The Prediction of PM2.5 Concentration with an Intelligent Hybrid Model

Xin Liu, Xiaoxi Zhang and Xuejing Zhao*
School of Mathematics and Statistics, Lanzhou University, Lanzhou, 730000, China

*Corresponding author: zhaoxj@lzu.edu.cn

Abstract. The objective of this work was to propose a hybrid model to predict the concentration of PM2.5 in three cities of China. PM2.5 is one of the most important pollution worldwide, therefore effective prevention and control are beneficial to human's production and life. A hybrid model, CEEMD-MFO-SVR- GRA-BPNN, was proposed to predict the concentration of PM2.5. The proposed model is the combination of (1) complementary ensemble empirical mode decomposition (CEEMD) to decompose the original PM2.5 concentration data; (2) support vector regression (SVR) to give a regressed prediction model in which parameters was optimized via moth-flame optimization algorithm (MFO); (3) grey relational analysis (GRA) to select atmospheric factors with distinguished effect on PM2.5; and (4) back propagation neural network (BPNN) to reduce the forecasting residual. The hybrid model was evaluated in three cities, Guiyang, Lijiang and Guangzhou of China, in which the environments and geographical locations are different. The implementation of the proposed model and well-known CEEMD-MFO-SVR, CEEMD- WOA-SVR, EEMD-MFO-SVR, EMD-MFO-SVR and MFO-SVR, BPNN-meteorology models, were compared. The results show that the prediction of the proposed hybrid model is more accurate than the compared models.

Keywords: PM2.5 concentration forecasting; Complementary ensemble empirical mode decomposition; Moth-flame optimization; Support vector regression; Grey relational analysis; Back propagation neural network.

1. Introduction

A report by United Nations Environment Organization states that: “Air pollution has become an unavoidable reality in the lives of urban residents around the world”. Nowadays, China have made great development in urbanization and industrialization, it also becomes a center of world manufacturing. However, the environment has suffered severe damage, especially air pollution. The main cause of haze pollution is the air containing high concentrations of PM2.5[1]. The higher the PM2.5 concentration, the more serious the air pollution. Air pollution is the main environmental health risk affecting human health, exposure to ambient air pollution, especially to particulate matter[2] could cause various diseases[3], such as allergic diseases[4], respiratory diseases[5], Alzheimer’s disease (AD)[6], bladder cancer[7], and cardiovascular disease[8], to name a few.

Recently, hybrid model has been widely used in the field of atmosphere. Sánchez et al[9] devoted support vector machine (SVM) method to establish a regression model for air quality in Spain. Dong et al[10] used SVM to estimate the concentration of PM2.5 in Shenzhen without considering the atmospheric factors. To overcome the drawbacks of single statistical model, hybrid models were now widely used, Zhou et al[12] used the EEMD-GRNN hybrid model, which preprocessed the data, to predict the PM2.5 concentration on a day in Xi’an, Niu et al[13] combined CEEMD and grey wolf...
optimization (GWO) algorithms to predict the concentration of atmospheric pollutants. Since the air concentration data is composed of complex linear and nonlinear models[15], we propose a hybrid model to improve the prediction accuracy. The hybrid model is the combination of CEEMD-MFO-SVR and GRA-BPNN, in which CEEMD-MFO-SVR considers the nonlinear influence of the historical PM2.5 concentration, and GRA-BPNN considered the effects of atmospheric factors on PM2.5 concentration. Figure 1 shows the process of model comparison. The main features and contributions of the CEEMD-MFO-SVR-GRA-BPNN hybrid model are as follows:

1) Complementary ensemble empirical mode decomposition (CEEMD) is used to decompose the original PM2.5 data of three cities into three intrinsic mode functions (IMF) (reference literature[13]) and a remainder series respectively. Without considering the influence of atmospheric factors, nonlinear SVM-based prediction will be established for each IMF and residual respectively, and finally a single prediction will be aggregated into a comprehensive prediction.

2) Support vector regression (SVR) is exploited to deduce a nonlinear time series regression of the decomposed data, where a RBF kernel function and the Moth-flame optimization algorithm (MFO) are chosen to improve the goodness of fit, thereafter improving the prediction accuracy of the model.

3) Grey Relational Analysis (GRA) is investigated to select the distinguished meteorological factors affecting PM2.5 concentration, so as to reduce the complexity of the model with many meteorological factors.

4) Back Propagation Neural Network (BPNN) is used to establish a three-layer neural network to give the prediction of the PM2.5 concentration. The GRA-filtered variables are used as the input to BPNN, and the difference between the true value and the CEEMD-MFO-SVR prediction as the predicted output.

Figure 1. Procedure of PM2.5 forecasting.

Figure 2. The structure of the proposed model

Figure 2 describes the structure of the proposed model as a highlight of this paper.
CEEMD-MFO-SVR (Model1, the predicted value corresponding is represented by P) is taken into account to give a prediction of nonlinear times series with lag 4, and GRA-BPNN (Model2, the predicted value is represented by R) is utilized to give more accuracy, in which the chosen meteorological factors in GRA are considered to be input parameters in BPNN. Since Model2 predicted the residual (R=y−P) of Model1(P) and the true value(y), the results are then added to obtain the final prediction result (P+R).

The rest of the paper is organized as follows. Materials and main methods of the hybrid model, such as cities’ descriptions, CEEMD, SVR, MFO, GRA and BPNN are described in Section 2. The empirical studies included descriptions of evaluation criteria, analysis of prediction results and model comparison are discussed in Section 3. And conclusions are addressed in Section 4.

2. Materials and Methods

Three cities with different geographical environment and economic levels, Lijiang, Guiyang and Guangzhou, have been selected to evaluate the effectiveness of the proposed hybrid model and conducted PM2.5 concentration prediction.

The training set contained the data from January 1, 2017 to December 1, 2017 (335 data) and the data from December 2, 2017 to December 31, 2017 (30 data) as testing data, calculated the error of the hybrid model by predicting the PM2.5 concentration for the next 30 days as 30 independent samples. The meteorological factors are daily average air pressure, daily average temperature, daily average relative humidity, daily precipitation, daily large evaporation, daily average wind speed, daily sunshine hours and daily average surface air temperature.

In order to avoid the impact of the default parameters in the experiment, we use the same value of parameters for the three cities: the number of search agents are 20, max iteration of MFO is 100, max iteration of CEEMD is 5000, number of IMFs are all equal to 3, standard deviation of the Noise is 0.4, number of realization is 200. The following outlines the statistical methods for analysis used in this paper.

2.1. Complementary Ensemble Empirical Mode Decomposition (CEEMD)

Empirical mode decomposition (EMD) proposed by Huang et al.[15] was to deal with signal decomposition based on time scale characteristics. By EMD decomposition, a complex time sequence (PM2.5concentration) could be decomposed into a certain number of mode functions (IMFs) and a remainder series. The decomposed IMFs contained some features of different scales of the original signal (PM2.5concentration). The final prediction result could be obtained by integrating the post-decomposition IMFs prediction results.

CEEMD[16] combined the advantages of the EMD algorithm and the EEMD algorithm[17]. The algorithm mainly added positive and negative paired auxiliary noise signals to the initial signal. It could solve the model aliasing phenomenon that EMD prone to decomposition, also it could avoid the problem of EEMD with reconstruction error, low efficiency and long operation time in decomposition.

2.2. Support Vector Regression (SVR)

Support vector regression was proposed to implement linear regression by constructing linear decision functions in high-dimensional space. For each decomposed series IMF(i)(i = 1,2, ..., M) aforementioned, a SVR model is used to give it’s autoregressive forecast. Denote by xi = {yi−p, y(i−p+1, ..., y(i−1}, then yi = f(xi) + ε = f(yi−p, y(i−p+1, ..., y(i−1)) + ε, where p is the order of lag.

The SVR function in the feature space could be written as:

\[ f(x) = \omega \ast \phi(x) + b, \]

(1)

where \( \phi(x) \) represents the feature vector mapping from x, \( \omega \) is the weight vector and b is the adjustable factor.

In order to minimize the loss function of the model, SVR can be expressed as follows:

\[ \min_{\omega, b} \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{m} l_\varepsilon(f(x_i) - y_i), \]

(2)
where \( c \) is the regularization constant and \( l_e \) is the \( \epsilon \)-insensitive loss function.

### 2.3. Moth-flame Optimization Algorithm (MFO)

The MFO algorithm was inspired by the natural moth lateral positioning navigation mechanism and updated the moth-flame position through the moth-flame spiral flight mathematical model, and finally converged to the flame (light) position of the new group intelligent bionic algorithm. The MFO algorithm selects the spiral function as the update operator of the moth’s spatial position.

\[
S(M_i, F_j) = |F_j - M_i| \cdot e^{bt} \cdot \cos(2\pi t) + F_j,
\]

(3)

Where \( M_i \) represents the \( i \)-th moth position, \( F_j \) represents the \( j \)-th flame, the optimal parameter value obtained at present; \( b \) is a spiral shape constant; \( t \) is a random number between \((-1, 1)\).

The moth flame optimization algorithm enabled the moth to continuously update the position around the flame in space. In order to ensure the convergence speed, the number of flames is gradually reduced during the iterative process, so that the global optimal solution can be achieved.

### 2.4. Gray Relational Analysis (GRA)

Grey Relational Analysis (GRA) is a system analysis method used to quantitatively study the complex interactions among multiple factors in a system[18]. Since there are many meteorological factors affecting the concentration of PM2.5, if all the influencing factors are taken into account, there will be too many input variables in the model, which will lead to a complicated model structure, a slow learning speed, and ultimately poor affection of popularization and application. In order to reduce the complexity of the model, the Grey relation analysis method is used to extract the main factors affecting PM2.5 as input variables of the prediction model.

### 2.5. Back Propagation Neural Network (BPNN)

BPNN is a supervised learning algorithm. It is a neural network method that propagates the signal forward by inversely propagating the error using the gradient descent method. The BP network could transform the original input and output problem of the sample into a nonlinear optimization problem. The input signal is forwardly propagated. First, the predicted value is output, and the output error is propagated from output neurons to input neurons. Then, according to the gradient descent method, the weights of each layer are continually updated, and the iteration is continued until the predicted output is as close as possible to the real value [19]. In this paper, the data fitted by the CEEMD-MFO-SVR model is compared with the real data, and the remained information in the original data is predicted by GRA-BPNN to give a more accurate prediction.

### 3. Results

#### 3.1. Evaluation Criteria

There is currently no standard general method for evaluating the accuracy of model prediction. In regression prediction, the evaluation criteria widely used are the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and the mean square error (MSE). The lower the RMSE and MAE values were, the better the fitted model was. MAPE is often used to calculate the accuracy of simulations and predictions related to time series. Also we suggested a MDI evaluator to measure the relative improvement of our proposed method compare to the other well-known algorithms:

\[
MDI = \frac{\text{RMSE}_{\text{compare}} - \text{RMSE}_{\text{propose}}}{\text{RMSE}_{\text{compare}}},
\]

(4)

#### 3.2. Analysis of Predicted Results and Model Comparison

For the comparative models in the paper, such as CEEMD-MFO-SVR, CEEMD-WOA-SVR, EEMD-MFO-SVR, EMD-MFO-SVR, we have adopted MFO and WOA optimization algorithms to...
optimize the parameters. Therefore, the models in Table 1 is the results under the conditions of optimal parameters.

Figure 3 shows the decomposition results of the sample data of Guiyang, Lijiang and Guangzhou through CEEMD. Using CEEMD to decompose PM2.5 concentration data, since too many IMFs may lead to model complexity and computational cost, the original data is decomposed into three IMFs and one remainder in this paper.

After the decomposition of the data, each IMFs and the reminder can be analyzed and the prediction can be archived by the proposed and compared model. Table 1 describes the comparison of the performance of each model.

![Graphs showing decomposition results](image)

**Figure 3.** The decomposition results in Guiyang(a), Lijiang(b) and Guangzhou(c).

This paper considered the impact of PM2.5 concentration in the past few days on future, and also considered the impact of meteorological factors on concentration. An intelligent hybrid model CEEMD-MFO-SVR-GRA-BPNN was used to predict the PM2.5 concentration in three cities: Guiyang, Lijiang and Guangzhou, and to compare with different methods of decomposition EEMD, EMD, different optimization algorithms WOA and other intelligent models MFO-SVR, BPNN. CEEMD can save processing time when data is decomposed, and as the amount of added noise increases, the residual amount of noise in the final reconstructed data gradually decreases or even neglects; When using MFO to optimize parameters, the optimization ability is better than other algorithms, which is proved to be a promising algorithm; When SVR is used for regression, the RBF kernel function can be used to map the original data to high dimensional, which improves the fitting effect on the data; GRA analyzes the grey relational degree of the data, reduces the dimension of data input, and simplifies the complexity of the model; BPNN neural network adds a backward propagation algorithm to the structure of the feed forward network. By using back propagation error, updating continuously and the gradient descent algorithm, it can optimize the parameters to reduce the error and get more excellent results.
Table 1. The comparison of forecasting performance.

| City         | Models                             | MAE     | MAPE (%) | MSE     | RMSE    | R       | MDI(%) |
|--------------|------------------------------------|---------|----------|---------|---------|---------|--------|
| Guiyang      | CEEMD-MFO-SVR-GRA-BPNN             | 4.6919  | 0.1002   | 33.9362 | 5.8252  | 0.9603  |        |
|              | CEEMD-MFO-SVR                      | 5.2866  | 0.1190   | 46.7394 | 6.8836  | 0.9577  | 14.55  |
|              | CEEMD-WOA-SVR                      | 5.3074  | 0.1994   | 46.7394 | 6.8836  | 0.9577  | 14.79  |
|              | EEMD-MFO-SVR                       | 6.0873  | 0.1132   | 53.9106 | 7.3424  | 0.9543  | 20.66  |
|              | EMD-MFO-SVR                        | 12.1793 | 0.2591   | 216.608 | 14.717  | 0.8114  | 60.42  |
|              | MFO-SVR                            | 13.3485 | 0.4064   | 263.988 | 16.247  | 0.8876  | 64.15  |
|              | BPNN-meteorology                   | 15.2371 | 0.4670   | 384.036 | 19.596  | 0.7642  | 70.27  |
| Lijiang      | CEEMD-MFO-SVR-GRA-BPNN             | 5.4643  | 0.1703   | 59.5119 | 7.7144  | 0.9752  |        |
|              | CEEMD-MFO-SVR                      | 5.4871  | 0.1790   | 65.3036 | 8.0811  | 0.9625  | 4.54   |
|              | CEEMD-WOA-SVR                      | 5.7621  | 0.1927   | 67.4635 | 8.2136  | 0.9732  | 6.08   |
|              | EEMD-MFO-SVR                       | 5.8596  | 0.2124   | 63.5262 | 7.9703  | 0.9653  | 3.21   |
|              | EMD-MFO-SVR                        | 7.0951  | 0.2562   | 93.3028 | 9.6593  | 0.8834  | 20.13  |
|              | MFO-SVR                            | 10.9864 | 0.3623   | 234.563 | 15.315  | 0.7038  | 49.63  |
|              | BPNN-meteorology                   | 10.5642 | 0.3635   | 211.016 | 14.526  | 0.6146  | 44.37  |
| Guangzhou    | CEEMD-MFO-SVR-GRA-BPNN             | 6.2319  | 0.1355   | 54.6097 | 7.3898  | 0.9480  |        |
|              | CEEMD-MFO-SVR                      | 6.3784  | 0.1432   | 61.7004 | 7.8550  | 0.9406  | 5.92   |
|              | CEEMD-WOA-SVR                      | 6.3787  | 0.420    | 64.5319 | 8.0332  | 0.9388  | 8.01   |
|              | EEMD-MFO-SVR                       | 6.7868  | 0.1548   | 66.4649 | 8.1562  | 0.9474  | 9.40   |
|              | EMD-MFO-SVR                        | 9.2630  | 0.2302   | 135.062 | 11.621  | 0.8778  | 36.41  |
|              | MFO-SVR                            | 17.6574 | 0.4048   | 475.424 | 21.804  | 0.4488  | 66.11  |
|              | BPNN-meteorology                   | 19.7677 | 0.4507   | 584.973 | 24.186  | 0.2653  | 67.52  |

4. Conclusions
The results showed that the proposed model was smaller than the comparison models on MAE, MAPE, MSE, and RMSE. Also it is noticed that the MAE, MAPE, MSE and RMSE values of BPNN-meteorology model were almost the maximums among the seven models, so it is ineffective to build PM2.5 concentration forecasting model by considering only the meteorological factors. MDI measures the degree of improvement of the model or the degree of error reduction and the minimum and maximum MDI values were 3.21% and 70.27% in Table 1 respectively, which means that the
The proposed model could reduce the minimum model error of 3.21% and the maximum model error of 70.27% respectively. The correlation coefficient \( R \) was also higher than in case of other models, so the hybrid model’s fitting effect was significantly improved. Since the hybrid model took into account the impact of both aspects (including the impacts of previous few days and meteorological factors), it could be minimized in the RMSE indicator. Compensated for the unilateral consideration of the predicted concentration of meteorological factors, the same method could be used to predict other pollutant concentrations, such as PM10, SO2, etc.

5. Data Availability Statements
The raw daily average PM2.5 of three cities were retrieved from the website of the online air quality monitoring and analyze platform of China (https://www.aqistudy.cn/history data/) and data on various meteorological factors were obtained from the National Meteorological Science Data Sharing Service Platform (http://data.cma.cn/site/index.html). Derived data supporting the findings of this study are available from the corresponding author upon request.

Acknowledgements
X. Zhao was supported by National Natural Science Foundation of China (No. 11971214, 81960309) and Cooperation Project of Chunhui Plan of the Ministry of Education of China-2018, and sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China. The authors would also like to thank Edit-in-chief and the referees for their suggestions to improve the paper.

Conflicts of Interest
The authors declare that there were no conflicts of interests for the publication of this study.

References
[1] D. Yan, Y. Lei, Y. Shi, Q. Zhu, L. Li, Z. Zhang, Evolution of the spatiotemporal pattern of PM2.5 concentrations in china: a case study from the Beijing-Tianjin-Hebei region, Atmospheric Environment. 2018: S135223101830195X.
[2] A. Badyda, A. Gayer, P. Czechowski, G. Majewski, P. Dabrowiecki, Pulmonary function and incidence of selected respiratory diseases depending on the exposure to ambient PM10, International Journal of Molecular Sciences 2016, 17 (11): 1954.
[3] A. Badyda , P. Czechowski, P. Dabrowiecki et al, The Prevalence of Selected Respiratory Diseases and the Exposure to PM10 in the Ambient Air, American Journal of Respiratory and Critical Care Medicine 2016, 193: A5417
[4] G. Porebski, M. Wooniak, E. Czarnobilska, Residential proximity to major roadways is associated with increased prevalence of allergic respiratory symptoms in children, Annals of Agricultural & Environmental Medicine 21 (876253) (2014) 760–766.
[5] S. Zhu, X. Lian, L. Wei, J. Che, X. Shen, L. Yang, X. Qiu, X. Liu, W. Gao, X. Ren, PM2.5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors, Atmospheric Environment 183: 20-32.
[6] L. Calderónarcidueus, J. Avilaramirez, A, et al, Cerebrospinal fluid biomarkers in highly exposed PM2.5 urbanites: The risk of Alzheimer’s and Parkinson’s diseases in young Mexico city residents., Journal of Alzheimer’s disease: JAD 2016, 54 (2): 597–613.
[7] H. L. Yeh, S. W. Hsu, Y. C. Chang, T. C. Chan, H. C. Tsou, Y. C. Chang, P. H. Chang, Spatial analysis of ambient pm2.5 exposure and bladder cancer mortality in Taiwan, International Journal of Environmental Research & Public Health 2017, 14 (5): pii508.
[8] T. Woodruff, J. Parker, K. Schoendorf, Fine particulate matter (PM2.5) air pollution and selected causes of post neonatal infant mortality in California, Environmental Health Perspectives 2006, 114 (5) : 786.
[9] A. Sánchez, P. Nieto, P. Fernández, J. Díaz, F. Iglesias-Rodríguez, Application of an svm based regression model to the air quality study at local scale in the avilés urban area (spain), Mathematical & Computer Modelling 2011, 54 (5-6) : 1453–1466.
[10] Y. Dong, H. Wang, L. Zhang, K. Zhang, An improved model for PM2.5 inference based on support vector machine, in: IEEE ACIS International Conference on Software Engineering, Artificial Intelligence, NETWORKING and Parallel distributed Computing, 2016, pp. 27–31.

[11] Q. Zhou, H. Jiang, J. Wang, J. Zhou, A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network, Science of the Total Environment 2014, 496 (2): 264–74.

[12] M. Niu, Y. Wang, S. Sun, Y. Li, A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting, Atmospheric Environment 2016, 134: 168–180.

[13] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. T. Chi, H. H. Liu, The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, Proceedings Mathematical Physical & Engineering Sciences 1998, 454 (1971): 903–995.

[14] J. Yeh, J. Shieh, N. Huang, Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis methods, Advances in Adaptive Data Analysis 2010, 2 (02): 135–156.

[15] Y. Xue, J. Cao, H. Du, G. Zhang, Y. Yao, Does mode mixing matter in emd-based highlight volume methods for hydrocarbon detection? experimental evidence, Journal of Applied Geophysics 2016, 132: 193–210.

[16] R. Kolhapure, V. Shinde, V. Kamble, Geometrical optimization of strain gauge force transducer using GRA method, Measurement 2017, 101: 111–117.

[17] M. Buscema, Back propagation neural networks, International Journal of the Addictions 1998, 33 (2): 233–270.