations and weight distribution. The forecasting results of Model2 are presented in Fig. 12. The performance of Model1 and Model2 cannot be easily evaluated by merely comparing Figs. 11 and 12. Therefore, specific forecasting effects are compared in the following subsection. To justify the validity of the clustering process, Model3 is implemented in an environment without clustering analysis. The forecasting results of this model are illustrated in Fig. 13. A comparison of Model1 and Model3 reveals that the forecasting value of the latter is not consistent with the actual value. Based on the fitting degree shown in the above figures, we can roughly conclude that the forecasting performance of Model3 is worse than that of Model1.

After discussing the effects of influencing factors analysis and clustering, we analyse the impact of the parameter optimisation algorithm on forecasting accuracy. In Model4, we input the same variables and data used in Model1 into the GA–SVM model. The forecasting results of Model4 are illustrated in Fig. 14. We adopt different ways to display the results of Model4 due to the differences in the properties of the algorithms. The ordinate of Fig. 14 represents the class label whilst its abscissa represents the number of training sets. Similar to Model2, we cannot directly compare the forecasting performance of Model4 and Model1. In the following subsection, we use forecasting error and training time to illustrate the forecasting accuracy of each model.

5.2. Comparison of the proposed model and contrastive models

This section compares the proposed model with the contrastive models. Although the forecasting performance of these models has been discussed in the previous subsection, their similarities and differences remain unclear. We discuss in this subsection the forecasting errors and training time of these models.

Table 6 presents the forecasting errors of all models. Each value in columns 2 to 5 of Table 6 represents the MAPE in the corresponding cluster of models. The value in the MAPE column is calculated from the arithmetic mean of MAPE in the previous clusters. No MAPE value is found under each cluster in Model3 due to the absence of a clustering process. The difference in MAPE value is calculated by subtracting the MAPE value of Model1 from those of the other models. A comparison of the results presented in columns 2 to 5 reveals that the MAPEs of Model1 are always significantly better than those of Model4. Moreover, although the gap between Model2 and Model1 is small, Model1 clearly outperforms Model2. The positive value of MAPE difference also confirms that Model1 outperforms all the other models.

The granular window forecasting error of each model is illustrated in Fig. 15. The error range of Model1 is $[-0.7\%, 0.7\%]$. If we set a benchmark $[-0.7\%, 0.7\%]$ of errors, the most consistent performers

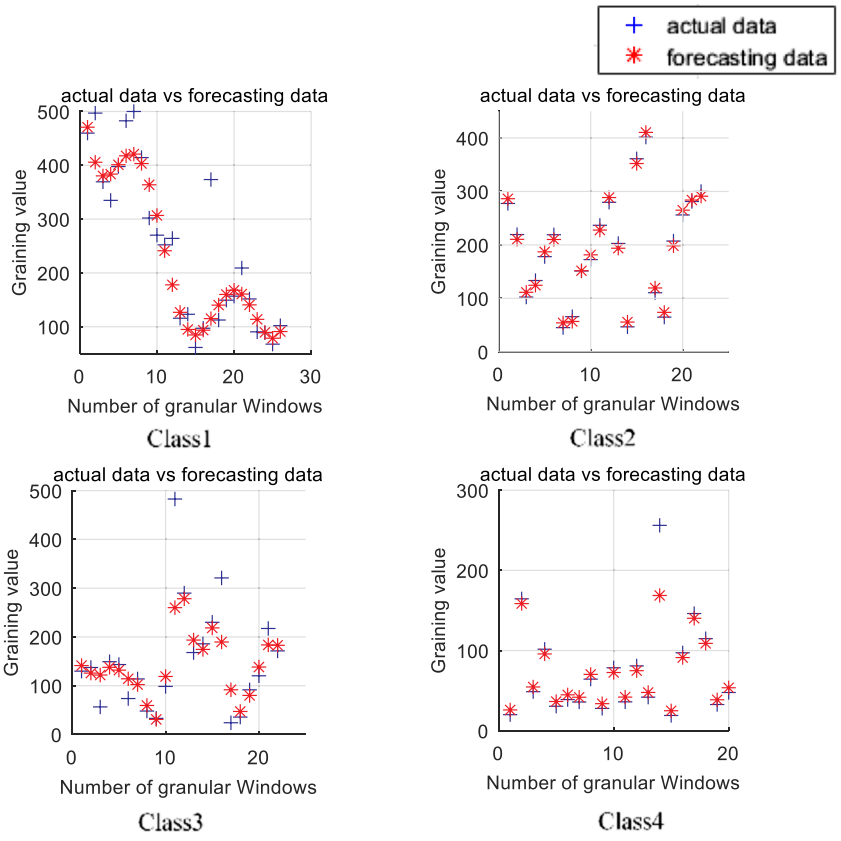

Fig. 12. Forecasting results of Model2.

can be displayed intuitively amongst the employed models. All granular windows in Model1 are within the range $[-0.7\%, 0.7\%]$ whilst only 93%, 81% and 90% of the granular windows in Model2, Model3 and Model4 fall within this range, respectively. Some forecasting errors in the contrastive models even exceed 1.5% and far exceeds the range of 0.7% in Model1. In other words, the error range of the proposed model is minimal amongst all compared methods. Therefore, our proposed model is relatively accurate and stable.

Some differences can also be observed in the training time of these four models (25, 53, 106 and 36 s for Model1, Model2, Model3 and Model4, respectively). The time spent on training data shows that the proposed model outperforms all the other models. Model2 spends more time than Model1 as it considers all variables. Meanwhile, Model3 has a longer forecasting process compared with the other models as it needs to read all information given its lack of a clustering process. Such long forecasting time may also affect the forecasting accuracy due to some extreme data. The time spent by Model4 does not greatly differ from that of Model1, but this model shows a low forecasting accuracy.

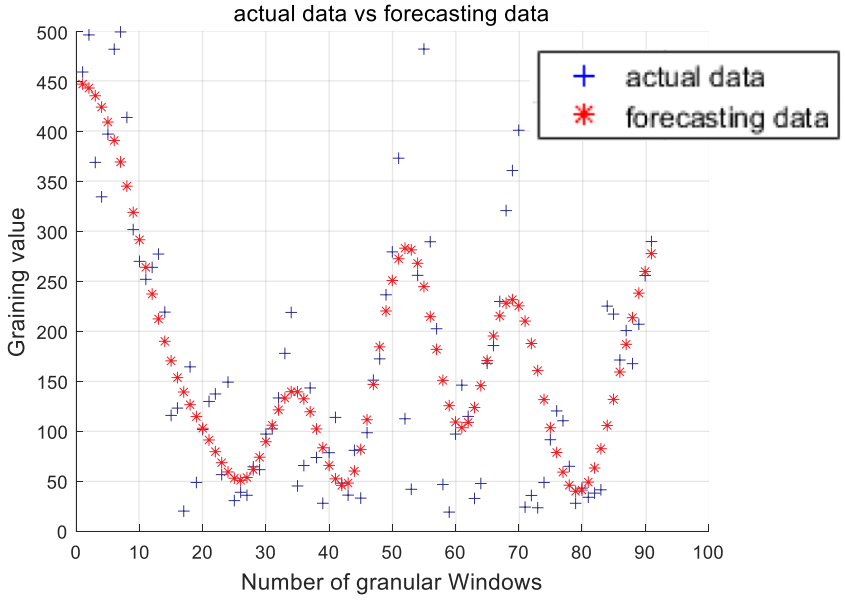

Fig. 13. Forecasting results of Model3.

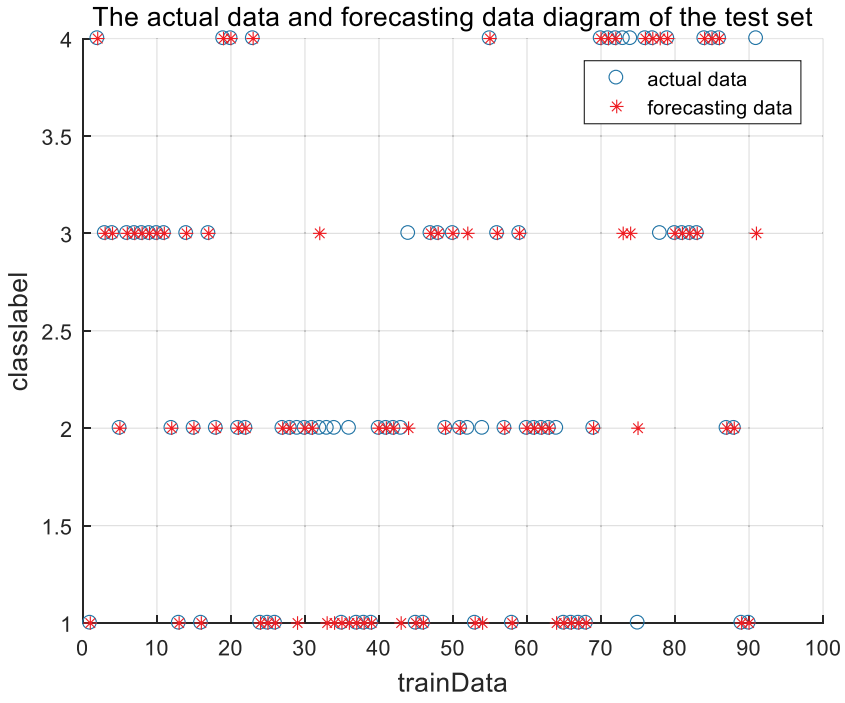

Fig. 14. Forecasting results of Model4.

Table 6
Forecasting errors of models.

No.	Cluster 1	Cluster 2	Cluster 3	Cluster 4	MAPE	MAPE difference
Model1	0.1568	0.0611	0.1391	0.1449	0.1255	0
Model2	0.1568	0.0628	0.1601	0.1451	0.1312	0.0057
Model3	/	/	/	/	0.7799	0.6544
Model4	0.1859	0.2447	0.2742	0.2642	0.2423	0.1168

Therefore, our proposed hybrid model is superior to the other models in terms of both forecasting accuracy and running time.

5.3. Statistical testing

We perform a non-parametric test (NPT) to test the statistical significance of the proposed model. Although NPT is not as powerful as a parametric test, this approach demonstrates its superiority when making assumptions about the population distribution is impossible. We perform NPT by comparing the proposed model with some conventional methods, such as BP, SVM, ANN and WNN.

Firstly, we test the normal distribution of the results obtained by each model. By performing the Shapiro–Wilk test, we find that the obtained results do not follow a normal distribution. The Q–Q plots in Fig. 16 also confirm this conclusion. We then perform the Kruskal–Wallis test (KWT) to test the null hypothesis that none of the groups is dominant over the others. The KWT results are presented in Fig. 17, where the bold line represents the median, the height represents the quartile distance and the highlighted circles represent the outliers. Multiple comparisons of models are presented in Table 7, in which columns 3 to 5 represent the minimum, group mean and maximum values of the models whilst the last column represents the p-value obtained via KWT. The null hypothesis is rejected when $\alpha = 0.05$. In other words, some differences can be observed between the proposed model and the conventional methods, thereby verifying the statistical significance of our proposed model.

6. Discussions

We perform several experiments to test the forecasting performance of our proposed hybrid model and the results demonstrate that this model outperforms the other models in terms of accuracy and running time. Table 6 shows that the forecasting accuracy of Model1 is better than that of Model2, but the differences in their accuracy are not obvious, especially in cluster 1. This finding indicates that by inputting all variables, the corresponding forecasting effect can be achieved after a certain period. However, in terms of runtime, Model1 shows the greatest advantage over the other models. The input variables analysed by the influential factors can also help trace the source of pollution

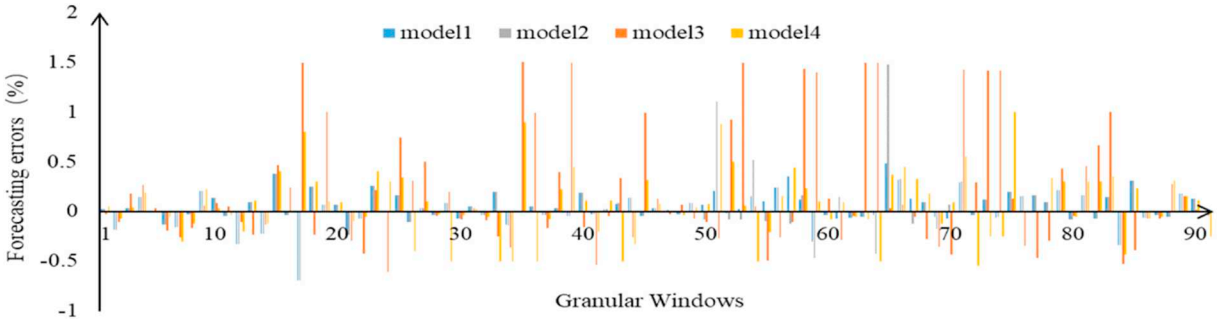

Fig. 15. Granular window forecasting errors of the models.

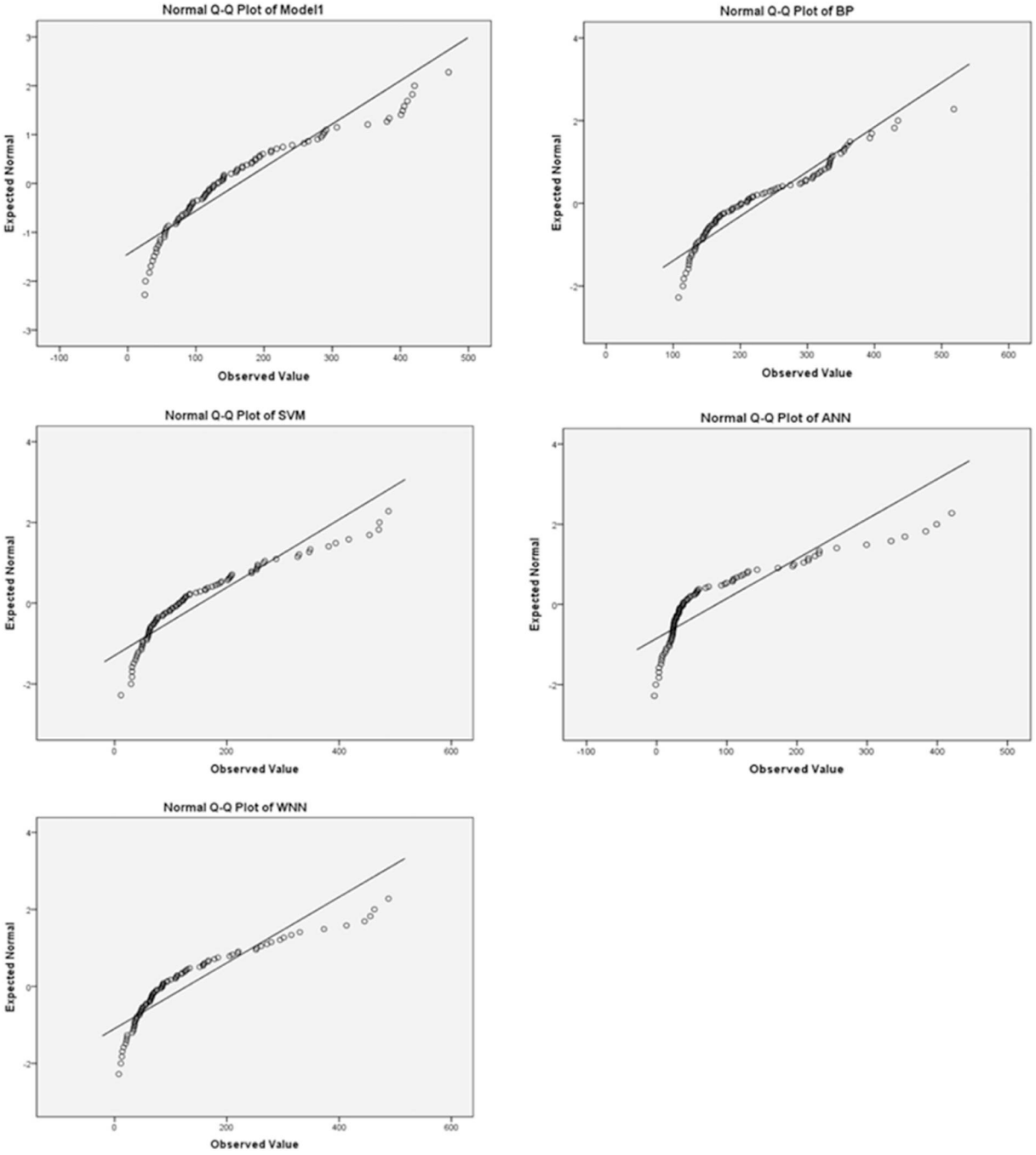

Fig. 16. Q–Q plots of the test models.

based on various information (e.g. weight), thereby helping us effectively control air pollution. We verify the effectiveness of the clustering algorithm by comparing Model1 with Model3. Fig. 13 and Table 6 show that Model1 has a better forecasting performance compared with Model3 due to the combination of the clustering algorithm in the proposed model. The results obtained by K-means algorithm can achieve higher similarity between data. The consistency between the forecasting data and actual data can therefore be improved. At the same time, the small datasets obtained by clustering meet the demand of SVM for a small sample data volume. The reduction in the sample data volume will also inevitably shorten the runtime.

We also compare the performance of our proposed hybrid model with another widely used model called GA-SVM. Table 6 and Fig. 15 highlight a forecasting performance gap between Model1 and Model4. The MAPE values in each cluster of the proposed model are smaller than those of Model4 because the GA algorithm has many parameters, including crossover and mutation rates. The choice of these parameters seriously affects the quality of the solution. Model1 also has a shorter runtime compared with the GA–SVM model, thereby highlighting the good convergence of the proposed hybrid model.

7. Conclusion

SAPCF has received much research attention due to its effectiveness in controlling air pollution. This paper develops a hybrid model for SAPCF based on influential factors analysis, K-means clustering, SVM and PSO and then tests its performance by conducting a case study in Temple of Heaven, Beijing. The forecasting process conducted in this

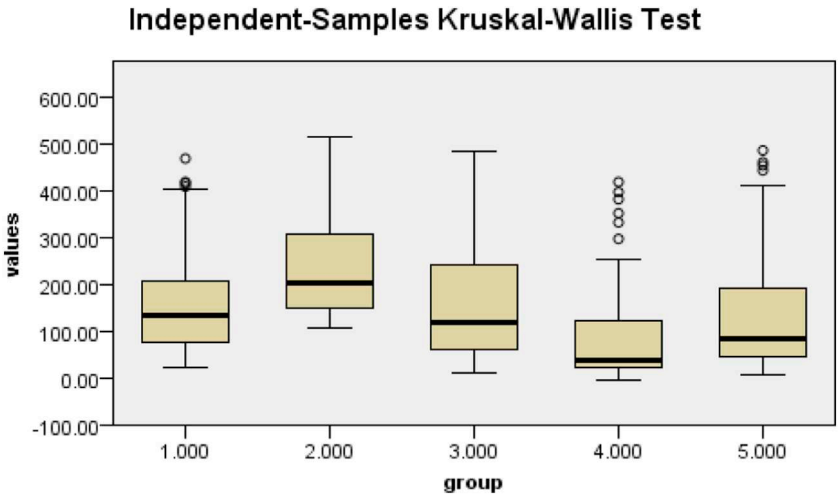

Fig. 17. Box plot of the KWT results.

Table 7
Multiple comparisons of KWT.

Group	Comparing	Minimum	Group mean	Maximum	p-Value
1	Model1 vs. BP	24.9253	196.7081	518.2727	0.001
2	Model1 vs. SVM	11.280	162.9646	487.9690	0.0359
3	Model1 vs. ANN	−3.326	128.0256	470.7976	0.00
4	Model1 vs. WNN	7.4023	150.4739	487.9327	0.013

study is summarised as follows. The data are initially pre-processed, the influencing factors are analysed and the optimal combination of variables is inputted into the hybrid PSO–SVM model based on K-means. All these processes sue the AQI as an example. We also perform a comparative analysis to verify the performance of our proposed hybrid model. The comparison results highlight the advantages of our proposed model in terms of forecasting accuracy and runtime. Meanwhile, our statistical tests verify the statistical significance of our proposed hybrid method. The major contributions of our work can be summarised as follows:

(1) We introduce influential factors analysis into our proposed model to obtain the optimal combinations of input variables and to help us reduce the forecasting time and trace the sources of pollution.
(2) We introduce an unsupervised clustering algorithm into our model to enhance the regularity of our modelling data. We then obtain a high similarity dataset and reduce the amount of running data to improve the forecasting accuracy of our proposed hybrid model and shorten its running time.
(3) We develop a new PSO algorithm to obtain the optimal parameters of SVM. Our proposed PSO–SVM model can realise automatic parameter selection and overcome the premature convergence problem of SVM.

The findings presented above reveal that our proposed APCF strategy is novel and effective. However, there are still some shortcomings of this work. For example, temporary emergencies, such as major holidays, can be used as influential factors and some seasonal factors may also need to be considered. Future studies may investigate how these factors can be considered comprehensively. Furthermore, short-term forecasting models include but are not limited to the ones we have mentioned. How to combine other effective methods and make full use of their advantages also require further discussion.

Acknowledgements

The authors thank the editors and anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China (No. 71871228) and Postgraduate Survey Project of Central South University (No. 2018dcyj035).

References

Abdul-Wahab, S.A., Bakheit, C.S., Al-Alawi, S.M., 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ. Model. Softw. 20 (10), 1263–1271.
Bai, Y., Li, Y., Wang, X.X., Xie, J.J., Li, C., 2016. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmospheric pollution research 7 (3), 557–566.
Bai, L., Wang, J., Ma, X., Lu, H., 2018. Air pollution forecasts: an overview. Int. J. Environ. Res. Public Health 15 (4), 780.
Barman, M., Dev Choudhury, N.B., Sutradhar, S., 2018. A regional hybrid GOA-SVM model based on similar day approach for short-term load forecasting in Assam, India. Energy 145, 710–720.
Benmouiza, K., Cheknane, A., 2013. Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models. Energ Convers Manag 75, 561–569.
Bollen, J., 2015. The value of air pollution co-benefits of climate policies: analysis with a global sector-trade CGE model called WorldScan. Technol Forecast Social 90, 178–191.
Chen, Q., Cao, G.N., Chen, L., 2010. Application of support vector machine to atmospheric pollution prediction. Comput Tec Dev 32 (12), 61–65.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Cortina-Januchs, M.G., Quintanilla-Dominguez, J., Vega-Corona, A., Andina, D., 2015. Development of a model for forecasting of PM10 concentrations in Salamanca, Mexico. Atmospheric Pollution Research 6 (4), 626–634.
Coussement, K., Lessmann, S., Verstraeten, G., 2017. A comparative analysis of data preparation algorithms for customer churn prediction: a case study in the telecommunication industry. Decis. Support. Syst. 95, 27–36.
Duan, J.C., Tan, J.H., Hao, J.M., Chai, F.H., 2014. Size distribution, characteristics and sources of heavy metals in haze episod in Beijing. J. Environ. Sci. 26 (1), 189–196.
Garg, H., 2015. A hybrid GA-GSA algorithm for optimizing the performance of an industrial system by utilizing uncertain data. In: Handbook of Research on Artificial Intelligence Techniques and Algorithms. IGI Global, pp. 620–654.
Garg, H., 2016. A hybrid PSO-GA algorithm for constrained optimization problems. Appl. Math. Comput. 274, 292–305.
Garg, H., 2019. A hybrid GSA-GA algorithm for constrained optimization problems. Inform Sciences 478, 499–523.
Ghaemi, Z., Farnaghi, M., Alimohammadi, A., 2015. Hadoop-based distributed system for online prediction of air pollution based on support vector machine. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 40 (1), 215.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a k-means clustering algorithm. J. R. Stat. Soc.: Ser. C: Appl. Stat. 28 (1), 100–108.
Hong, W.C., 2010. Application of chaotic ant swarm optimization in electric load

forecasting. Energ Policy 38 (10), 5830–5839.
Huber, P.J., 1985. Projection pursuit. Ann. Stat. 13 (2), 435–475.
Inman, R.H., Pedro, H.T.C., Coimbra, C.F.M., 2013. Solar forecasting methods for renewable energy integration. Progress in Energy & Combustion Science 39 (6), 535–576.
Jordehi, A.R., 2015. Enhanced leader PSO (ELPSO): a new PSO variant for solving global optimisation problems. Appl. Soft Comput. 26, 401–417.
Kumar, A., Goyal, P., 2011. Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research 2 (4), 436–444.
Li, Y., Tao, Y., 2018. Daily PM10 concentration forecasting based on multiscale fusion support vector regression. J. Intell. Fuzzy Syst. 34 (1), 3833–3844.
Nebenzal, A., Fishbain, B., 2018. Long-term forecasting of nitrogen dioxide ambient levels in metropolitan areas using the discrete-time Markov model. Environ. Model. Softw. 107, 175–185.
Niu, M.F., Wang, Y.F., Sun, S.L., Li, Y.W., 2016. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. Atmos. Environ. 134, 168–180.
Patwal, R.S., Narang, N., Garg, H., 2018. A novel TVAC-PSO based mutation strategies algorithm for generation scheduling of pumped storage hydrothermal system incorporating solar units. Energy 142, 822–837.
Rahman, N.H.A., Lee, M.H., Latif, M.T., 2015. Artificial neural networks and fuzzy time series forecasting: an application to air quality. Qual. Quant. 49 (6), 2633–2647.
Rohde, R.A., Muller, R.A., 2015. Air pollution in China: mapping of concentrations and sources. PLoS One 10 (8), e0135749.
Russo, A., Lind, P.G., Raischel, F., Trigo, R., Mendes, M., 2015. Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmospheric Pollution Research 6 (3), 540–549.
Schlechtingen, M., Santos, I.F., Achiche, S., 2013. Using data-mining approaches for wind turbine power curve monitoring: a comparative study. IEEE Transactions on Sustainable Energy 4 (3), 671–679.
Shah, H., Tairan, N., Garg, H., Ghazali, R., 2018. A quick gbest guided artificial bee colony algorithm for stock market prices prediction. Symmetry 10 (7), 292.
Soh, P.W., Chang, J.W., Huang, J.W., 2018. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access 6, 38186–38199.
Sun, W., Sun, J.Y., 2017. Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. J. Environ. Manag. 188, 144–152.
Wang, Y., 2011. Quality analysis of meteorological data from Huabei expressway in Tianjin. Meteorol Sci Tec 6, 413.
Wang, P., Liu, Y., Qin, Z.D., Zhang, G.S., 2015a. A novel hybrid forecasting model for PM10 and SO2 daily concentrations. Sci. Total Environ. 505, 1202–1212.
Wang, Z.Y., Lu, F., Lu, Q.C., Wang, D.S., Peng, Z.R., 2015b. Fine-scale estimation of carbon monoxide and fine particulate matter concentrations in proximity to a road intersection by using wavelet neural network with genetic algorithm. Atmos. Environ. 104, 264–272.
Wang, P., Zhang, H., Qin, Z.D., Zhang, G.S., 2017. A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting. Atmospheric Pollution Research 8 (5), 850–860.
Wang, Y., Kung, L., Byrd, T.A., 2018a. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Social 126, 3–13.
Wang, Y.F., Wang, H.Y., Chang, S.H., Avram, A., 2018b. Prediction of daily PM2.5 concentration in China using partial differential equations. PLoS One 13 (6), e0197666.
Whiteman, C.D., Hoch, S.W., Horel, J.D., Charland, A., 2014. Relationship between particulate air pollution and meteorological variables in Utah's Salt Lake Valley. Atmos. Environ. 94, 742–753.
Wu, L.F., Li, N., Yang, Y.J., 2018. Prediction of air quality indicators for the Beijing-Tianjin-Hebei region. J. Clean. Prod. 196, 682–687.
Xiao, Y.C., Kang, N., Hong, Y., Zhang, G.J., 2017. Misalignment fault diagnosis of DFWT based on IEMD energy entropy and PSO-SVM. Entropy 19 (1), 6.
Xie, R., Wei, D.H., Han, F., Lu, Y., Fang, J.Y., Liu, Y., Wang, J.F., 2018. The effect of traffic density on smog pollution: evidence from Chinese cities. Technol Forecast Social. https://doi.org/10.1016/j.techfore.2018.04.023.
Yousefzadeh, B., Shalmany, S.H., Makinwa, K.A., 2017. A BJT-based temperature-to-digital converter with ± 60 mK (3~\sigma) inaccuracy from −55° C to +125 °C in 0.16-μm CMOS. Ieee J Solid-st Circ 52 (4), 1044–1052.
Zendehboudi, A., Baseer, M., Saidur, R., 2018. Application of support vector machine models for forecasting solar and wind energy resources: a review. J. Clean. Prod. 199, 272–285.
Zeng, X.L., Song, Q.B., Li, J.H., Yuan, W.Y., Duan, H.B., Liu, L.L., 2015. Solving e-waste problem using an integrated mobile recycling plant. J. Clean. Prod. 90, 55–59.
Zhai, B., Chen, J., 2018. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. Sci. Total Environ. 635, 644–658.
Zhang, L.Y., Lin, J., Qiu, R.Z., Hu, X.S., Zhang, H.H., Chen, Q.Y., Tan, H.M., Lin, D.T., Wang, J.K., 2018. Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. Ecol. Indic. 95, 702–710.
Zheng, L.G., Zhou, H., Wang, C.L., Cen, K.F., 2008. Combining support vector regression and ant colony optimization to reduce NO_x emissions in coal-fired utility boilers. Energ Fuel 22 (2), 1034–1040.
Zhou, H.M., Deng, Z.H., Xia, Y.Q., Fu, M.Y., 2016. A new sampling method in particle filter based on Pearson correlation coefficient. Neurocomputing 216, 208–215.
Zou, Y., Chan, S.-C., Ng, T.-S., 2000. Least mean M-estimate algorithms for robust adaptive filtering in impulse noise. IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing 47 (12), 1564–1569.