A knowledge-guided and manual intervention-based gene expression programming for PM2.5 concentration prediction

Chaoxue Wang, Xiaoli Jia*, Fan Zhang, and Yuhang Pan

College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China

Abstract. In view of the lack of interpretation and inability to know the occurrence mechanism of PM2.5 concentration by deep learning algorithm in solving PM2.5 concentration prediction problem, this paper adopts a knowledge-guided and manual intervention-based gene expression programming (KMGEP) to solve it. The KMGEP algorithm not only has strong model learning ability, but also can obtain the explicit function relationship between PM2.5 concentration and its influencing factors. In the process of algorithm implementation, knowledge guidance and manual intervention are introduced to GEP for predicting PM2.5 concentration ability and convergence speed. In this paper, the daily PM2.5 concentration in winter (from December to February) in Xi'an region is taken as an example, and the KMGEP algorithm is compared with the artificial neural network back propagation algorithm (BP-ANN) and the convolutional neural network and long short-term memory neural network combined model (CNN-LSTM). Experimental results show that the KMGEP algorithm not only has high prediction accuracy in solving the PM2.5 concentration prediction, but also the obtained function expression can reveal the occurrence relationship between PM2.5 concentration and its influencing factors.

1 Introduction

Air pollution is a serious threat to human health. Especially in the air, particulate matter (PM2.5) with a diameter of less than $2.5 \,\mu m$ can adhere to the deep respiratory tract due to its small size, and affect blood circulation by penetrating lung cells, thereby affect human health [1,2]. Studies have shown that the increase in PM2.5 concentration will increase the risk of airway obstructive diseases, chronic bronchitis, asthma, lung cancer and various cardiovascular diseases [3]. Therefore, accurate PM2.5 concentration prediction and the functional relationship between PM2.5 concentration and its influencing factors have become the key research problems for experts and scholars at home and abroad.

In recent years, with the development of artificial intelligence technology, many scholars have used machine learning algorithms to predict PM2.5 concentrations. Chen Cheng etc. used a multi-instance genetic neural network to predict indoor PM2.5 concentration, and its results were better than linear regression, support vector, random forest and other methods [4]. Zheng Guowei etc. established a combined prediction model based on support vector machine-wavelet neural network based on the non-linear and time-varying characteristics of PM2.5 concentration changes, and its prediction results were better than the single model of support vector machine [5]. Chen Qiang etc. used artificial neural network back propagation algorithm (BP-ANN) and multiple linear

regression model to predict the PM2.5 concentration in Zhengzhou, and the results showed that the BP-ANN algorithm was more effective in predicting the PM2.5 concentration in Zhengzhou [6]. Sun Yibo etc. proposed a PM2.5 concentration prediction model based on a deep neural network (DNN), and the results showed that DNN could greatly improve the prediction accuracy by using only aerosol optical depth data and meteorological observation data [7]. Zhao Wenfang etc. established predictive model based on deep learning, and the results showed that the model could effectively improve the prediction accuracy of PM2.5 concentration in the next 24 hours [8]. Li Taoying etc. proposed a combination model of convolutional neural network and long short term memory neural network (CNN-LSTM), and compared the univariate CNN-LSTM model and the multivariate CNN-LSTM model to prove the prediction of the multivariate CNN-LSTM model better [9]. Liu Xulin etc. proposed a deep learning prediction model based on convolutional neural networks and sequence-to-sequence, and the results showed that the model could effectively improve the prediction accuracy of PM2.5 concentration in the next hour and had a high generalization ability [10].

In summary, the deep learning algorithm already has been a hot spot method for PM2.5 concentration prediction, but it cannot get an explicit function expression. Gene Expression Programming (GEP) is a new type of machine learning algorithm proposed by Ferreira on the basis of genetic algorithm and genetic

^{*} Corresponding author: j997487801@163.com

programming, which is inspired by the open reading frame in genetics [11]. GEP not only has the same powerful model learning capabilities as deep learning algorithms, but also can obtain explicit functional relations. At present, the GEP algorithm has been successfully applied in the predictive modelling of software reliability [12], dew point [13], energy loss [14], and concrete mechanical properties [15].

In this paper, KMGEP algorithm is used to solve the problem of PM2.5 concentration prediction in the short-term. Based on the previous research results in PM2.5 concentration prediction, the influencing factors and related functions of PM2.5 concentration are designed according to the requirements of GEP algorithm so as to build a prior knowledge base of GEP algorithm. On the basis of natural evolution, manual intervention is added to enhance the efficiency of the algorithm and improve the quality of the solution. This paper takes the daily prediction of PM2.5 concentration in Xi'an in winter as a case, and finally obtains the functional relationship between PM2.5 concentration and various influencing factors, which can reveal the internal mechanism of PM2.5 concentration and various influencing factors. The degree of fit (R^2) , mean absolute error (MAE) and root mean square error (RMSE) are used as model prediction evaluation indicators. The effectiveness of the KMGEP prediction model is proved by comparing with the artificial neural network back propagation algorithm [6] (BP-ANN) and the combined model of convolutional neural network and long short term memory neural network [9] (CNN-LSTM) in the literature.

2 Methods

2.1 Gene expression programming algorithm

Gene expression programming (GEP) algorithm absorbs the coding form of genetic algorithm (GA) and the tree structure of genetic programming (GP). Its advantage lies in that it relies on its powerful search and evolution ability to find the optimal mathematical expression in line with the training data under the condition of not knowing the internal mechanism of things and only having the training data. The flow chart of GEP algorithm is shown as in Fig. 1.

The steps are as follows:

Step 1: [Initialization] including algorithm parameter setting, and the generation of initial population;

Step 2: [Abort condition] evaluate the fitness of each individual in the population to determine whether it has reached the maximum fitness value or the maximum number of iterations. If yes, output the final result, otherwise, go to the next step;

Step 3: [Mutation] randomly select an individual according to the mutation probability, and perform corresponding mutation operation on the selected individual randomly selected gene location;

Step 4: [(IS, RIS, gene) transposition] according to transposition probability randomly select individuals for the corresponding transposition operation;

Step 5: [Single point recombination, two points recombination and gene recombination] randomly select individuals for the corresponding recombination operation according to the recombination probability;

Step 6: [Selection] use the tournament selection to select individuals to form the next generation population, then go to step 2.



Fig. 1. The flowchart of GEP

2.2 PM2.5 concentration prediction model based on KMGEP

The setting of the influencing factors of the traditional gene expression programming algorithm is subjective to some extent. In this paper, the results of previous studies on the influencing factors of PM2.5 concentration and related functions are summarized and designed as the priori knowledge base of KMGEP. At the same time, for the slow convergence of traditional gene expression programming, the superior individuals in the current population are selected to form a posteriori knowledge base to guide the evolution direction of KMGEP algorithm, and the manual intervention is introduced to improve the evolutionary efficiency of KMGEP algorithm.

2.2.1 Knowledge guidance

Knowledge guidance is realized through the knowledge base. The knowledge base consists of two parts, a priori knowledge base and a posterior knowledge base.

(1) The establishment of a priori knowledge base

The priori knowledge base is obtained from previous studies on PM2.5 concentration, which influencing PM2.5 the factors contains of concentration and related functions. Research by Lu Debin etc. showed that the influencing factors of PM2.5 were meteorological factors (relative humidity, rainfall, wind speed, temperature), pollution sources (sulfur dioxide emissions, smoke and dust emissions), urbanization and industrial structure (population density, GDP per capita, ratio of primary, secondary and tertiary industries to GDP), and corporate pollution control and technology optimization (green area, green

coverage, sulfur dioxide removal, soot removal, R&D expenditure as a percentage of GDP) [16]. The research of Wu Zhuang etc. showed that in a short period of time, a region's development scale, geographical industrial pollution emissions conditions. and automobile exhaust emissions were relatively fixed, therefore, the change of PM2.5 concentration was mainly related to local meteorological conditions [17]. Through the research of Lv Baolei etc. and Wu Yuan etc., it is found that the meteorological factors affecting PM2.5 are wind, temperature, humidity, and air pressure [18][19]. Liu Suixin etc. in their research found that atmospheric pollution wind speed and precipitation were the factors affecting the PM2.5 concentration in Xi'an [20]. Meng Zhaowei etc. in their research found that daily average temperature, daily average pressure, daily average wind speed, daily average relative humidity, precipitation and the lowest temperature were significantly related to the mass concentration of PM2.5 [21]. Research by Peng Yan etc. showed that the influencing factors of PM2.5 concentration were PM10, SO2, NO2, O3, minimum temperature, average relative humidity, maximum wind speed, maximum wind direction, and sunshine time [22]. Research by Qu Chao etc. showed that the influencing factors of PM2.5 concentration were PM10, SO2, CO, O3, relative humidity, temperature and wind speed [23]. Research by Aydin Shishegaran etc. found that GEP could use nonlinear equations such as power functions and trigonometric functions to express the impact of meteorological parameters on air quality [24].

Table 1. Prior knowledge base

A prior knowledge base	Parameters	
	PM10(<i>x</i> ₀)	
	$SO2(x_1)$	
	$NO2(x_2)$	
	$O3(x_4)$	
Influencing factors	minimum temperature (x_5)	
	average temperature (x_6)	
	relative humidity (x_7)	
	average wind speed (x_8)	
	air pressure (x_9)	
	sunshine duration (x_{10})	
	Precipitation (x_{11})	
	maximum wind speed (x_{12})	
	the direction of the maximum	
	wind speed (x_{13})	
	+, -, *, /	
	log (logarithm based on 10)	
	exp (exponent of e)	
function sets	ln (logarithm based on e)	
	\sim (the exponent of 10)	
	x^{2} (the second power of x)	
	sqrt (root)	
	sin (sine)	
	cos (cosine)	
	abs (absolute value)	

According to the above literature research, it is found that the main factors affecting the PM2.5 concentration in the short-term prediction are the concentration of air pollutants and meteorological factors. Predictive models contain basic functions such as power functions and trigonometric functions with high probability. Based on the above analysis and the characteristics of short-term prediction in this paper, a prior knowledge base is constructed, as shown in Table 1.

(2) The establishment of a posteriori knowledge base

The current population was ranked by fitness, and the first *m* individuals were selected as the members of the posterior knowledge base.

2.2.2 Manual intervention

Manual intervention consists of individual intervention and population intervention.

(1) Individual intervention

Individual intervention includes two kinds of operations: repairing operator and strengthening operator.

The individual intervention operation, which aims to improve the quality of population individuals, consists of a repairing operator that removes the morbid genes in individuals and a strengthening operator that spreads eminent genes to the individuals of population. The specific operation method is as follows:

Repairing operator: for the inferior gene sites contained in the non-feasible solution in the population, such as making the divisor 0 and the gene site less than 0 under the quadratic radical, the gene loci leading to these inferior gene traits were recorded in the *inferiorgene* matrix, which was constantly changing in the course of evolution. The specific operations for determining inferior gene sites are as follows:

1) Calculate the effective length of each gene in the individual;

2) Traverse each gene location from right to left to find one or two operands of the gene location, and use the function operator on the gene location to calculate the operands;

3) If "ZeroDivisionError", "ValueError", and "Over-flowError" appear in the calculation process, then their gene sites are an inferior gene site.

The specific steps of repairing operator are as follows:

1) To find the inferior gene sites of the infeasible individuals stored in the *inferiorgene* matrix;

2) A symbol is randomly selected from the function character set F and the termination character set T to replace the symbols of inferior gene loci one by one;

3) Calculate the fitness of the individual after replacement. If the calculation result is real, the replacement is successful. Otherwise, continue with (2).

Strengthening operator: the specific steps of the strengthening operator are as follows:

1) For each individual in the population, if the number between 0 and 1 randomly generated is less than the probability of strengthening operator, the

individual will be optimized. Otherwise, the optimization operation is not performed.

2) For the individuals in the posteriori knowledge base, the gene fragment [s:t] from the position *s* to the position t of the *i*-th individual is randomly selected and transplanted to the position *s* to the position *t* of the *j*-th individual selected in step (1) to form a new individual *k*, and evaluate the fitness of *k*. If the fitness is greater than the fitness of the original individual *j*, then replace the original individual j with the new individual *k*, otherwise keep the original individual *j* unchanged.

(2) Population intervention

Population intervention is aimed at the phenomenon of premature convergence caused by the lack of population diversity in the evolution process, by replacing the same number of poor individuals with new randomly generated feasible individuals and mirror individuals with large differences generated by mirror mapping in the population, in order to increase the diversity of the population, thereby effectively improving the global optimization ability of the algorithm. Population intervention diversifies the function between the expression of PM2.5 concentration and influencing factors, and selects evolution through a more comprehensive function set, so that the algorithm can more accurately express the functional relationship between PM2.5 concentration and various influencing factors.

The paper uses information entropy as a measure of population diversity, and judges whether the current population needs intervention based on the threshold of information entropy. The information entropy is solved as follows:

1) Count the number of occurrences of the *i*-th function or terminator on the same gene position j of the population C_{ij} ;

2) Calculate the probability P_{ij} of the appearance of the *i*-th function or terminator on the same gene position *j* of the population, where *N* is the size of the population. The calculation formula is as Formula 1:

$$P_{ij} = \frac{c_{ij}}{N} \tag{1}$$

3) Calculate the information entropy of the population. The specific calculation formula of information entropy is as Formula 2:

$$H = \sum_{j=1}^{L} \sum_{i=1}^{S} \frac{1}{L} - P_{ij} \log P_{ij}$$
(2)

In Formula 2, L is the total length of individuals, and S is the total number of function and terminator.

If H is greater than or equal to the set threshold, the original population will remain unchanged. If it is less than the set threshold, population intervention will be carried out. The specific operation method is as follows:

Population intervention is to sort the population from large to small according to its fitness. For the penultimate *b* individuals, the *b* mirror individuals are replaced, and the penultimate b+c individuals are replaced by *c* random individuals to form a new population.

Mirror individual: function symbol set $F=\{+, -, *, /, sqrt, x^2, exp, cos, sin, ln, log, ~ (base 10 exponent)\}$, mirror function symbol set mirror_ $F=\{-, +, /, *, x^2, sqrt, ln, sin, cos, exp, ~, log\}$, traverse the individuals that need to be mirrored, if the *i*-th gene is the *j*-th element in F, the *i*-th gene of the individual after replacement is the j-th element in mirror_F. According to the abovementioned rules, a new individual is formed after traversing all the gene positions of the individual.

Random individual: the same rules as the initial individual generation.

2.2.3 KMGEP algorithm steps

The article adds knowledge guidance and manual intervention to the GEP to improve the evolution efficiency of the algorithm and optimize the quality of the solution. In the selection operator, the tournament selection operator with elite strategy is used to save the best individuals in the current population to ensure the algorithm converges to the global optimum [25]. KMGEP algorithm steps are as follows:

Pseudo code of KMGEP

Input: the data PM2.5 concentration and various influencing factors; population size N; The number of iterations G; genetic probability; set information entropy H_g ; Output: functional relationship between PM2.5 concentration and each influencing factor;

- 1. Initialize each chromosome in the population
- 2. Solve the fitness value of each individual
- 3. M individuals with greater fitness in the population were selected as members of the posteriori knowledge base.
- 4. While iteration termination condition does not satisfy DO
- 5. mutation operation
- 6. transposition operation
- 7. recombination operation
- 8. Evaluate the fitness
- 9. Removing inferiority
- 10. Increase superiority
- 11. Compute the current population Information entropy H
- 12. If *H*<*H*_g
- 13. b random individuals replace b individuals with the worst fitness
- 14. c individuals replace c individuals with inferior fitness
- 15. Update the posterior knowledge base
- 16. Tournament selection with elite strategy
- 17. Evaluate the fitness
- 18. End while

3 Comparative experiment

In order to verify the performance of the algorithm in this paper, taking the winter PM2.5 concentration prediction in Xi'an as an example, a comparative experiment was carried out with literature [6] and literature [9].

3.1 Data setting

The influencing factors of PM2.5 concentration established in this paper are air quality data (PM10, SO2, NO2, CO, O3) and meteorological data (minimum temperature, average temperature, relative humidity, average wind speed, air pressure, sunshine duration, precipitation, maximum wind speed, the direction of the maximum wind speed). The daily monitoring data of the Xi'an area from 2017 to 2018 (December-February) obtained through the China National Environmental Monitoring Centre. The collected data are daily averages. 70% of the data is used as the training set, and 30% of the data is used as the test set.

3.2 Fitness function selection

In the paper, the degree of fit $R^2=1$ -SSE/SST is used as the fitness function, which is the multiple correlation coefficient in statistics. Among them, SSE is calculated as Formula 3, and SST is calculated as Formula 4.

$$SSE = \sum_{j=1}^{n} \left(y_j - \hat{y_j} \right)^2$$
(3)

$$SST = \sum_{j=1}^{n} \left(y_j - \bar{y} \right)^2 \tag{4}$$

Which y_j represents real data, y_j represents predicted data, and \overline{y} represents the average value of real data. When R² is closer to 1, it indicates that the prediction accuracy is higher.

3.3 Performance evaluation

This article uses three indicators of fit (R²), root mean square error (RMSE), and average absolute error (MAE) to evaluate the prediction results. The root mean square error (RMSE) is calculated as Formula 5, and the average absolute error (MAE) is calculated as Formula 6. Where *n* is the length of the prediction set, y_i is the *i*-th true value in the prediction set, and \hat{y}_i is the *i*-th predicted value obtained by the model.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$
(5)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$
(6)

3.4 Initial parameter setting

The algorithm parameter settings are shown in Table 2.

 Table 2. Initialize parameter Settings

Parameter	Setting value	
Maximum evolutionary algebra	200	
Population size N	100	
Posterior knowledge base size	15	
Function symbol set F	+ - * / log exp ln \sim x ² sqrt sin cos abs	
Terminator set T	influencing factors of PM2.5 concentration	
Connector	+	
head length	10	

Number of genes	6
Point mutation rate	0.3
Recombination rate	0.2
Transposition probability	0.1
Length of IS	{1,2,3,4,5}
Length of RIS	{1,2,3,4,5}
Strengthening operator probability	0.2
Mirror replacement individuals b	20
Random replacement individuals c	30
Competition scale	2

4 Results and Discussion

The KMGEP model prediction model expression is shown in Formula 7, and the model prediction performance results are shown in Table 3.

$$y = e^{x_3} + (\cos \ln x_7) * \ln x_9 + 2 * x_2 + (\cos \ln x_0) * \\ \ln 10^{\ln x_0} + x_6 * x_{10}$$
(7)

Among them, y represents the concentration of PM2.5, x_0 represents PM10, x_2 represents NO2, x_3 represents CO, x_6 represents average temperature, x_7 represents humidity, x_9 represents air pressure and x_{10} represents sunshine hours. From Formula 7, the functional relationship between PM2.5 concentration and its factors can be clearly seen, and it can be known that through the evolution of the algorithm. The influencing factors of PM2.5 concentration are changed from the original 14 influencing factors of PM10, SO2, NO2, CO, O3, minimum temperature, average temperature, relative humidity, average wind speed, air pressure, sunshine, precipitation, the maximum wind speed and direction of maximum wind speed change to 7 influencing factors of PM10, CO, NO2, humidity, sunshine time, air pressure and average temperature. Related functions changed from the original 13 functions of +, -, *, /, log, exp, ln, ~, x², sqrt, sin, \cos , and abs to 5 functions of exp, \cos , \ln , + and *. Finally, the PM2.5 concentration in winter in Xi'an area is mainly related to PM10, CO, NO2, humidity, sunshine time, air pressure and average temperature. When the concentration of air pollutants CO and NO2 increase, the PM2.5 concentration also increases, which is consistent with the research results in the literature. From the above analysis, it can be seen that exponential function, logarithmic function and trigonometric function can better reveal the relationship between PM2.5 concentration and its factors. It shows that the KMGEP algorithm finally obtains functions and influencing factors related to PM2.5 concentration through selection and evolution, survival of the fittest, and the obtained function expression can effectively explain the relationship between PM2.5 concentration and its influencing factors, internal mechanism. For the control of Xi'an winter haze concentration, Formula 5 can be used to control the corresponding variables so as to achieve the purpose of effective control of PM2.5 concentration.

Comparing KMGEP with BP-ANN and CNN-LSTM models, the prediction curves are shown in Figure 1-3. In the figure, The horizontal axis shows the sample points, and the vertical axis shows the concentration of PM2.5, and r v is the true value, and p_v is the predicted value. The prediction performance indicators are shown in Table 4.

Table 3. KMGEP prediction performance

Performance indicators	R ²	MAE	RMSE
training set	0.88	16.75	21.86
test set	0.86	18.06	22.43

 Table 4. Comparison of KMGEP algorithm, BP-ANN and CNN-LSTM in prediction

Model	R ²	MAE	RMSE		
KMGEP	0.86	18.06	22.43		
BP-ANN	0.83	20.14	25.86		
CNN-LSTM	0.84	19.90	24.75		



Fig. 2. Prediction curve of KMGEP algorithm



Fig. 3. Prediction curve of BP-ANN algorithm





From Figure 2-4, it can be found that KMGEP algorithm in this paper is better than the BP-ANN and CNN-LSTM algorithms. From Table 4, it can be seen that the KMGEP calculation is 0.03 higher than the BP-ANN fit, and the average absolute error is 2.08 lower. The root mean square error is lower by 3.43. The fit is 0.02 higher than that of CNN-LSTM, and the average absolute error is lower by 1.84, and the root mean square error is lower by 2.32. The degree of fit, root mean square error and average absolute error are better than the other two models. BP-ANN and CNN-LSTM are based on deep learning models. The final model is a correlation parameter matrix, and the relationship of PM2.5 concentration and its factors is not visible. KMGEP algorithm finally obtains the specific functional relationship between PM2.5 concentration and its influencing factors, which can reveal the mechanism of PM2.5 concentration. CNN-LSTM and BP-ANN are determining the PM2.5 concentration influencing factor. The above is mainly based on correlation analysis, while correlation analysis mainly analyzes linear relationships. It is not enough for nonlinear relationship analysis. The application of the results of predecessors in PM2.5 concentration research is insufficient, which makes the influencing factors of the input model incomplete, and the KMGEP algorithm is based on the prior knowledge base established by previous achievements is used as the influencing factor, and the determination of influencing factors is more comprehensive, and the experiment proves that the accuracy is higher. In proves summary, the above analysis the competitiveness and advancement of the KMGEP algorithm in PM2.5 concentration prediction.

5 Summary

PM2.5 concentration prediction is of great significance to the prevention and control of haze. This paper presents a knowledge-guided and manual interventionbased gene expression programming. KMGEP introduces knowledge guidance and manual intervention on the basis of GEP. KMGEP algorithm establishes a prior knowledge base of influencing factors and related functions of PM2.5 concentration by mining previous research results of PM2.5 concentration prediction. A posteriori knowledge base is established by preserving high quality genes in the evolutionary population, and the algorithm can optimize the evolution more accurately and effectively through the guidance of knowledge. Then, the KMGEP algorithm maintains the diversity of the population by introducing manual intervention, and finally finds the optimal solution of the problem. In this paper, the proposed algorithm was applied to the PM2.5 concentration prediction in Xi 'an, and compared with the neural network-based algorithm in the literature, the results show that KMGEP algorithm not only effectively improves the accuracy of PM2.5 concentration prediction, but also can clearly see the relationship between the influencing factors of PM2.5 concentration and haze, which provides an important reference for haze control in China. In the future, the KMGEP algorithm can also be applied to other air mass concentration studies and other intelligent prediction fields.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62072363) and the Natural Science Basic Research Program of Shaanxi Province (No. S2019-JC-YB-1191).

References

- Perrone, M.G., M. Gualtieri, V. Consonni, L. Ferrero, G. Sangiorgi, E. Longhin, D. Ballabio, E. Bolzacchini, M. Camatini, Particle size, chemical composition, seasons of the year and urban, rural or remote site origins as determinants of biological effects of particulate matter on pulmonary cells, *Environ. Pollut.*, **176**, 215-227(2013).
- [2] R. Bono, R. Tassinari, V. Bellisario, G. Gilli, M. Pazzi, V. Pirro, G. Mengozzi, M. Bugiani, P Piccioni, Urban air and tobacco smoke as conditions that increase the risk of oxidative stress and respiratory response in youth, *Environmental Research*, 137, 141–146(2015).
- [3] H. Jianjun, G. Sunling, Y. Ye, Y. Lijuan, W. Lin, M. Honjun, S. Congbo, Z. Suping, L. Hongli, L. Xiaoyu, L. Ruipeng, Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major Chinese cities, *Environ. Pollut.*, 223,484–496(2017).
- [4] C. Cheng, W. Hongjie, L. Weisheng, F. Qiming, T. Ye, Indoor PM2.5 Prediction Based on Multi-Instance Genetic Neural Network (In Chinese), *Computer Applications and Software*, 36(5),235-241(2019).
- Z. Guowei, W. Tengjun, PM2.5 Concentration Prediction Model Based on SVM-wavelet Neural Network (In Chinese), *Sichuan Environment*, 37(6), 141-144(2018).
- [6] C. Qiang, M. Kun, Z. Huimin, C. Xianlei, Z. Minghua, Study on Spatiotemporal Variability of PM2. 5Concentrations and Prediction Model over Zhengzhou City (In Chinese), *Environmental Monitoring in China*, 31(3),105-112(2015).
- [7] Y. Sun, Q. Zeng, B. Geng, X. Lin, B. Sude, L. Chen, Deep Learning Architecture for Estimating Hourly Ground-Level PM2.5 Using Satellite Remote Sensing, *IEEE Geoscience and Remote Sensing Letters*, 16(9), 1343-1347(2019).
- [8] Z. Wenfang, L. Runsheng, T. Wei, Z. Yong, Forecasting Model of Short-Term PM2.5 Concentration Based on Deep Learning (In Chinese), *Journal of Nanjing Normal University* (*Natural Science Edition*), **3**, 32-41(2019).
- [9] T. Li, M. Hua, X. Wu, A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5),

IEEE ACCESS, (8), 26933-26940, (2020).

- [10] L. Xulin, Z Wenfang, T Wei, Forecasting Model of PM2. 5 Concentration one Hour in Advance Based on CNN-Seq2Seq (In Chinese), *Journal of Chinese Computer Systems*, **41**(05),1000-1006(2020).
- [11] F. C., Gene Expression Programming: a new adaptive algorithm for solving problems, *Complex System*, 13(2), 87-129(2001).
- [12] L. Haifeng, L Minyan, Z Min, H Baiqiao, Application of Gene Expression Programming in Software Reliability Modeling (In Chinese), *Journal of Frontiers of Computer Science and Technology*, 5(6), 534-546(2011).
- [13] M. Saeid, B. Javad, K. Keivan, Application of gene expression programming to predict daily dew point temperature, *Applied Thermal Engineering*, **112**, 1097-1107(2017).
- [14] S. Hr. Aghay Kaboli a, A. Fallahpour a, J. Selvaraj a, N.A. Rahim a b, Long-term electrical energy consumption formulating and forecasting via optimized gene expression programming, *Energy*, **126**, 144-164(2017).
- [15] I. Muhammad Farjad, L. Qingfeng, A. Iftikhar, Z. Xingyi, Y. Jian, J. Muhammad Faisal, R. Momina, Prediction of mechanical properties of green concrete incorporating waste foundry sand based on gene expression programming, *Journal of Hazardous Materials*, **384**,121322(2020).
- [16] L. Debin, X. Jianhua, Y. Wenze, M. Wanliu, Yang. Dongyang, W. Jinzhu, Response of PM 2.5 pollution to land use in China, *Journal of Cleaner Production*, 2020,244, 118741-118741(2020).
- [17] W. Zhuang, Z. Shuo, Study on the spatialtemporal change characteristics and influence factors of fog and haze pollution based on GAM, *Neural Computing and Application*, **31**(05), 1619-1631(2019).
- [18] B. Lv, W. G. Cobourn, Y. Bai, Development of nonlinear empirical models to forecast daily PM 2.5 and ozone levels in three large Chinese cities, *Atmospheric Environment*, **147**, 209-223(2016).
- [19] W. Yuan, K. Wang, X. Bo, L. Tang, J Wu, A novel multi-factor & multi-scale method for PM2.5 concentration forecasting, *Environmental Pollution*, 255(1), 113187 (2019).
- [20] L. Suixin, C. Junji, A. Zhisheng. Characterization of Ambient Fine Particles (PM2.5) Concentration and Its Influential Factors (In Chinese), *The Chinese Journal of Process Engineering*, 9(S2), (2009)
- [21] M. Zhaowei, L. Peiyu, Z. tongjun, C. Changfeng, Characteristics and Meteorological Influencing Factors of PM2.5 Mass Concentration in Two Urban Districts of Xi'an During 2015-2018(In Chinese), *Journal of Hygiene Research*, 49(01),75-79(2020).
- [22] P. Yan, Z. Ziru, W. Tingxian, W. jie,Prediction of PM2.5 Concentration Based on Ensemble

Learning (In Chinese), Journal of Beijing University of Posts and Telecommunications, doi:10.13190/j.jbupt. 2019-153.

- [23] Q. Chao, C. Tingting, L. Jia, L. Yudong. Spatio-Temporal Characteristics of PM (2.5) and Influence Factors in Typical Cities of China (In Chinese). *Research of Environmental Sciences*, **32**(07), 1117-1125(2019).
- [24] S. Aydin, S. Mohsen, K. Anikender, G. Hossein, Prediction of air quality in Tehran by developing the nonlinear ensemble model, *Journal of Cleaner Production*, 259, 120825, (2020).
- [25] Y. Changan, T. Changjie, Z. Jie, Function Mining Based on Gene Expression Programming Convergency Analysis and Remnant-guided Evolution Algorithm (In Chinese), Advanced Engineering Sciences, 36(6),100-105(2004).